Finding Meaning Through Downstream Analysis of Embeddings: A Case-Study of Knowledge Discovery for the Media and Publishing Industry

RYAN THOMAS HODGSON

Finding Meaning Through Downstream Analysis of Embeddings: A Case-Study of Knowledge Discovery for the Media and Publishing Industry

Ryan Thomas Hodgson
Supervisor: Prof. Alexandra I. Cristea

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
2024

Abstract

In recent years, the publishing sector has undergone a significant transformation in the way news and journalistic material is consumed, shifting from traditional print formats to digital platforms. This shift in the consumption of journalistic content poses both opportunities and challenges for the industry. Although traditional print publications relied on a combination of sales and advertising revenue, the expectation of free access to online media has impacted the profitability of these organisations. Consequently, publishing houses and newspapers have significantly reduced budgets, including those allocated to journalists. This reduction in journalistic staff has limited the time and resources available to produce high-quality content.

To address this issue, the research presented in this thesis explores methods to address the time-costly nature of many of the tasks that journalists and publishers perform, in order to contribute innovative tools to streamline many manual processes. Conducted in conjunction with an industry sponsor, Distinctive Publishing, this research contributes to the publishing domain through a focus on the leveraging of unsupervised learning techniques, to enable the expediting of common processes, which would often be performed manually.

The structure of this research project can be summarised in four main aspects. 1. The proposal of meta-embedding based semantic similarity searches, to enhance the quality of semantic searching of databases of social media influencers based on full-text queries. 2. An exploration of the feasibility of topic modelling algorithms, for the identification of academic topics from large volumes of literature. 3. Based on the outcomes of the exploration of topic modelling, an end-to-end framework for assisting literature analysis is proposed, and evaluated in an experimental setting. It indicates the breadth and generalisabilty of our findings, and the value that automating the analysis of literature can have for researchers, beyond journalists and media stakeholders. 4. The task of parametric dimensionality reduction using attention mechanisms in neural network encoders is proposed as a method to improve cluster-based topic models, by enhancing the dimensionality reduction processes required in such algorithms.

The findings presented in this thesis are initially based on applied research, by adapting existing algorithms to the specific domain of the media and publishing industry. However, during the research, key aspects of the algorithmic processes were identified, specifically in relation to the dimensionality reduction process necessary for cluster-based topic modelling algorithms. Based on this, a novel paradigm of research is presented and explored, through the consideration of architectural design in parametric dimensionality reduction as an effective method to improve the quality of topic modelling. The consequences of this discovery introduce new areas for future investigation, which are proposed for further exploration in ongoing research.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I would like to express my deepest gratitude to everyone who contributed to making this thesis possible.

I would like to acknowledge with great appreciation my academic supervisor, Prof. Alexandra I. Cristea, for her ongoing guidance, recommendations, and thorough proofreading throughout my research. Your help in understanding the demands of industrial research was crucial in shaping the content of this thesis, and I would not have been able to do it without your help.

I also express my sincere gratitude to Dr. Jingyun Wang for her thorough support during many of the trials of the research undertaken, for her contribution in stimulating suggestions and encouragement, and helping me to coordinate my projects from concept to publication. Her counsel during difficult times was crucial in motivating me to continue my studies. Furthermore, her assistance was invaluable in establishing a sound scientific method throughout the human studies involved in evaluating the results of my research.

To my industrial sponsors, John, and Kerry, I am deeply grateful for the opportunity to conduct this research through their company, Distinctive Publishing, and for enabling the industrial implementation of many of the findings which were identified during the course of my research. I look forward to our continued work together in bringing the fruits of our labour into the public domain.

To all University staff and administrators who have helped throughout the course of my research, I extend my deepest gratitude.

Outside of academia, I would like to thank my parents, friends, and most importantly, my partner Louis, for putting up with me throughout the years of my studies.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Industrial Context and the Necessity for Datamining in the Publishing Industry

In recent years, the publishing industry has experienced a significant shift in the consumption of news and journalistic content from print-based to online. Sales of national and locally printed newspapers have decreased by approximately half from 2007 to 2017 [7]. In comparison, journalistic material has seen significant uptake in online consumption. In 2018 74% of UK adults used some form of online method to find news, increasing to 91% of 18-24 year olds [8]. More Recently, in 2020, 70% of individuals reported to have read or downloaded online media in the UK [9]. In a similar study conducted by Ofcom[1], the regulator for communications services in the United Kingdom, it was found that in 2023 83% of 16-24 year-olds reported using online news sources of any variety, including social media, podcasts, and all other websites, with this being driven by social media, where 71% of this age group reporting using it for news. However, this study also reported that for survey participants

[1] https://www.ofcom.org.uk/about-ofcom/what-we-do/what-is-ofcom/

aged 25-34, and 35-44, the reach of broadcast television and print newspapers is extended by additional online consumption through Broadcaster Video-on-Demand services, and newspaper websites and apps. In older audiences ranging 45-54, 55-64, 65-74, and 75+, the percentage of responses indicating online consumption of news steadily decreases to 65%, 61%, 53%, and 38%, respectively [10]. This indicates that while there is a steady shift amongst younger audiences towards online and social media consumption, there remains a potential for traditional media outlets to leverage online consumption to enhance their reach. To address this shift in the medium of access to journalism, many publishers have found it necessary to change their revenue focus to online advertising, or subscription-based models, with online advertising undergoing an increase from 16% to 48% of the total UK advertising spend from 2007 to 2017 [7].

The fundamental change in the way journalistic content is consumed presents unique opportunities and challenges to the industry. While print publications had typically earned through a mixture of sales and advertising, the assumption that access to online media such as news articles should be free has had an impact on the profitability of such organisations. This, in turn, has led to publishing houses and newspapers drastically reducing the budgets of their internal staff. One such instance of this is the reduction of budgets for journalists. By reducing the number of journalists active within a publishing house, the amount of time and resources that may be spent on producing meaningful journalistic content is reduced. To address this problem, the initial scope of the Ph.D. project was defined *to investigate predictive and prescriptive analytics to assist journalists and publishers by providing novel tools that can accelerate many of the processes that are typically carried out manually.*

1.2 Problem Space and Research Motivation

With regard to the publishing industry, the overarching objectives to be solved by this research were, first, the expediting of common processes that are time-costly, due to the current need for manual intervention; and secondly, to benefit advances in AI to enrich customer experience (here, specifically the experience of their journalists and that of their readers). Through discussion with the industry sponsor, it became clear that many processes could be automated, or semi-automated, to free up more working time to focus on other aspects of the journalistic process. For this research project, the focus on AI application research is made on two main categories, which were identified by the industry sponsor as being directly beneficial to the industry; *(1) the identification of relevant audiences for journalistic content, and the identification of themes, and (2) relationships between these themes, based on the semantic content of large volumes of text.* While both of these overarching categories may at first appear to be very different, they are approached within this research from the perspective of textual semantics, which is a persistent theme throughout. In the following, we explain for both of these categories the reasons why the task is valuable in an industrial context, and why it is, in its standard form, normally time-consuming for journalists and publishers.

1.2.1 Identification of Relevant Audiences for Journalistic Content

Given an item of journalistic material, which could be a magazine or news article, a common process within the industry is the promotion of material to relevant audiences through social media channels to enhance the readership of an article. Publishers may promote an article on social media, hoping followers will share it. However, this is reliant upon the number of followers that a publisher's social media account has, and whether these followers would be interested in the subject topic of the article. To mitigate this reliance upon a large social following, a publisher may also seek to identify other 'influential' accounts, which discuss topics similar to the subject of the article, through *manual identification*, with the aim of contacting them to see if they are interested in the article, and would consider sharing it with their audience.

This involves the manual searching of social media platforms using the available interfaces and keyword terms - something that requires both a prior understanding of the subject topic of the article and enough time to be dedicated to the evaluation of any identified accounts in terms of both their ‘influence’ (commonly based on the number of followers, but also on more advanced measures [11]), and the relevance of the account to the given article in terms of the subject topic - the *semantic similarity*. The process of manually searching and evaluating potential candidate influencers is inherently time-intensive. This is attributed to the requirement for an individual, who may not be intimately familiar with the subject matter of an article if they are not its author but are tasked with its promotion, to thoroughly read and comprehend the article’s content. Subsequently, the individual must apply this knowledge to conduct a manual assessment of candidate accounts, focusing on the nature of posts made by the candidate. Additionally, the evaluator must determine whether the candidate holds sufficient influence within the subject area to warrant engagement concerning the article.

Thus, it is proposed that there is a potential for expediting this process through the leveraging of neural-embedding techniques to shortlist potential candidate influencer accounts based on semantic similarity. This is based on the assumption that both the content posted by an account, and the user-defined biography of the account can be taken into consideration when evaluating whether a social media account is a suitable candidate for contacting regarding a given article during the manual process. The research carried out in seeking to address this problem is presented in Chapter 4, where a meta-embedding focus is adopted for the construction of hybridised neural-embeddings produced from different data sources.

1.2.2 Assisting Readers in the Comprehension of Literature

In contrast to identifying relevant audiences for content promotion, which is more closely beneficial to *publishers*, the identification of themes in large volumes of textual content, such as social media posts or recent news publications can provide a direct benefit to *journalists* by providing a means for detection of topics which may indicate a general public interest in a particular theme. Subsequently, this serves as an effective instrument for identifying topics on which to compose writings, with the anticipation that such efforts may garner a larger readership.

Software solutions such as *Google Trends*[2] and the X (formally Twitter) *Trending Now* page[3] are valuable tools used to detect current trends in public interest, and were reported by the industry sponsor as being used within their business when planning journalistic articles. In the case of *Google Trends*, trends are defined based on the search volume of groups of words or phrases which share a common theme within a given time frame, with 'bursting' trends being highlighted to the user when the search volume is higher than usual. For the *X Trending Now* page, trending topics are identified based on the volume of keywords and hashtags mentioned within posts on the platform, where a hashtag is a user-defined keyword prepended by the '#' symbol.

Both of these tools are valuable in enabling the identification of trending topics of public interest which can assist in the planning of journalistic material, however, rely upon two major organisations - namely *Google*, and *X* - to process and provide results to users, with little information being provided on how the search volumes and trend criteria are defined. Focusing only on *Google* and *X* may lead to a bias in any analysis, because the results take into account only the traffic of a particular platform. Furthermore, there are many alternative sources of textual content available online and internally within the industry.

Thus, the second aspect of this research project was proposed, which is based upon investigating whether analysis of large volumes of unstructured and typically unlabelled data could provide a benefit to the industry by the extraction of topics and subsequent analysis of these topics from a semantic and temporal perspective. To achieve this, the paradigm of topic modelling was leveraged, and subsequently extended, to facilitate the semantic and temporal analysis of unstructured texts. Within this research, the focus is made upon academic literature with the objective of enhancing literature analysis for researchers and students, due to the readily accessible data and easy dissemination of research findings without the risk of sharing data or processes which are vital to the sponsor's industry position. Still, the findings of this research, which are the focus of Chapters 5 and 6, could be transferred to

[2] https://trends.google.com/trends/

[3] https://x.com/explore/tabs/trending

any textual domain including sources of journalistic content or diverse multi-source social media data with relative ease.

1.3 Necessity of Unsupervised Learning in Addressing the Needs of Publishers

Over the course of the research, several domains were identified to address the issues mentioned above, formed through collaboration with the industrial sponsor. A wealth of data is available for investigation within these areas, with data taking the form of magazine articles, news articles, and social media sources. However, much of this data is made up of largely unstructured texts with few instances of labelled data. Therefore, this presents difficulties in the application of commonly used supervised learning techniques, which, in turn, implies the need to use unsupervised learning approaches.

To address this lack of valuable labelled data, several avenues are available for consideration. Semantic search [12] of vector representations of text is a valuable tool in the retrieval of semantically similar documents, for example, in recommender systems, and this is one such method of making sense of unstructured data that was considered in Chapter 4.

In contrast, topic modelling can address some of the issues currently faced by the transition from print to online media, by assisting in making sense of large volumes of text online, which would be infeasible by a human. By introducing topic modelling to aid in understanding extensive amounts of text, and then utilising explainable AI to illustrate connections and patterns of semantic similarities between topics (Chapters 5 and 6), it was shown that these techniques can be beneficial in uncovering fresh insights within textual data. For example, both algorithmically different topic modelling approaches, Latent Dirichlet Allocation [13], and Top2Vec [6], achieve the identification of coherent topics in a case-study literature dataset. However, simply finding topics in the literature does not directly improve the literature review process or expedite comprehension of large volumes of literature. Consequently, the temporal and semantic analyses of the identified topics, presented in Sections 5.4

and 6.4.1, enhance the comprehensibility of these topics. These analyses provide an overview of the general temporal trends within the literature, thereby facilitating the identification of areas that are either underexplored or oversaturated. In addition, such analyses illuminate potential interconnections between topics that may have previously been overlooked. This proposed approach is additionally particularly valuable for publishers and journalists with limited resources, where the identification of temporal trends and semantic relationships would be beneficial in providing cues for novel writing topics or subjects that merit a deeper investigation.

Based on the results of this research, an in-depth analysis was conducted into dimensionality reduction (DR), which is a key step in the topic modelling process presented in Chapters 5 and 6. Subsequently, investigations were conducted to address improvements in downstream clustering of word embeddings (a key step in topic modelling) through attention mechanisms using a parametric version of the Uniform Manifold Approximation and Projection (UMAP) [4,14] algorithm (Chapter 7). The findings presented in Chapter 7 represent the culmination of the Ph.D project, through the exploration of a novel paradigm in topic modelling research, wherein an analysis of the transformer-encoder architecture in parametric dimensionality reduction indicates the potential for the improvement of cluster-based topic modelling techniques.

1.4 Introducing the Industry Sponsor

The industrial sponsor and partner for the research presented in this thesis is **Distinctive Media Group**[4], a publishing house based in Newcastle upon Tyne, England. The sponsor provides the **Reveela**[5] content community, which aims to connect newsworthy content directly to the media, through leveraging applied research into Natural Language Processing (NLP) to provide benefit to the media and publishing industries.

The research presented in this thesis was sponsored in part by Distinctive Me-

[4] www.distinctivegroup.co.uk

[5] www.reveela.com

dia Group through the European Regional Development Fund, Intensive Industrial Innovation Programme, with the aim of designing and advancing applied research to benefit the publishing industry. As part of this, some of the research has been commercially implemented and is already being used in industry, with other aspects of this thesis being continually expanded upon to contribute directly to industry.

1.5 Generic Research Questions

Overarching research questions are defined and designed to address the general requirements outlined by the industry sponsor and the initial project brief. The order of these research questions follows the overall progress of the Ph.D. project, where exploratory analysis was focused on general, unsupervised learning techniques to contribute to the publishing industry by a journalist retrieval system for social media data which is presented by the first general research question, and addressed in Chapter 4. This in turn contributed to the subsequent area of research, which focused on applying topic modelling techniques to contribute to large-scale literature analysis, which, is presented in the second general research question and analysed in Chapters 5 and 6. Following this, the final general research question is defined, and focuses on investigating techniques for improving dimensionality reduction in cluster-based topic modelling algorithms. This is formed of contributing research related to a transformer-encoder architecture which can be used in dimensionality reduction pipelines, which is the main focus of Chapter 7. For each research question, the chapters where these general research questions are addressed are provided, wherein each chapter provides a fine-grained breakdown of specific research questions and objectives. A general overview of the concepts explored within this thesis is presented in Figure 1.1.

GRQ1: How can journalistic influencers be retrieved with accuracy for a publishing enterprise?

For the publishing industry, real-world data is commonly unstructured and unlabelled, where, for example, the industry sponsor possesses a significant portion of

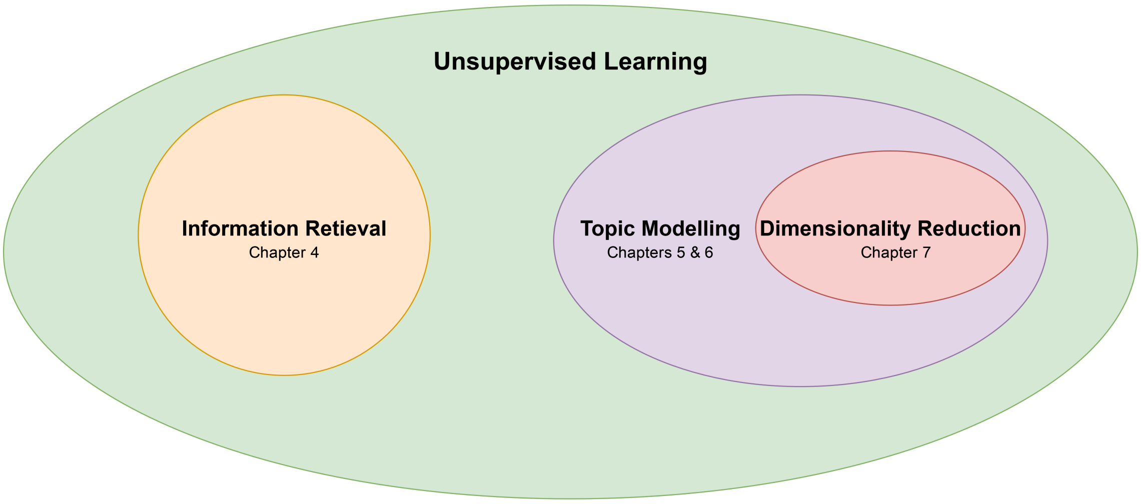


Figure 1.1: Overview of the concepts explored within this thesis.

unstructured textual data in the form of magazine articles and press-releases, and is seeking to investigate the leveraging of external data sources such as social media streams. The unstructured nature of these diverse sources of textual data, which is commonly unlabelled, makes the application of common supervised learning techniques, such as classification or regression, difficult. Moreover, supervised learning techniques are inherently constrained by the data on which they are trained, rendering them less effective for discerning novel patterns and trends within textual media.Thus, in order to provide meaningful industry impact through data analysis, unsupervised analysis techniques are necessary; however, the best approach to address this needs to be explored. One of the first investigations into this is discussed in Chapter 4, where semantic search is proposed as a method for the retrieval of relevant accounts on social media platforms, based on the provision of an item of full-text journalistic content. For this research question, the term 'journalistic influencer' refers to an influential social media account, with a potential for collaboration with journalists or publishing houses to help promote or disseminate journalistic content. Although accurate recommendations based on zero-shot vector searches using information retrieval techniques presented promising results, it was found that the construction of meta-embeddings from both user-defined biographies, and posted content of social media influencers indicated a positive influence upon the accuracy of recommendations. These are discussed in greater detail in Chapter 4.

A key limitation for journalists and publishers is the *identification of relevant audiences and key players within those audiences, to whom an item of content is relevant.* For publishers, the promotion of journalistic content can serve as an effective method of increasing the reach of the material, and it is this domain that was established as a suitable area of investigation at the beginning of the Ph.D. This is formed of an information retrieval approach to the recommendation of social media influencer accounts, which are relevant to a given piece of content.

Several prerequisites were established by the industry sponsor and have subsequently contributed to the general direction of the research project. Firstly, the sponsor requested the necessity for the recommendation of influencer accounts based on the provision of a full-text article. Secondly, recommendations must account for a degree of 'recency' based on influencer discussions. In this context, 'recency' entails ensuring that recommendations take into consideration the temporal aspect of an identified relevant result, thereby also prioritising the recommendation results that have been posted in the most recent period. A consequence of incorporating this consideration is the necessity to balance the prioritisation of temporal relevance against the precision of retrieval results. Thirdly, the commercial implementation must provide relevant results, without the need for extensive retraining, and adapt to the 'cold start' problem [15] commonly experienced by existing recommender systems.

Based on these criteria, the following in-depth research questions were formulated.

- **RQ1.1:** *How may information retrieval techniques be adapted to facilitate the identification of relevant social media accounts for a given article of journalistic content, based on a user's posted content?* (Answered in Chapter 4.)

- **RQ1.2:** *How does hybridisation of tweet and user biography embeddings contribute to recommendation quality in the retrieval of social media accounts based on full-text querying?* (Answered in Chapter 4.)

- **RQ1.3:** *Can forward time-decay functions be applied to recommendation ranking to improve the temporal relevance of results, and how does this affect the* *quality of recommendations?* (Answered in Chapter 4.)

GRQ2: What AI techniques can be defined to assist readers in the comprehension of literature?

While the recommendation of relevant social media influencers, based on full-text querying, presents a benefit to the publishing domain, it is necessary to consider alternative methods of full-text analysis. One such unsupervised learning technique is topic modelling, which can mitigate the lack of labelled data made apparent throughout the research in the publishing industry while still providing a mechanism for extracting valuable information from text. Therefore, in addressing the second general research question, it was deemed necessary to analyse how topic modelling can assist with analysis of large volumes of unstructured texts. Specifically, in relation to academic literature, it was deemed necessary to investigate how topic modelling, and subsequent analysis of identified topics within academic literature can assist in the comprehension of literature. The selection of academic literature as the focus for this aspect of the project was due to the intellectual property restrictions associated with the sharing of data provided by the industry sponsor, in which the publication of resources related to industry data could risk affecting patent applications. Thus, a focus on academic literature, which may be freely accessed in agreement with academic publishers was performed.

In addressing this general research question, the following specific questions can be defined:

- **RQ2.1:** *How may topic modelling algorithms be applied to assist in the comprehension of large volumes of academic data in an automatic manner?* (Answered in Chapter 5.)

- **RQ2.2:** *How can the semantic relationships between topics in the literature contribute to understanding of topics during exploratory literature analysis?* (Answered in Chapter 5.)

- **RQ2.3:** *How can analysing the temporal trends of identified topics in literature contribute to understanding of topics during exploratory literature analysis?*

(Answered in Chapter 5.)

- **RQ2.4:** *Can rule-based extraction of named biomedical resources contribute to the comprehension of automatically generated literature topics?*

- **RQ2.5:** *Can an end-to-end framework, built upon cluster-based topic modelling confer a benefit to researchers through assisting the literature review process?* (Answered in Chapter 6.)

- **RQ2.6:** *What are the effects on cognitive load, technology acceptance, and general acceptance of users when using a software implementation based on our framework?* (Answered in Chapter 6.)

- **RQ2.7:** *How does the proposed literature analysis framework and software implementation influence the accuracy of the identified results in practical use cases for medical researchers?* (Answered in Chapter 6.)

GRQ3: How can dimensionality reduction be used to improve the accuracy of text clustering and, subsequently, topic modelling?

During the investigation of **RQ2**, the Top2Vec algorithm [16] was identified as a suitable method for the identification of topics within the academic literature. An introduction into the algorithmic process of Top2Vec is provided in Section 3.2.3, however, for now, it is worth noting that a key step for this algorithm is the reduction of highly dimensional text embeddings, into a lower-dimensional representation, to ensure that they can be effectively clustered to identify topics. Therefore, as part of the investigation into the effects of topic modelling upon the comprehension of literature, it was proposed to consider how to improve clustering performance through adapting the dimensionality reduction process. This is made up of two main objectives, which are discussed in Chapter 7. Initially, a comprehensive evaluation of existing methods for dimensionality reduction was performed, followed by the proposed adaptation of parametric dimensionality reduction within a supervised learning framework using attention mechanisms. Based on the results of this initial evaluation, the proposed parametric dimensionality reduction pipeline is adapted to facilitate enhancements to topic modelling directly. This avenue of research is formed of the hypothesis that the transformer-encoder architecture, which has demonstrated success across many domains in NLP in recent years, may be a viable method for the parametric dimensionality reduction of highly dimensional embeddings through the UMAP framework, where its influence upon clustering is investigated using benchmark datasets. On the basis of the findings, an investigation is then performed to analyse whether the transformer-encoder can subsequently contribute to improvements in cluster-based topic modelling, which was the primary method of obtaining topics within the literature investigated in this thesis. This can contribute to improving the quality of topic modelling when used in frameworks such as Top2Vec. As part of this investigation, the phenomenon of the "vanishing gradient problem" [17] was taken into consideration, where, during backpropagation in neural networks, the multiplication of layer gradients can cause an exponential decrease in gradients, leading to a slowdown or halt in training progress. It has been observed that an increase in the sequence length within a model generally results in a reduction in the magnitude of the gradient [18]. Given that the objective of the study is to enhance the DR process, and input sequences will therefore be large, it was considered necessary to investigate potential contributions to this domain. This was performed by incorporating additional residual connections into the conventional transformer-encoder architecture. A residual connection in this context is defined as the summation of the initial high-dimensional embedding and the output from the transformer encoder. This approach aims to directly enhance the quality of topic modeling outcomes by alleviating the effects associated with the vanishing gradient problem [17], as discussed in Section 7.6.1.

In addressing this general research question, the following specific questions can be defined:

- **RQ3.1:** *How do current dimensionality reduction algorithms affect accuracy in text clustering tasks?* (Answered in Chapter 7.)

- **RQ3.2:** *How does output dimensionality of dimensionality reduction models influence performance in text clustering tasks?* (Answered in Chapter 7.)

- **RQ3.3:** *Can small portions of labelled data used in the metric learning of dimensionality reduction contribute to improvements in downstream clustering accuracy?* (Answered in Chapter 7.)

- **RQ3.4:** *Can the introduction of attention mechanisms within neural networks (further) improve the metric learning of dimensionality reduction algorithms, in terms of clustering accuracy?* (Answered in Chapter 7.)

- **RQ3.5:** *Can the transformer-encoder neural network positively influence downstream clustering when applied to cluster-based topic modelling tasks?* (Answered in Chapter 7.)

- **RQ3.6:** *What are the implications of introducing additional residual connections into the transformer-encoder pipeline for dimensionality reduction, when dimensionality reduction is used for cluster-based topic modelling?* (Answered in Chapter 7.)

1.6 Research Contributions

The following research contributions have been made based on the research conducted in this thesis:

- The demonstration of semantic search as an effective method for social media account retrieval based on full-text querying, in the novel, and niche, domain of recommender systems for full-text journalistic queries. (**Chapter 4**.)
  - Meta-embeddings derived from social media data effectively enhance semantic search capabilities, which is a novel approach to the domain of influencer recommendation for the publishing industry.
  - The experimental evaluation of forward-time decay when ranking retrieval results is shown to enhance the capabilities of the overall information-retrieval approach for the industry application by ensuring to suggest recommendation candidates which have recently discussed a semantically-similar topic. It is observed that this negatively influences retrieval ac-

curacy, and an evaluation of the weighting of the influence of temporal-relevance during re-ranking supports this empirically.

- The proposal of a novel framework for assisting in the discovery of new knowledge in medical literature. (**Chapters 5 and 6**.)
  - Comparison of Latent Dirichlet Allocation, and Top2Vec topic models on the same task of automated literature analysis, which highlights the advantages of Top2Vec in expediting topic analysis by eliminating the need for manual intervention and ensuring the leveraging of semantic-aware language models.
  - The novel proposal of semantic similarity methods, as a means for identifying relationships between topics in academic literature, thus enhancing knowledge discovery during literature analysis. Analysis of the relationships identified between topics reflect those found within the literature by domain-experts.
  - Evaluation of a proposed literature analysis framework through a study conducted with medical experts and students, which demonstrates the effectiveness of the framework in highlighting novel comparisons between literature topics, and enables knowledge discovery through facilitating the identification of new topics and items of research which are relevant to the literature analysis task.
- The proposal of parametric dimensionality reduction as a method to enhance cluster-based topic models. (**Chapter 7**.)
  - A comparative study of existing dimensionality reduction algorithms on downstream clustering by the K-Means algorithm provides a novel evaluation of the influence of target output dimensionalities for well-established dimensionality reduction algorithms.
  - The novel proposal of a transformer-encoder-based parametric dimensionality reduction pipeline, which is unexplored within the field of DR

and contributes to improved performance in the clustering of neural-embeddings.

– Adaptation of the transformer-encoder through the introduction of additional residual connections enhances the quality of parametric dimensionality reduction when applied to cluster-based topic models, which is demonstrated through both quantitative and qualitative evaluation of topic modelling results.

1.7 Publication List

The following publications were produced during the thesis and have direct impact on its content, as described below.

Hybrid Weighted Retrieval of Twitter Users for Temporally Relevant Full-Text Querying in the Media Industry

13th International Congress on Advanced Applied Informatics Winter, 2022

URL: `https://ieeexplore.ieee.org/abstract/document/10123489`

Relevant Chapters: 4

Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis

arXiv preprint, 2024

URL: `https://arxiv.org/abs/2403.01921`

Relevant Chapters: 5

Wide-scale Automatic Analysis of 20 Years of ITS research

Intelligent Tutoring Systems, 2021 (**Core A ranking**)

URL: `https://link.springer.com/chapter/10.1007/978-3-030-80421-3_2`

Relevant Chapters: 5

A Topic-Centric Crowdsourced Assisted Biomedical Literature Review Framework for Academics

Educational Datamining, 2022

URL: `https://educationaldatamining.org/edm2022/proceedings/2022.EDM-posters.76/`

Relevant Chapters: 6

A co-designed Framework for the Analysis of Large Volumes of Biomedical Literature

Submission: *Journal of Biomedical Informatics, 2024* (**IF: 6.32**)

URL: `https://www.sciencedirect.com/journal/journal-of-biomedical-informatics`

Relevant Chapters: 6

Partially-Supervised Metric Learning via Dimensionality Reduction of Text Embeddings using Transformer Encoders and Attention Mechanisms

Accepted: *IEEE Access Journal, 2024* (**IF: 3.9**)

URL: `https://ieeexplore.ieee.org/abstract/document/10536728`

Relevant Chapters: 7

Enhancing Cluster-Based Topic Models through Transformer-Encoder Facilitated Dimensionality Reduction with Additional Residual Connections

Under Review: *The 10th International Conference on Machine Learning Optimization & Data Science, 2024*

URL: `https://lod2024.icas.events/`

Relevant Chapters: 7

1.7.1 Publications Not Discussed in this Thesis

Additionally, I have collaborated in other publications during my PhD, whilst expanding techniques and methods for my thesis, as follows.

Using AI, ML and Sentiment Analysis to Increase Diversity and Equity in Technology Training and Careers

Educational Datamining, 2022

URL: `https://educationaldatamining.org/EDM2022/proceedings/2022.EDM-doctoral-co`
`106/2022.EDM-doctoral-consortium.106.pdf`

MEMORABLE: A Multi-playEr custoMisable seriOus Game fRAmework for cyBer-security LEarning

Intelligent Tutoring Systems, 2022 (**Core A ranking**)

URL: `https://link.springer.com/chapter/10.1007/978-3-031-09680-8_29`

1.8 Prologue

After establishing and introducing the general scope of the thesis and the research questions which will be investigated, the next chapter will address the general overview of background, algorithms, and problems which require necessary prior introduction, as well as an analysis of the literature where required.

CHAPTER 2

Background and General Literature Review

This chapter serves as a general overview of the themes, algorithms, and problems which will be covered throughout the thesis. Specific and in-depth analyses of the literature are additionally provided in the respective chapters where they are discussed.

2.1 Language Models and Embedding Techniques

In Natural Language Processing, the representation of language in a vector of real numbers is essential in facilitating subsequent analyses, given that computers, in essence, process information numerically. Thus, the numerical representation of language, characterised as an embedding [19], typically involves the mapping of vocabulary into n-dimensional vectors. Initial approaches to text embedding handled this task using techniques such as hot encoding [20], where the categorical features of text (the words or characters), are encoded as sparse vectors containing all zeros except for a single 1 in the i^{th} index of the vector, such that all unique categorical features are encoded by a unique sparse vector. This results in N sparse vectors with N being the number of unique categorical features within the data. Based on these vectors, sequences of text can be represented as n sparse vectors, where n

corresponds the number of unique categorical features in the text. Conversely, the bag-of-words [21, 22] methodology for text representation characterises text based on word frequency within a document. This method constructs a vector wherein each element signifies a word from the vocabulary, and the value of each element reflects the frequency of occurrence of the respective word within the document. For both of these early approaches to text embedding, a significant limitation is the failure to account for the context in which the words are used in a text, given that the individual frequencies or positions of words are encoded.

In order to mitigate the limitation of early frequency-based embedding techniques in failing to account for context, more recent research has sought to represent text in embeddings where the distance between words, sentences, or documents encoded into vectors is smaller if the terms are semantically similar. In these cases, for example, if the words *"boy"*, *"man"*, and *"grapes"* are encoded into individual embeddings, the distance between the embedding for *"boy"* and *"man"* would be the smallest [19]. Such embeddings are of great value to text analytics, where downstream analysis using text embeddings has been demonstrated in the domains of text classification, sentiment analysis, named entity recognition, biomedical text mining, and topic modelling [23], to name a few.

This vector or matrix representation of the content of a word, sentence, or document has progressed from context-free methods such as bag-of-words [21], and term-frequencies [24], to context-dependent representations which can capture the syntactic and semantic properties of language [1, 25, 26]. In recent years, the advent of pre-trained language models, which can produce contextual embeddings of language using unsupervised learning techniques [26, 27], has led to significant improvements in a range of tasks. Notably, the pre-trained transformer network BERT [26] was instrumental in pioneering an architecture founded on the transformer network [28], utilising an unsupervised pre-training approach on an extensive corpus to attain state-of-the-art results across various downstream benchmarking tasks, such as GLUE (General Language Understanding Evaluation), SQuAD (Stanford Question Answering Dataset [29]), and SWAG (Situations With Adversarial Generations [30]). Furthermore, the embeddings produced by pre-trained language models like BERT [26], and subsequent derivatives [31], have been demonstrated in a number of downstream text mining tasks to contribute to improved performance. Examples include topic modelling [16,32], clustering [33], and sentiment analysis [34], to name a few.

2.1.1 Early Neural Embedding Techniques

Unlike prior classical techniques for the generation of word embeddings such as bag-of-words, term-frequency, and N-gram [35] based approaches, the seminal work of Word2Vec introduced two methods for the computation of continuous vector representations of words [1], titled Continious Bag-Of-Words (CBOW), and Skip-gram models respectively. These approaches are essentially a method of training a Feed-Forward Neural Network Language Model (NNLM) via the prediction of the current word, based on the words before and after it in a sequence in the case of CBOW, or the prediction of surrounding words given the current word in the case of the skip-gram model. A visual representation of these training strategies is presented in Figure 2.1. By training the NNLM over a large corpus of 6 billion words, the model is able to learn semantic relationships between words [1]. This provides the unique advantage of the generation of embeddings which take into account the context of the surrounding words; however, it does not account for the ordering of contextual words. While this technique is relatively outdated by current standards, it is worth providing a brief overview due to the prevalence that this training technique will have in relation to subsequent work into neural embedding techniques. Notably, this is a precursor to the masked language modelling technique (MLM) that was pioneered in the work presenting BERT [26].

The CBOW and Skip-Gram learning methods presented under the umbrella term of Word2Vec [1] have been demonstrated to be effective in the generation of embeddings which can account for the context in which words are used, taking into account the surrounding words during computation of the embedding. However, these are intended for the purpose of computing embeddings based upon a single word. In comparison, in many NLP domains, it is preferable to compute a vector based on an entire sequence, paragraph, or document. In the case of Word2Vec,

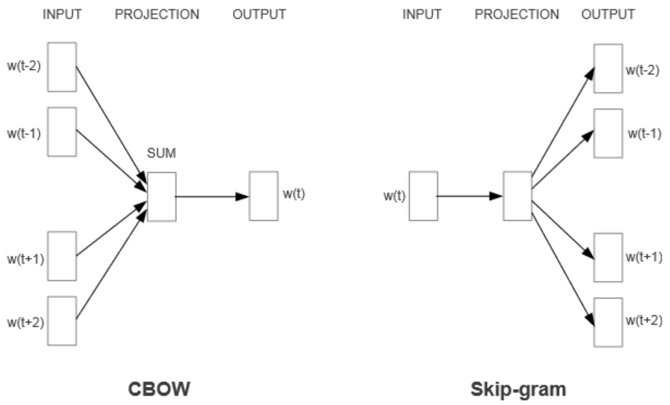
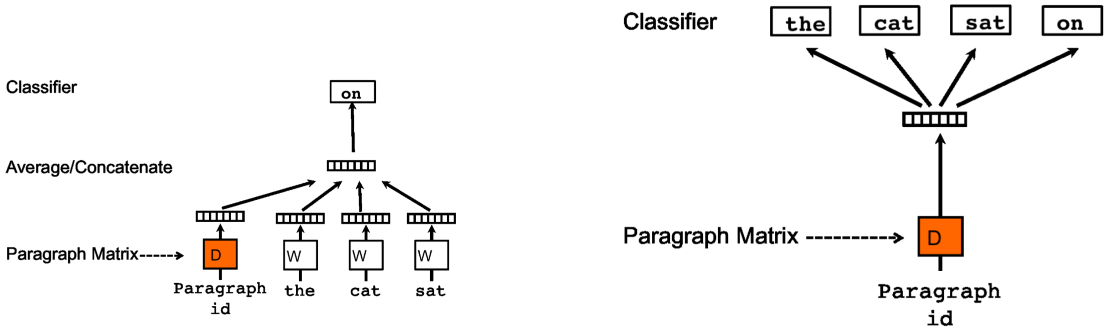


Figure 2.1: Overview of the strategies of CBOW and Skip-Gram for training Word2Vec embeddings, taken from [1]

a common technique to achieve this is the averaging of all single word embeddings into one embedding representing the entire text sequence [36]. In comparison, an extension to the Word2Vec framework was presented through Doc2Vec [2], which permits the learning of an overall embedding representing the entire text sequence. From the perspective of the CBOW training strategy, this is performed through the implementation of an additional paragraph vector, such that when training the word vectors, the document vector is also trained. I.e., the training of the paragraph vector happens along with word vectors when predicting the next word in a text window. In the original work, this strategy is titled Paragraph Vector Distributed Memory (PV-DM). In comparison, building upon the skip-gram model of Word2Vec, the Paragraph Vector Distributed Bag of Words method for document embeddings (PV-DBOW) technique for variable length document embeddings trains the paragraph vector by ignoring the context of the words in the input, instead tasking the model for predicting words randomly sampled from the paragraph. This presents the advantage of requiring a reduced memory overhead compared to PV-DM, as individual word vectors do not need to be stored during training. An overview of these training strategies is presented in Figure 2.2.At the time of using this for the thesis research, the paragraph vector (now Doc2Vec) model achieved state-of-the-art results on several text classification and sentiment analysis tasks, including sentiment analysis on the Stanford Sentiment Treebank Dataset [37, 38], and IMDB dataset [39]. Although these neural embedding methods are considered outdated by current standards, the Doc2Vec method of generating embeddings remains valid for tasks where the length of documents varies considerably, given that current pre-trained transformer based embedding models experience a limit in the number of input tokens [40, 41]. Furthermore, document embeddings can be easily trained using modern hardware, which permits the learning of embeddings based on a provided corpus, which is especially valuable in industry where costs incurred when computing embeddings must be minimised. This presents an advantage compared to pre-trained language models, which can fail to perform adequately when exposed to vocabulary which they were not trained upon [42]. It is for these reasons the Doc2Vec embedding technique is applied in Chapters 5 and 6 as an embedding model for handling variable length inputs and niche vocabulary.



(a) Paragraph Vector, Distributed Memory (PV-DM)

(b) Paragraph Vector, Distributed Bag-of-Words (PV-DBOW)

Figure 2.2: Overview of the training strategies for Doc2vec, where the document embedding is optimised, taken from [2]

2.1.2 Transformer Networks and the Emergence of Pre-Trained Language Models

In recent years, one of the most notable contributions to the field of NLP can be found in the proposal of the seminal work, Pre-training of deep bidirectional transformers for language understanding [40] (BERT hereafter). One important aspect of the BERT design is that it employs transformer networks as the core aspect of the neural network. These were introduced by [5] in "Attention Is All You Need", where the transformer network is introduced, being made up only of attention mechanisms, and fully-connected layers. Attention mechanisms perform the computation of a context vector, representing the relationship between the layer output and the input, where the context vector is a weighted sum of the hidden states of the network at each time step. This guides a model to focus on specific components of the input sequence, rather than the whole vector sequence. In language modelling tasks, this allows a model to focus on specific words within a sentence or speech, which may provide the most contextual information. In [28], the transformer network is based solely on attention mechanisms and leverages the proposed "Scaled Dot-Product Attention", and Multi-Head Attention, to achieve improved performance in machine translation tasks [28]. Subsequently, the transformer architecture has contributed to improvements in benchmark performance in several NLP tasks, including question answering, sentence continuation, named entity recognition, and language understanding [26] [31]. This has been advanced through the introduction of the pre-trained transformer in works such as BERT [26]. As part of the research presented in this thesis, the attention mechanism and transformer-encoder architecture is investigated, to consider whether contributions can be obtained with regard to the reduction in dimensionality of highly dimensional embeddings, as introduced in **Research Question GRQ3** and Chapter 7. A detailed explanation of the transformer is presented in Section 3.2.6.

2.1.3 Information Retrieval Systems from the Perspective of Vector Searches

One advantage of representing text in neural embeddings is the opportunity to utilise subsequent analysis methods upon the embeddings. From an unsupervised learning perspective, which was outlined in Section 1.3 as being best suited to contribute to the media and publishing industry, the task of information retrieval (IR) demonstrates one such downstream application of neural embeddings. IR can generally be defined as the retrieval of information from large data repositories [12], and broadly covers tasks related to the representation, storage, and searching of data for the purpose of knowledge discovery as a response to a given query [43, 44].

When searching large volumes of text, one such method of IR is presented by measuring the distance between embeddings of the overlying text. At the most basic level, this could be the measurement of a straight-line distance between a query embedding, and all possible candidate document embeddings, using Euclidean distance, where the distance is calculated as the square root of the sum of the squared differences between the two vectors [45]. Alternatively, the measurement of the cosine similarity between the query and candidate embeddings can be calculated through measurement of the cosine of the angle between a query and candidate vector [46], detailed in Section 3.2.1. After calculating all possible distances using either method, it is possible to select candidate documents with the smallest distance as having the closest degree of semantic similarity to the original query. This can be termed a brute-force approach, as it requires the exhaustive calculation of a similarity measure for all candidate embeddings. Such brute-force comparison of semantic similarity can be costly in large datasets, particularly since when comparing a vector with a candidate set of n vectors, there will be $O(n)$ comparisons necessary, before even taking into account the complexity of the distance calculation. Thus, in addressing this costly nature of semantic similarity comparison, Approximate-Nearest-Neighbour (ANN) searches have gained ground. ANNs approach the task of the nearest-neighbour problem by focusing on a slight reduction in accuracy, in order to greatly enhance computational efficiency and retrieval time [47]. Generally, ANNs can be divided into four major categories: hashing-based [48, 49], tree-based [50, 51], quantisation-based [52, 53], and graph-based [3, 54, 55]. Of these, graph-based algorithms have emerged as an effective method of ANN searching in recent years, due to the ability to provide a superior trade-off in terms of accuracy versus efficiency [3, 54, 56–58]. Notably, graph-based ANNs have merited research into vector searching in large tech companies such as Microsoft [59, 60], Alibaba [61, 62], and Yahoo [63–65]. For these reasons, graph-based ANNs were considered for the vector search strategy provided in Chapter 4, however it is worth establishing a deeper understanding of the underlying principles of these approaches to retrieval tasks.

Graph-based approaches to ANN searching generally approach the task of retrieval with the objective of efficiently retrieving approximate nearest-neighbours from data through the construction of a graph [55], where each note of the graph represents a data point, with edges connecting these nodes based on their proximity determined by a distance measure, such as Euclidean distance, or Cosine Similarity (both of these measures are discussed in Section 3.2.1). Within this domain, the base graph construction can generally be divided into four categories, which are *Delauney Graph* (DG), *Relative Neighbourhood Graph* (RNG), *K-Nearest Neighbour Graph* (KNNG), and *Minimum Spanning Tree* (MST), with the underlying principles of each presenting differing advantages and caveats [55]. Regarding DG, advantages include a high degree of certainty in ensuring precise retrieval results [3], however imply an almost fully connected graph structure in highly dimensional data leading to a large search space [61]. RNG provides some advantage to DG based approaches, through the elimination of redundant neighbours to reduce the number of required distance calculations [3]. KNNG based graphs provide the advantage of limiting the number of neighbours of each node to a specified maximum parameter K, which is beneficial in ensuring to limit memory requirements, however, there is a significant problem given that there is no guarantee of global connectivity in the graph which could make some searches infeasible. Finally, MST based approaches are generally underrepresented in current graph-based ANNs, however, the approach presents the advantage of using the least edges in ensuring global connectivity. This does imply that in some instances, a lack of shortcuts through the graph means that searches may involve a detour through a larger number of nodes [58, 61].

Of the four base graph types, it is DG and RNG which we now focus. The Navigable Small Worlds (NSW) algorithm [58] approximates DG by building an undirected graph through continuous insertion of nodes while ensuring global connectivity. By iterative insertion of graph nodes, long edges are initially formed at the beginning of graph construction, ensuring small world connectivity within the graph. Later insertions to the graph then enable the creation of shorter range edge connections, which ensures search accuracy. This has been demonstrated to enable a high degree of retrieval accuracy, however implies a poly-logarithmic search complexity [66]. In order to improve this poly-logarithmic search complexity, the Hierarchical Navigable Small Worlds (HNSW) algorithm addresses complexity issues through the construction of a hierarchical graph. Graph neighbours are separated into different hierarchical levels based on their proximity, with the granularity of each graph increasing at deeper levels of the hierarchy. This implies that the search process is conducted iteratively through the hierarchical stacked graph, proceeding from the general level to increasingly finer-grained hops with each iteration. This methodology approximates not only DG, as achieved with NSW, but also considers the distribution of node neighbours, thereby approximating RNG. For the study focus of Chapter 4, HNSW was selected as a suitable candidate for approximate nearest-neighbour search and is discussed in greater detail in Section 3.2.2.

2.1.4 High Dimensionality in Embeddings

A caveat of pre-trained language models is the high degree of dimensionality which is prevalent in the embeddings produced by such models. The dimensions for the embeddings produced by BERT-base are 768 [26], with the output of the GPT-3 davinci embedding model being 1536 dimensions [67]. This high degree of dimensionality can ensure that large amounts of information, such as from the context in which a word or sentence is used in an item of text, can be encoded as a vector.

However, as the dimensionality of embeddings continues to grow with larger language models, so does the prevalence of the "Curse of Dimensionality" [68], wherein increasing dimensionality leads to several detrimental effects. First, the volume of computational memory required for storage and processing of highly dimensional vectors is large and grows exponentially. Second, a greater computational complexity is observed in algorithms as the number of dimensions increases [68]. Furthermore, the distance measurements necessary for determining distances between embeddings tend to become meaningless in high-dimensional spaces, wherein the ratio between nearest and farthest points approaches one, such that the points essentially become equidistant from each other [69–71]. This can be represented by the following equation.

$$\lim_{d \to \infty} \frac{\text{MaxDist - MinDist}}{\text{MinDist}} = 0 \tag{2.1}$$

In such cases, the notion of the nearest neighbour of a point is meaningless [72]. As such, downstream tasks which involve calculating distances between embeddings of highly dimensional data can be negatively affected by this curse of dimensionality. One such example of this can be observed in clustering [72], where distance measurements are essential in determining the distance between points needed for assigning them to clusters. The well-known K-means algorithm identifies a predefined number of clusters (determined prior as K) through finding K centroids, and subsequently assigning each point to a cluster associated with the *nearest* centroid, where *nearest* is determined by a distance measurement, such as euclidean distance. Centroids themselves are determined as the mean or median of the points within its cluster. This is summarised in pseudocode in Algorithm 1.

Algorithm 1 K-Means Clustering

```
1: Initialize $k$ centroids randomly
2: while not converged do
3:     Assign each data point to the nearest centroid using a distance function
4:     Update the centroid location as the mean of all assigned data points
5: end while
```

Although much of the algorithmic process of K-means is not relevant to the goal of the thesis, a key point of interest is the *distance function* used on line 3 of the algorithm. As a distance measure, Euclidean distance, and Manhattan distance are the most common [72,73], where Manhattan distance computes the sum of absolute distance if a grid-like path is followed, and Euclidean distance measures the direct distance, calculated based on the sum of the squares, and then finding the square root of that sum. This is analogous to finding the length of a hypotenuse in a triangle. Generally, Euclidean distance is one of the most commonly used distance measurements in clustering [74–76], however, it is adversely affected by the "Curse of Dimensionality" [68,72], which has been proven by [77], where it was identified that Equation 2.1 holds true for Euclidean distance. In addition to this, empirical investigations have shown that this phenomenon appears for dimensionalities greater than 10 [69]. Given that the dimensionalities of language models can be in excess of 1024, it is clear that the curse of dimensionality will have considerable impact upon any applied techniques relevant to this thesis, where clustering techniques are used on embeddings generated by language models - for example, the cluster-based topic modelling techniques discussed in Chapters 5, 6, and 7. Therefore, a chapter has been dedicated to investigating some methods to address this, based on expanding state-of-the-art research into dimensionality reduction techniques, which is detailed in Section 7.6.

2.1.5 Topic Modelling

Due to the predominantly unstructured nature of textual data in the publishing sector, it is essential to explore analysis methods that could use unsupervised learning to address the absence of explicit labelling in the data. Of these, topic modelling is of particular value to the company involved, due to the ability of such techniques to generate new knowledge out of otherwise unstructured data.

Traditionally, topic models were defined by Blei and Lafferty as *"probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts"* [78]. Effectively, this allows for the discovery of patterns of words among documents and for the discovery of information in otherwise unstructured data. Of these, the most well known, Latent Dirichlet Allocation (LDA) [13], works on the assumption that given a collection of documents, K topics are present, with each document expressing topics with different proportions. Such probabilistic approaches are, in essence, bag-of-words techniques, since they model the probabilities of individual words and the likelihood of their appearance within documents and collections of documents [13]. Thus, these techniques ignore implicit semantic and syntactic structure within text. For this Ph.D, it was identified early on that such approaches struggle to leverage valuable information obtained from neural language models, discussed in Sections 5.2, and 5.3.

Recent proposals to address the lack of consideration of context and semantics during topic modelling have focused on a clustering approach [6, 32], wherein dense clusters of embeddings are assumed to represent underlying topics. This enables leveraging of existing techniques for generating embeddings such as those discussed in Section 2.1, to generate a contextual vector representation of the underlying corpus, before applying clustering of embeddings, thus ensuring that clusters representing topics account for contextual semantics in the original texts. This provides not only the advantage of ensuring that any identified topics are semantically and contextually similar, which has been shown to produce highly coherent topics [6], but also permits the downstream analysis of topic and document embeddings in terms of semantic similarity, a practice which is investigated in Chapters 5 and 6. A detailed explanation of this cluster-based topic modelling process is provided in Section 3.2.3, as one of these algorithms, Top2Vec [6], has been a significant focus of the research included in this thesis. However, it is worth establishing some of the advantages of cluster-based topic modelling. Such approaches can ensure the leveraging of neural-embedding models in order to generate embeddings that capture the semantic and syntactic structure of the text, something which would not be possible with probabilistic models. Furthermore, the clustering algorithm used to identify topics from dense areas of embeddings, HDBSCAN [79], provides the advantage of automatic identification of the optimal number of clusters within the data, while when using LDA, it is necessary to define the number of topics to identify, or exhaustively compare across a range of topic numbers to identify an optimal solution. Furthermore, cluster-based topic models are model agnostic and allow the use of any available technique for embedding the text. This ensures that when working with niche domains of text, embedding models can be easily interchanged to evaluate the best approach to analysing the data, such as using embedding models pre-trained on domain-specific literature [80, 81].

A unique necessity of cluster-based topic modelling is encountered through the "*curse of dimensionality*" (detailed in Section 2.1.4), where the heirarchical clustering algorithm HDBSCAN [79] cannot efficiently identify clusters in high-dimensional data [6, 82]. In order to address this, cluster-based topic models first apply a step of dimensionality reduction, commonly using the established state-of-the-art Uniform Manifold Approximation and Projection (UMAP) algorithm [14]. This ensures the reduction of original, high-dimensional embeddings to a low-dimensional representation that can be easily clustered. This dimensionality reduction technique is something which merits a further analysis, to evaluate whether contributions to cluster-based topic models can be facilitated through adapting dimensionality reduction, which is the main focus of the work presented in Chapter 7.

2.1.6 Dimensionality Reduction

Based on the issues previously discussed with the "Curse of dimensionality", the highly-dimensional nature of embeddings produced by language models, which have performed so well in many NLP tasks [26, 31, 33], present difficulties when it comes to applying them in unsupervised learning tasks which involve clustering, and subsequently rely upon distance measures like Euclidean distance. In addressing this, dimensionality reduction techniques have been proposed. Dimensionality reduction can be defined as the transformation of high-dimensional data into a meaningful representation, with reduced dimensionality [83].

Traditionally, linear techniques, such as Principal Component Analysis [84] have been applied to DR, however, these were found to be inadequate when applied to complex, non-linear data. The supervised alternative, Linear Discriminant Analysis, involves a generalisation of Fischer's linear discriminant [85], seeking to identify linear combinations of features as a means to characterise or separate objects, or documents.

More recently, t-distributed stochastic neighbour embedding [86] was proposed, as a nonlinear means of dimensionality reduction for visualisation purposes. t-SNE was based upon Stochastic Neighbour Embedding [87], wherein a Gaussian is centred on high-dimensional objects, ensuring that a probability distribution may be defined over potential neighbours of the object. t-SNE expanded upon this, through the implementation of a Student-t distribution in place of a Gaussian, when computing the similarity between points in low-dimensional space. In this method, reduction was typically performed to a dimensionality of 2 or 3, with the resulting vectors being applied as coordinate points in the visualisation. One issue with this is that t-SNE observed a significant decrease in performance as dimensionality increased, and as such was generally best applied in the generation of low-dimensional vectors of 2-3 dimensions [88]. Furthermore, there is a significant limitation due to the computational complexity of the algorithm which scales at a degree of $O(N^2)$, where N is the number of data points [89]. The application of the Barnes-Hut algorithm as an approximation method for the gradient calculation algorithm can improve efficiency to $O(\log N)$ time complexity, as demonstrated in [90, 91]. However, it is important to note that this approach is applicable only when the output dimensionality is less than or equal to three dimensions.

An extension to t-SNE is presented in [92], where a parametric dimensionality reduction technique is introduced, based on the strategies employed in t-SNE. This works upon the assumption that a neural network possessing sufficient hidden layers is capable of achieving an approximation of the non-linear functions employed by t-SNE, when mapping a high-dimensional representation to a lower-dimensional representation. In this work, the authors discuss that directly training a neural network through backpropogation is not feasible, due to the tendency for backpropogation to encounter a local minimum, given the complex interactions between layers in the network, which entail a large number of parameters. To address this, the authors applied a training strategy involving the training of autoencoders based upon Restricted Boltzmann Machines (RBMS). In this process, a stack of RBMs is trained, and then used to generate a pre-trained feed-forward network, which can subsequently be fine-tuned using backpropogation. The resulting network represents an approximation of the functions of t-SNE. The work demonstrates through experimentation that the parametric model can outperform PCA and an autoencoder in the reduction of the dimensionality of the MNIST [93] and 20 Newsgroup datasets [94]. It is worth noting that this local minimum phenomenon is something that is considered in Chapter 7, Section 7.6, where, in order to address this, additional residual connections are integrated into the transformer-encoder employed for parametric dimensionality reduction, as outlined in RQ3.6.

Uniform Manifold Approximation and Projection

In recent years, Uniform Manifold Approximation and Projection [14], has demonstrated a significant improvement to t-SNE, through an improvement in scalability, making it more accessible for use in data transformation pipelines. UMAP sought to better represent the local structure of the original data, while preserving the global structure. Similarly to other dimensionality reduction techniques, the algorithm serves as a suitable tool for visualisation of high-dimensional data; however, it has been demonstrated to be an efficient tool for general purpose dimensionality reduction for use in machine learning, with applications including topic modelling [16], text clustering [95], and genetics research [96, 97]. In the case of topic-modelling research, this is of particular value to the research conducted throughout the Ph.D project, where cluster-based topic modelling, largely based around the Top2Vec [16] algorithm, involves the reduction of highly dimensional document embeddings to a low-dimensional representation using UMAP in order to efficiently identify clusters. Thus, it is necessary to introduce a sound foundation of UMAP, which is presented in Section 3.2.4. Currently, UMAP, along with its derivatives, represents the state-of-the-art in dimensionality reduction, substantiated by several key factors best demonstrated by comparing with its closes competitor, *t*-SNE. Firstly, UMAP exhibits significant scalability advantages over *t*-SNE [14, 89], thereby enabling the efficient processing of large datasets. Additionally, UMAP operates without encountering computational constraints related to the embedding dimension, a limitation encoutered with *t*-SNE [14]. Lastly, empirical evidence suggests that UMAP more effectively preserves the global structure of original data in comparison to *t*-SNE, however this has been highlighted as being an artefact of UMAP initialisation, with *t*-SNE performing just as well when the same initialisation strategy was used [98]. Given these advantages, it is evident why UMAP has been extensively embraced as a method for diverse dimensionality reduction tasks, including the enhancement of clustering processes [95], visualization of genetic data [97], and applications in molecular biology [99], among others.

Parametric UMAP

A subsequent extension of UMAP has been presented through parametric UMAP [4]. While UMAP performed optimisation of the low-dimensional representation using stochastic gradient descent, parametric UMAP introduced a neural network in its place, which learns a parametric relationship between the original high-dimensional data and the embedding [4]. This provided a significant improvement in the speed of inference of new embeddings, once the parametric model has been trained. The authors also analysed the performance of parametric UMAP in clustering tasks, through the evaluation of normalised mutual information (NMI) in k-Means clustering, with findings indicating this to be comparable to UMAP. The introduction of a neural network in parametric UMAP presents a novel avenue for investigation; Namely, the hypothesis that neural network architecture may have an impact upon the outcomes of dimensionality reduction. This is something that is addressed as part of **RQ3**, where investigations into the introduction of a transformer encoder architecture demonstrate an improvement in downstream clustering tasks. This is presented in detail in Chapter 7.

CHAPTER 3

Technical Background and Methodology

After the highlighting of the general research questions and objectives of the project in Chapter 1, presenting a generalised overview of the state of the literature at the outset of the project, and the general themes covered in Chapter 2, it is necessary to provide an overview and explanation of the general research methodologies necessary to address the research questions and meet the outlined objectives. This is formed of a general discussion of each general research question, and the reasoning behind it in Section 3.1. Following this, an introduction to the methodological approaches used throughout the project is performed in Section 3.2. These can be broken down into the methods used for the collection of any data used in experiments, which was not previously available, the framing of the general algorithmic techniques which are utilised throughout the project, and any methods which were necessary for the evaluation of findings from the studies conducted throughout the project.

3.1 Research Outline

Generally, the methodology for investigating how to contribute to the media and publishing industry was made up of three main aspects, which are detailed as *influ-*

encer retrieval, *literature analysis through topic modelling*, and *advancing research in dimensionality reduction techniques, which can directly contribute to improving topic models*. In the initial stages of the project, the prevalence of unstructured and largely unlabelled data implied the necessity for investigation of techniques which can handle unlabelled data. Moreover, if the outcomes of the project were to have any lasting commercial impact, the techniques needed to demonstrate a robustness and ability to handle new data without extensive human intervention or retraining. Therefore, unsupervised learning techniques, in which algorithms seek to learn patterns exclusively from unlabelled data [100], were the best suited. In Chapter 4 the focus is upon semantic search techniques, which serve as grounds for investigation into the construction of meta-embeddings from multiple sources, and how this affects retrieval accuracy. Subsequently, in Chapters 5, and 6 the focus of the research is adapted to applications of topic modelling algorithms, to identify how these can contribute to the identification of topics within literature, and thus give rise to the provision of an automated literature analysis framework. These chapters take into consideration the findings of the previous Chapter 4, by investigating the analysis of semantic similarity of the identified topics using distance measures first introduced in Chapter 4. Following this, in Chapter 7, a key step in the process for cluster-based topic models is investigated through parametric UMAP [4], where it is identified that the introduction of the transformer-encoder [5] architecture can contribute to better quality topic models through enhancing dimensionality reduction.

3.1.1 Addressing GRQ1: *How can journalistic influencers be retrieved with accuracy for a publishing enterprise?*

In addressing this general research question, the following specific questions were defined:

- **RQ1.1:** *How may information retrieval techniques be adapted to facilitate the identification of relevant social media accounts for a given article of journalistic content, based on a user's posted content?*

- **RQ1.2:** *How does the hybridisation of tweet and user biography embeddings contribute to recommendation quality in the retrieval of social media accounts based on full-text querying?*

- **RQ1.3:** *Can forward time-decay functions be applied to recommendation ranking to improve the temporal relevance of results, and how does this affect the quality of recommendations?*

These in-depth research questions are discussed in detail in Chapter 4; here, a high-level explanation follows. Regarding RQ1.1, an information retrieval approach is defined, to identify relevant social media (influencer) accounts. By taking an information retrieval approach using zero-shot learning based on vector retrieval, it is possible to mitigate the "cold start problem", which is encountered in many collaborative filtering recommendation systems, wherein a lack of initial training data will mean that initial recommendations are poor quality for new users or data [101]. Further to this, a vector retrieval approach ensures to address the criteria set forward by the industry sponsor, as a vector representation of the full-text of an article, computed using language embedding techniques, can be used for retrieval. During research on semantic search, the phenomenon known as the "curse of dimensionality" was encountered, which is introduced in Section 2.1.4. This became a significant aspect of investigation in later chapters, where the effects on cluster-based topic modelling methods became apparent. In this case, this was caused by the high degree of dimensionality of embeddings used in the vector search, meaning that brute-force retrieval using cosine similarity would be computationally expensive for a large dataset.

In investigating methods to improve the accuracy of recommendations, it was proposed to explore whether a user's posted content could augment retrieval based on their bio. This gave rise to the conjecture that merging embeddings from diverse data sources, such as a user's bio and their posted content via a "meta-embedding" technique [36, 102], could directly boost the accuracy of influencer results retrieved. This concept is encapsulated in RQ1.2, with an investigation into the hybridisation of tweet and biography embeddings, and an assessment into how the bias in tweet-bio weighting can impact the accuracy of recommendations.

Lastly, in RQ1.3, the application of time-decay functions is explored, to give preference to "recency" in recommendations, a criterion specifically outlined during discussions with the industry sponsor. In this context, it is preferable to consider the timing of when recommended users have engaged in a semantically similar topic, and bias this towards more recent recommendations. Similarly to RQ1.2, this aspect is examined to ascertain how "influential" the time-decay function should be, and how this affects the recommendation results in terms of accuracy.

Overall, this aspect of research, which was conducted in the early stages of the Ph.D., investigated the use of zero-shot learning through vector retrieval at its core. Through the investigation of meta-embeddings and time-decay functions, improvements in the accuracy of recommendations were demonstrated. This has since led to a **commercially available product: https://reveela.com/**, which is produced by the industry partner, for the recommendation of semantically similar social media influencers and journalists based on a full-text item of journalistic content.

3.1.2 Addressing GRQ2: *What AI techniques can be defined to assist readers in the comprehension of literature?*

In addressing this general research question, the following specific questions were defined:

- **RQ2.1:** *How can Existing Topic Modelling Technologies be adapted to facilitate a semi-automated analysis of academic literature?*

- **RQ2.2:** *How can the semantic relationships between topics in literature contribute to an understanding of topics during exploratory literature analysis?*

- **RQ2.3:** *How can analysing the temporal trends of identified topics in literature contribute to an understanding of topics during exploratory literature analysis?*

- **RQ2.4:** *Can rule-based extraction of named biomedical resources contribute to the comprehension of automatically generated literature topics?*

- **RQ2.5:** *Can an end-to-end framework, built upon cluster-based topic modelling confer a benefit to researchers through assisting the literature review process?*

- **RQ2.6:** *What are the effects on cognitive load, technology acceptance, and general acceptance of users when using a software implementation based on our framework?*

- **RQ2.7:** *How does the proposed literature analysis framework and software implementation influence the accuracy of the identified results in practical use cases for medical researchers?*

To further tackle the constraints associated with unstructured and unlabelled data in the publishing sector, it was considered essential to explore the use of unsupervised learning techniques to uncover significant connections within unlabelled data. Textual media is continuously produced in the publishing industry, and key players within the industry must maintain an up-to-date understanding of trends and relationships within topics. This formed the basis of the second overarching research question, where topic modelling was proposed as a method to assist in the comprehension of large volumes of textual content. This was carried out with a focus on academic literature, which is easily accessible and enables access to a large amount of data which may be accessed through API resources. To address this research question, a two-pronged approach was adopted involving an initial exploratory phase outlined in Chapter 5, and a subsequent experimental investigation based on the proposal of a definitive method for the comprehension of large volumes of academic literature in Chapter 6. Restrictions to the sharing of Intellectual Property owned by the industry sponsor led to the necessity to adopt a case-study approach which does not reveal sensitive industry data, and for this, a case-study into the topic modelling of academic literature was devised.

This channel of research, encapsulated by addressing RQ2, commences with an exploration of the probabilistic topic model, Latent Dirichlet Allocation through a case-study analysis of capabilities of using the algorithm for supporting the findings of a manual literature review, detailed in Chapter 5. After establishing the caveats of probabilistic topic modelling in this domain, a further case-study evaluates the alternative technique of cluster-based topic modelling through the Top2Vec algorithm, and compares this with Latent Dirichlet Allocation in terms of quantitative evaluation of established measures of topic quality, and a qualitative analysis of the topics, which shows that Top2Vec presents many advantages in the applied task of literature analysis.

After establishing the value of cluster-based topic modelling for literature analysis, a framework for the automated analysis of litertaure is proposed in Chapter 6, which is evaluated first through a preliminary case-study conducted with biomedical experts on their domain of expertise. Following this, a wide-scale study is conducted through a participant study of medical experts into their experience when using the literature analysis framework with respect to the established concepts of cognitive load, technology acceptance, and general satisfaction.

Contributions from this aspect of research can subsequently be transferred into the publishing domain by focusing upon the wealth of textual data available in the domain, such as news publications, magazine articles, and social media streams, where the findings of this research could directly enable knowledge discovery, thus ensuring to expedite many processes within the publishing sector.

3.1.3 Addressing GRQ3: *How can dimensionality reduction be used to improve the accuracy of text clustering and, subsequently, topic modelling?*

In addressing this general research question, the following specific questions were defined:

- **RQ3.1:** *How do current dimensionality reduction algorithms affect accuracy in text clustering tasks?*

- **RQ3.2:** *How does output dimensionality of dimensionality reduction models influence performance in text clustering tasks?*

- **RQ3.3:** *Can small portions of labelled data used in the metric learning of* *dimensionality reduction contribute to improvements in downstream clustering accuracy?*

- **RQ3.4:** *Can the introduction of attention mechanisms within neural networks (further) improve the metric learning of dimensionality reduction algorithms, in terms of clustering accuracy?*

- **RQ3.5:** *Can the transformer-encoder neural network positively influence downstream clustering when applied to cluster-based topic modelling tasks?*

It was identified during the course of research into topic modelling for the thesis that an essential process in cluster-based topic modelling is prior reduction in the dimensionality of embeddings, in order to permit efficient clustering. Therefore, a novel research paradigm is proposed as answer to RQ3, addressed in Chapter 7, to investigate how dimensionality reduction techniques could directly improve cluster-based topic modelling within the domain. This was based on the recently proposed parametric UMAP algorithm [4], which enables the specification of deep-learning neural-networks for the task of dimensionality reduction. Regarding this avenue of research, it was first necessary to establish the implications of existing dimensionality reduction algorithms upon downstream clustering. To achieve this, four benchmark algorithms were selected, and evaluated with respect to the effects that the specification of output dimensionality has upon downstream clustering with the k-Means algorithm. The transformer-encoder architecture, known for its significant impact in NLP, was then integrated into a metric-learning pipeline with the parametric UMAP algorithm, resulting in improved clustering accuracy compared to state-of-the-art dimensionality reduction algorithms.

Although not directly related to cluster-based topic models, the focus of RQs3.1-3.4 was undertaken for several reasons, as investigated in Section 7.5. By first approaching a comparison of DR, by assessing clustering accuracy, it becomes possible to demonstrate direct quantitative evaluation of various dimensionality reduction techniques, by leveraging established benchmark datasets. The selection of k-Means, in this context, is justified by its provision to pre-define the number of clusters to be extracted, corresponding to the known count of unique labels within the evaluation dataset. Conversely, in cluster-based topic modeling, employing HDBSCAN for clustering implies an absence of assurance regarding the identification of the correct number of clusters, alongside the likelihood of a substantial portion of the data being categorised as outliers and thus excluded from the evaluation. Thus, the aim of the initial evaluation of parametric dimensionality reduction in Section 7.5 is to compare existing dimensionality reduction algorithms across various output dimensionalities, and then introduce the proposed modified pipelines through the adoption of a transformer-encoder architecture, to ensure a precise comparison.

After establishing the value of the architecture in a metric-learning methodology using accuracy measures, it remains necessary to consider whether the transformer-encoder could also directly benefit cluster-based topic modelling. Thus, in Chapter 7, Section 7.6, an adaptation to the Top2Vec algorithm is made, to facilitate the introduction of the transformer-encoder through parametric UMAP. To quantify the influence that this adaptation has upon the quality of the topic modelling solution, an evaluation is conducted from both the perspective of established evaluation metrics, and also the general human interpretation of the topics produced. During the literature evaluation of RQ3, it was identified that the introduction of residual connections in the transformer network confers a benefit in the training of deep learning architectures. Thus, a contribution to the transformer-encoder for parametric DR is proposed through the introduction of further residual connections, with the results of these demonstrating a further benefit to the task of topic modelling than that of the conventional transformer-encoder.

3.2 Methodological Approaches

3.2.1 Distance Measures of Semantic Similarity

The concept of measuring similarity between text embeddings is an essential consideration throughout the research discussed in this thesis, with Chapters 4, 5, 6 and 7 all entailing this strategy to some degree. The general objective of measuring similarity of embeddings is to identify semantically similar instances of text, which presents value in tasks such as information retrieval and clustering. Moreover, measuring semantic similarity presents the opportunity to identify relationships between texts which would otherwise be unstructured, a concept which is evaluated in Chapters 5 and 6.

An effective measure of similarity between embeddings is the *cosine similarity.* Cosine similarity measures the angle between the cosines of vectors, computed based on the dot product of vectors, and dividing this by the product of their magnitudes. This can be summarised in the following equation, given two vectors a and b:

$$Cos_{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||} \tag{3.1}$$

Where $||\mathbf{a}||$, and $||\mathbf{b}||$ correspond to the respective lengths of each vector. Cosine similarity is between the bounds of -1 and 1, where 1 indicates an angle of 0 such that the vectors are as close as they can be, indicating a high similarity between vectors.

3.2.2 Information Retrieval via Hierarchical Navigable Small Worlds

As detailed in the general related works in Section 2.1.3, data retrieval by exhaustive comparison of semantic similarity, a brute-force search, is a computationally expensive method of information retrieval. Thus, it becomes impracticable as a search strategy when applied to large datasets, such as those used in industry. To address this limitation, Chapter 4 explores a framework for retrieving social media influencers using journalistic text queries based on meta-embeddings, which depends on a potentially expanding dataset of $390,750$ social media posts as new data is collected in industry. To facilitate this, an Approximate-Nearest-Neighbour (ANN) algorithmic approach based on Hierarchical Navigable Small Worlds (HNSW) is adopted. No modifications of the underlying algorithm are performed in this thesis, however it is necessary to provide an overview of the algorithm.

HNSW is a graph-based ANN which, compared to the polylogarithmic complexity of precursor Navigable Small Worlds (NSW) algorithms [103–105], provides a logarithmic complexity scaling [3]. It was for this reason that HNSW was selected as the method for semantic search in Chapter 4.

HNSW is based on the principle of a NSW, where data points are constructed within a graph structure, such that every node can be reached within a small number of "hops" from any other node. HNSW adopts a hierarchical graph structure, where data points are linked based on their proximity, or distance, to other points, starting with a base layer representing a graph of the entire dataset composed of short-distance links, with each higher layer of the graph consisting of longer-range connections between points. By constructing a graph in this manner, it is possible, during the search stage, to navigate through the hierarchical graph layers starting from the top layer, where links between graph nodes entail the largest degree of proximity in the data, and descending through each layer in order. This can be best represented in the diagram of Figure 3.1.

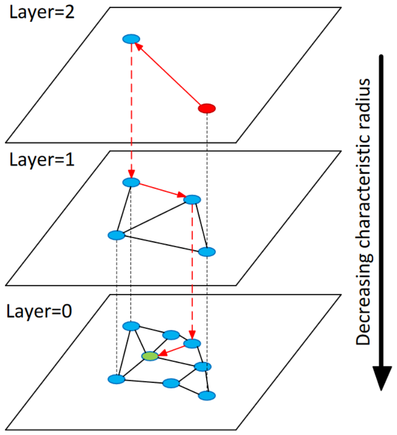


Figure 3.1: A Basic overview of the graph construction of HNSW, taken from [3].

HNSW presents many advantages, such as the logarithmic complexity scaling, which makes it well suited for large, datasets. Similarly, HNSW has been demonstrated to handle highly dimensional data well, which makes it suited for searches of highly dimensional text embeddings. However, the algorithm is not designed with the consideration of dynamic datasets, such as the industry application, which was the result of the research presented in Chapter 4. While it is possible to modify a HNSW index, it is reported in the literature that frequent modification of HNSW indexes through additions and deletions can lead to some data points becoming unreachachable - the '*unreachable points phenomenon*' [106]. Thus, it is necessary to recompute a new HNSW index as new data is added or deleted. This recomputation scales logarithmically in terms of complexity with regard to the size of the index, as it is effectively the creation of a new index. Thus, within an industrial context, such recomputation is permissible provided it is not undertaken with excessive frequency. To address this, in the industry application of the research presented in Chapter 4, recomputation of the index is performed in a batched manner on a weekly basis.

3.2.3 Cluster-Based Topic Modelling

Cluster-based topic modelling through the Top2Vec algorithm [6] is an essential technique adopted for the analysis of academic literature, which is the main focus of Chapters 5 and 6. Thus, it is necessary to provide a sound understanding of the concepts involved.

Essentially, Top2Vec is built upon the idea of semantic space, which is a spatial representation in which distance represents semantic association [107]. Semantic embeddings of words and sentences, derived in the manner discussed in Section 2.1 have been shown to capture the semantic regularities of language [1,108]. By taking the assumption that semantic embeddings can be measured using vector distance measures, Top2Vec performs the computation of a continuous representation of the semantics of the data by mapping document and word embeddings into a vector distribution. In the original work proposing Top2Vec, this is performed using Doc2Vec DBOW model [25]. After the computation of a matrix of jointly embedded document and word vectors, the words that are nearest to a document vector are the most semantically descriptive of the underlying topic of that document. Dense areas of document embeddings can then be interpreted as semantically similar areas, with the centroid of these dense areas being represented as a vector of the overall topic of the dense cluster of documents.

To identify dense clusters of documents within the distribution, Top2Vec employs density-based clustering of document vectors through Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [79]. The unique advantage of this density-based clustering of document vectors is made apparent through the fact that HDBSCAN enables the automatic identification of the number of clusters within the data, without the need for prior specification of a hyperparameter detailing how many clusters to identify. Furthermore, HDBSCAN labels sparse areas of data as noise, which is valuable in ensuring to only identify the dense clusters of document vectors as representing clusters of semantically similar topics. A limitation of applying HDBSCAN, however, is made prevalent through the "curse of dimensionality", discussed in Section 2.1.4. Thus, to enable effective clustering with HDBSCAN, the UMAP [14] algorithm must be applied, to produce a lower-dimensional representation of the original document vectors. The details of UMAP are discussed in Section 3.2.4, as chapters within this thesis will analyse contributions to this dimensionality reduction method.

After the computation of a low-dimensional representation of document embeddings and the subsequent clustering of these with HDBSCAN, the dense clusters of embeddings identified can be seen as representing topics of semantically similar documents. A topic vector is then defined as being the centroid of a dense cluster of documents. Subsequently, topic words can be identified as being the nearest word vectors to the topic vector. This technique of cluster-based topic modelling using Top2Vec is applied at several stages throughout this thesis, namely, Chapters 5, 6 and 7. However, in Chapter 7, an adaptation to the algorithm is proposed through the introduction of parametric UMAP [4] and a transformer-encoder based pipeline, to evaluate whether this contributes to improvements in the general quality of the topics produced by modified Top2Vec.

3.2.4 Dimensionality Reduction via Uniform Manifold Approximation and Projection

As detailed in Section 3.2.3, an essential technique in cluster-based topic modelling with Top2Vec [6] is the dimensionality reduction of embeddings to permit efficient clustering with HDBSCAN [79]. This is something which is considered in Chapter 7, where it is proposed to enhance the topic modelling procedure through the introduction of a parametric DR pipeline by exploring the architectural design of neural networks. Thus, while the research presented in this thesis does not directly modify UMAP [14], and instead focuses upon parametric UMAP [4], it remains essential to establish a sound understanding of UMAP.

The task of dimensionality reduction of high-dimensional embeddings with UMAP can be summarised as the following: Given a D-dimensional data set $X \in R^D$, produce a d-dimensional embedding $Z \in R^d$ such that points that are close together in X (e.g., x_i and x_j) are also close together in Z (z_i and z_j) [4,14].

Functionally, UMAP is a manifold learning technique based on Reimannian geometry and algebraic topology. UMAP performs two key steps in the computation of low-dimensionality vectors: First, the computation of a graph representation of data; and second, the optimisation of a low-dimensionality representation of the graph through stochastic gradient descent.

Graph Construction

In the first stage, UMAP performs the construction of a fuzzy simplicial complex (a topological representation of a local neighbourhood graph), containing a weighted graph where edge weightings represent the probability P that two points in X are connected [14]. Probabilities are first calculated as local, one-way probabilities $\left(P^{\text{UMAP}}_{i|j}\right)$ from a point to its neighbours, and then later symmetrised to create a final probability representation of the relationship between pairs of points. Local, one-directional probabilities are computed between a point and its neighbours to determine the probability that an edge (referred to as a simplex) exists, using the assumption that data is uniformly distributed across a manifold in a warped dataspace. Based on this assumption, the local concept of distance is determined by the distance to the k^{th} closest neighbour, and the local probability is adjusted based on this local distance measure:

$$p^{\text{UMAP}}_{j|i} = \exp\big(-\big(\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j) - p_i\big)/\sigma_i\big) \tag{3.2}$$

The variable p_i represents a local connectivity parameter that is defined as the distance from x_i to its closest neighbour, while σ_i is another local connectivity parameter that is adjusted to the local distance surrounding x_i based on its k nearest neighbours.

Following the computation of one-directional edge probabilities for each data-point, UMAP computes a global probability as the probability of either of the two local, one-directional probabilities occurring:

$$p_{ji}^{\mathrm{UMAP}} = \big(p_{j|i} + p_{i|j}\big) - p_{j|i}p_{i|j} \tag{3.3}$$

Graph Embedding

After the construction of a distribution of probabilistically weighted edges between points, an embedding is initialised corresponding to each data point, where a probability distribution (Q) is computed between points as was done with the distribution (P) in the input space. The objective of UMAP is subsequently to optimise this embedding to minimise the difference between P and Q.

Pairwise probabilities are computed directly, without first computing local, one-directional probabilities. UMAP computes pairwise probability q_{ij}^{UMAP} as:

$$q_{ij}^{\mathrm{UMAP}} = \left(\left(1 + a\, \|, z_i, -, z_j, (|\,|)^{2b}\right)\right)^{-1}, \tag{3.4}$$

Where a and b are hyperparameters which can be set based on the desired minimum distance between points in embedding space. It is worth noting that this probability distribution in embedding space is not normalised. UMAP then performs an optimisation to minimise the differences between Q and P via a cost function based on cross-entropy, optimised using gradient descent:

$$C_{\mathrm{UMAP}} = \sum_{i \neq j} p_{ij} log \left(\frac{p_{ij}}{q_{ij}}\right) + (1, -, p_{ij})\, log \left(\frac{1 - p_{ij}}{1 - q_{ij}}\right). \tag{3.5}$$

Attraction and Repulsion

Attempting to minimise the cost function of Equation 3.5 over every pair of points in the dataset would be computationally expensive in UMAP. Thus, UMAP assumes that the cost function can be broken down into a mixture of attractive forces between locally connected embeddings, and repulsive forces between nonlocally connected embeddings.

With regard to attractive forces, UMAP utilises an approximate nearest neighbour graph, based on the assumption that elements that are further apart in the data space will have very small edge probabilities, and can be treated as being zero. Thus, edge probabilities and attractive forces need only be calculated for the nearest neighbours of a point, with non-nearest neighbours being treated as having an edge probability of zero. As brute-force nearest neighbour graphs are computationally expensive, UMAP applies an approximate nearest neighbour graph at this stage, which may be any K-Nearest Neighbour Graph algorithm, with UMAP applying NN-Descent for this task [109].

As most of the data points will not be locally connected, UMAP assumes that it is not necessary to compute the cost function over most pairs of embeddings. To achieve this, UMAP performs negative sampling over embeddings, by, during training, iterating over positive, locally connected edges, and randomly samples edges from the remainder of the data, treating their edge probabilities as zero when computing cross-entropy. As most data points are not locally connected and will have a low edge probability, the negative samples are usually correct which permits UMAP to sample only sparsely over edges in the data.

In summary, UMAP relies upon the construction of a graph and subsequent embedding that preserves the structure of the graph. This is achieved by learning an embedding by minimising cross-entropy sampled over positively weighted edges, and negative sampling randomly over the dataset to minimise over sampled batches of the dataset. The explanation of the algorithmic properties of UMAP was based upon the main works discussing them [4, 14].

3.2.5 Parametric Uniform Manifold Approximation and Projection

As discussed in Section 2.1.6, a recent advancement based upon UMAP has been presented in the Parametric UMAP [14] algorithm. In this work, the task of DR is approached from the perspective of deep-learning, generally based upon the assumption that a parametric relationship can be learned between the original high-dimensional data, and a low-dimensional representation.

In this thesis work, UMAP loss is applied as a regularisation for stochastic gra-

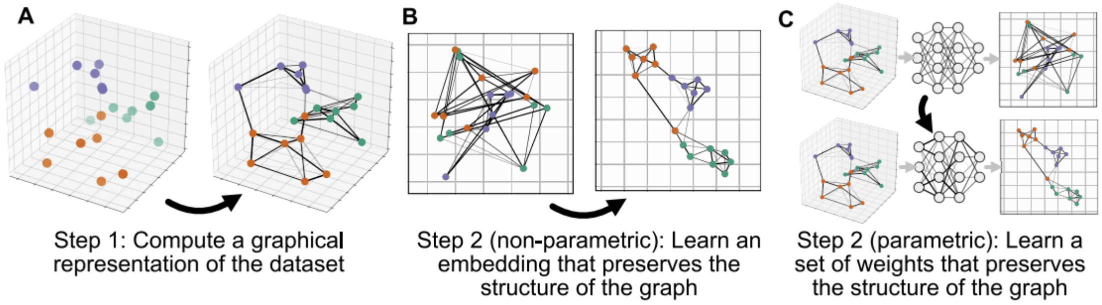

A
B
C
Step 1: Compute a graphical representation of the dataset
Step 2 (non-parametric): Learn an embedding that preserves the structure of the graph
Step 2 (parametric): Learn a set of weights that preserves the structure of the graph

Figure 3.2: An overview of UMAP ($A \rightarrow B$) and Parametric UMAP ($A \rightarrow C$), taken from [4].

dient descent in deep learning, enabling the introduction of an encoder network, which optimises deep-learning model weights based on UMAP loss while producing a low-dimensional representation of the data that maintains the overall structure of the graph. As presented in Figure 3.2, this differs from nonparametric UMAP by optimising the parameters of a neural network, in order to produce a low-dimensional representation that preserves the structure of the original graph representation of the dataset. It has been demonstrated that this parametric approach to DR can produce embeddings that are comparable to those produced by nonparametric UMAP [4], however, most notably, the introduction of a neural-network enables the possibility of considering the influence that architectural design of the encoder network can have upon the quality of DR. This is the main focus of the research outlined in Chapter 7.

3.2.6 Transformer Networks

The transformer network architecture has had a profound impact on the field of NLP since its proposal by Vaswani et al. [5]. Originally proposed as a novel architecture which can address the vanishing gradient problem, which becomes prevalent when modelling long sequences, the transformer makes use of novel (at the time of its inception) multi-head self-attention mechanisms within stacked transformer blocks. It is valuable to consider the architectural design of the transformer for two reasons. Firstly, a number of the embedding techniques used within the contributions of this thesis are built on transformer networks, such as BERT [40], MPNET [41], and

RoBERTa [31]. Furthermore, in Chapter 7, the encoder block of this architecture is considered as a method for enhancing the quality of dimensionality reduction techniques in order to directly contribute to topic modelling algorithms.

At its core, the transformer consists of an encoder block and decoder block, with the encoder block taking an input sequence and outputting a matrix representation of the input, the decoder block then takes the matrix representation output of the encoder as an input and produces an output sequence.

Within both the encoder and decoder network, a transformer block is itself made up of two distinct sub-layers. The first of these implements a multi-head self-attention mechanism, with the second consisting of a position-wise fully-connected feed-forward network. Between each of the sub-layers of this architecture, residual connections [17] are implemented, which have been demonstrated to assist in addressing the vanishing gradient problem, where during backpropogation, the multiplication of the gradients of each layer, if they are smaller than 1, leads to exponentially decreasing gradients. It is reported that as the sequence length of a model increases, the magnitude of the gradient typically decreases, which can slow or even stop the training process [18]. Notably, the consideration of residual connections and their implications on the problem of vanishing gradients with respect to long sequences is made in Chapter 7 when investigating dimensionality reduction for clustering and topic modelling. A general overview of the transformer architecture is presented in Figure 3.3.

Multi-Head Self-Attention

A notable contribution of the transformer is the proposal of multi-head self-attention, based on scaled dot-product attention mechanisms. Attention can be generally described as the mapping of a query, and a set of key-value pairs to an output, where all of these parameters are vectors. In essence, attention mechanisms allow a model to focus on specific aspects of an input sequence. In the field of NLP, this could be an input sentence, where the model may focus upon specific, important words. Given an input vector sequence of input vectors $\mathbf{X} = [x_1...x_n]$, self-attention projects each input vector x_i to a query vector q_i, and all other input vectors in X to produce

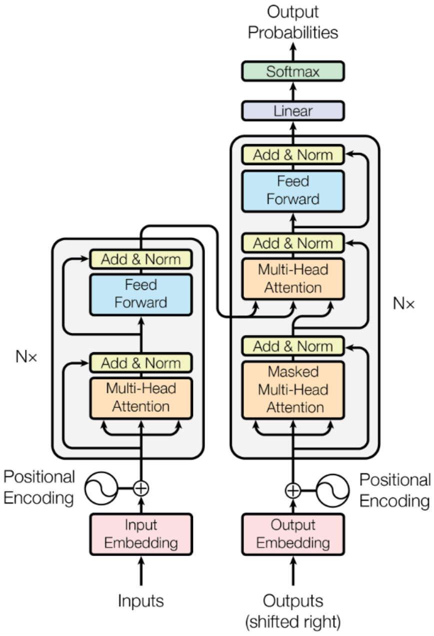


Figure 3.3: The overall architecture of the transformer network, taken from the original work [5].

attention weights. These weights then show how relevant each input vector in X is with respect to x_i. The resulting output is then computed as the weighted sum of these values.

For the original transformer network, a variant of attention named scaled dot-product attention was proposed, which entails the computation of the dot products of a query vector Q, with all keys K, with a dimension of d_k and the subsequent division of each by $\sqrt{d_k}$ and subsequent application of a softmax function to obtain the weights on the values V. This can be summarised mathematically by Equation 3.6.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3.6}$$

This scaled dot-product attention function is implemented within a further proposed multi-head attention. The advantage of this implementation was to ensure that the proposed model could jointly attend to information from different representation subspaces at different positions at the same time. Multi-head attention considers the linear projection of queries Q, keys K, and values V multiple (h) times, with the attention function performed in parallel. This is summarised mathematically in Equation 3.4, and visually in Figure 3.4. With regard to Equation 3.4, QW_i^Q, KW_i^K, VW_i^V, are projection matrices of individual attention heads corresponding to queries, keys, and values, and W^O is a final projection matrix for the whole multi-head attention head.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \\ \text{where head}_\text{i} &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3.7}$$

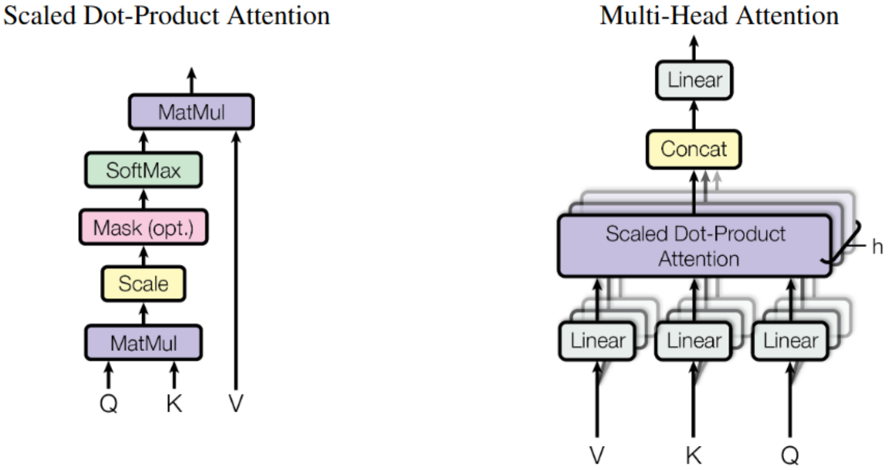


Figure 3.4: Scaled dot-product attention (left) and multi-head attention (right), taken from [5].

Residual Connections

While the introduction of residual connections [17], also known as layer addition, into the transformer architecture is little discussed in the original work [5], it is worth considering the implications of this architectural design choice. Residual connections were originally considered as a method to address the difficulty of training deep neural networks in the domain of computer vision, where a degradation in accuracy is sometimes observed as the depth of a network increases [17]. By the introduction of residual connections, in essence allowing reference to output sequences from higher in the network architectures, it was identified that the training and validation error was significantly reduced, even in networks with up to 34 layers. In the transformer, residual connections are employed within both the encoder, and decoder blocks following both of the distinct sub-layers of multi-head attention, and feed-forward network. For this, the output of each sub-layer is summed with the original input sequence, such that $LayerNorm(x+Sublayer(x))$, where $Sublayer(x)$ is the function for each sub-layer. This consideration is further expanded in Chapter 7, where a proposed adoption of additional residual connections is considered to contribute to the dimensionality reduction process of topic models.

3.3 Overview of Datasets

In order to clarify the datasets utilised in this thesis, the following table provides an overview of the significant datasets used in research throughout the project. This outlines their contextual applications and references the works where they may be found.

Table 3.1: Datasets used in this thesis, their associated task and relevant publications.

Datset Name	Chapter	Task	Relevant Works
Twitter Influencers	4	Information Retrieval	[110]
Arabic Sentiment Analysis Publications	5	Topic Modelling	
Intelligent Tutoring Systems Publications	5	Topic Modelling	
Human Kinome Publications	6	Topic Modelling	
Human Kinome Dataset	6	Publication Querying	[111, 112]
20 Newsgroups	7	Text Clustering & Topic Modelling	[94]
TREC	7	Text Clustering	[113]
AG News	7	Text Clustering	See Footnote In-text.
BBC News	7	Text Clustering & Topic Modelling	[114]

The datasets discussed can be generally attributed to three main tasks, which are information retrieval, topic modelling, and text clustering. Regarding information retrieval, it was necessary when addressing the General Research Question (GRQ 1) to perform the collection and curation of a suitable dataset of social media influencer accounts for the Twitter[1], which would be used in an information retrieval based-approach for an investigation into the provision of an influencer recommender engine based on full text querying. Given a lack of suitable existing data for this task, it was necessary to curate one using API resources. Details to the collection of data for this process are detailed in Chapter 4. Due to intellectual property restrictions, and the subsequent deployment of the collected data in a commercial setting, this dataset has not been published and is stored internally with the industrial sponsor of the project, Distinctive Publishing[2].

For the task of topic modelling, several different data sources were employed when evaluating topic models, and subsequently providing case-studies into the application of topic modelling methodologies. In Chapters 5, and 6, case-studies were conducted through the topic modelling of academic publications within niche academic domains. Data, in the form of academic publications and their associated metadata, was obtained for these tasks through the querying of publicly available API resources of major publishers and repositories. In the case-studies of topic-modelling of academic literature in Chapter 5, publications were extracted from the CrossRef API[3], Elsevier API[4], Springer API[5], Wiley API[6], Core API[7] and ArXiv preprints repository[8]. In the case-study of topic modelling of academic literature in the medical domain, Chapter 6, all literature used in the analysis was collected from PubMed API tools [115].

Outside of the case-studies conducted upon academic publication data, several benchmark datasets which are openly available were selected for use in the evaluation of topic modelling and clustering experiments, which is focused upon Chapter 7. Within these, the 20 Newsgroups, Text Retrieval Conference (TREC), AG's Newss[9],

[1] https://twitter.com/

[2] https://reveela.com/

[3] https://www.crossref.org/education/retrieve-metadata/rest-api/text-and-data-mining-for-researchers/

[4] https://dev.elsevier.com/

[5] https://dev.springernature.com/

[6] https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining

[7] https://core.ac.uk/services/api/

[8] https://pypi.org/project/arxiv/

[9] No publication is associated with this dataset, taken from: http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

and BBC News datasets are applied. The details on the selection of these datasets is provided in Chapter 7.

3.4 Ethical Considerations

3.4.1 Ethics of Social Media Data

It essential to consider ethical implications when working with data obtained from social media platforms, as was performed in Chapter 4. To ensure the removal of personal user information, only the unique identifier of a user, and embeddings derived from user biography, and tweet content is collected and stored, which ensures adherence to the developer agreement set forth under the X (formally Twitter) software agreement[10]. At the time when this research was conducted, an agreement was established through the Twitter Academic Research Program[11], which enabled the collection of Tweets through a sampled stream. When evaluating the accuracy of recommendations, the unique identifier associated with each user was used to generate a URL linking directly to the profile of that user, allowing labellers to evaluate whether the profile was a relevant recommendation.

3.4.2 Ethical Approval for Studies Involving Humans

In Chapter 6, when evaluating a framework for assisting literature review, a study based on the feedback of medical professionals was conducted through a participant questionnaire. It is essential to ensure that ethical approval was obtained for the participant study conducted in this work. This was approved by Durham University Ethics Board[12], approval code 1490. To ensure the adherence to ethical requirements, informed consent was obtained from all participants prior to participating in the study. Participants were provided with a written summary of the project, including how the information collected will be used in an information sheet; additionally, participants were provided with an oral verbal summary of the project,

[10] https://developer.twitter.com/en/developer-terms/agreement-and-policy

[11] https://developer.twitter.com/en/use-cases/do-research/academic-research

[12] https://www.durham.ac.uk/research/ethics--governance/

along with how the data will be used. No personal information was collected from the participants. The participant information sheet is provided in Appendix A.

CHAPTER 4

Construction of Meta-Embeddings to Enhance Journalist Retrieval Based on Full-Text Queries

This Chapter responds to GRQ1, and contains material which was published in the 13th International Congress on Advanced Applied Informatics Winter Conference, titled Hybrid Weighted Retrieval of Twitter Users for Temporally Relevant Full-Text Querying in the Media Industry.

4.1 Prologue

A key difficulty for principal actors within the publishing industry is the promotion of content to suitable audiences. Publishers, who are tasked with the promotion of content, may have little knowledge regarding a subject on which a particular article discusses, which can in-turn lead to difficulties in promoting content. Furthermore, barriers to the delivery of journalistic content to suitable media outlets present difficulties to both journalists and publishing houses. These may take the form of barriers to the identification of key individuals to whom the content is relevant and who may be influential within their domain topic.

To address this problem, a methodological approach to automated recommender systems for journalists or content-writers is proposed, which enables the identification of social media accounts, based upon information retrieval techniques. This addresses the first general research question In-depth research questions for this are discussed in Section 4.1.2.

This study investigates three areas, which to current knowledge have yet to be applied to such a domain; first, how user-defined descriptions, commonly associated with user profiles on the Twitter platform, may impact the perceived accuracy of recommendations. Second, the application of a full-text query facility to accept journalistic content queries. Finally, an investigation into time decay, as part of the recommendation-ranking methodology. This results in a proposed novel hybrid weighting methodology to design meta-embeddings, which provides advantages to baseline information retrieval techniques and demonstrates the influence of temporal decay ranking upon model performance. This is presented in the published work *Hybrid Weighted Retrieval of Twitter Users for Temporally Relevant Full-Text Querying in the Media Industry*, which was published in the 13th International Congress on Advanced Applied Informatics, Winter Conference [110].

As discussed in the Introduction section 1.1, the publishing industry has experienced a significant shift in the consumption of news and journalistic content from print to online, with sales of national and local printed newspapers in the UK falling by roughly half from 2007 to 2017 [7]. However, the industry currently remains reliant upon the manual promotion of material, from which advertising revenue is crucial. By alleviating the time-consuming nature of promoting journalistic content outside of the domain, publishers may benefit from reduced operating costs with the support of automated procedures.

On the other hand, the number of people consuming news and magazine media online has continued to grow, with 70% of individuals reporting to read or download online media in 2020 in the UK [9]. Moreover, Social Network Service (SNS) usage has grown across all age demographics, with more than 86% of 16-54 year olds now owning an SNS account [116], a considerable growth over the last 5 years in the UK. Content posted on these platforms provides a valuable data source, that can be applied in data mining techniques, to assist in addressing a variety of problems, such as identifying marketing opportunities or promotion of content or services.

Given an item of journalistic material, this chapter addresses the issue of searching for relevant users of online social networks. Traditionally, individuals in the publishing industry typically rely on manual searching. It is additionally useful, from a marketing or journalistic perspective, to provide results which are temporally relevant, based upon the premise that interest in topics may change over time, based on what is currently in the public interest. In this work, the aforementioned requirements are addressed through an information retrieval perspective, targeted at journalists, writers, and publishing houses. The resulting methodology provides several benefits to the industrial problem, including an improvement in accuracy, compared to information retrieval based on state-of-the-art transformer embeddings, and facility for the temporal ranking of results, which ensures to enable the retrieval of recommendations that have recently discussed a semantically similar subject to the query article of journalistic material.

4.1.1 Importance to the Publishing Industry

The work provided in this chapter contributes to the publishing industry by providing an end-to-end framework for the retrieval of relevant social media accounts. This addresses the issues identified by the UK government Cairncross review [7], through enabling publishers a means of the identification of relevant audiences based on the semantic content of the content of social media platforms.

4.1.2 Chapter Specific Research Questions

The following research-questions were defined for this aspect of research:

- **RQ1.1:** *How may information retrieval techniques be adapted to facilitate the identification of relevant social media accounts for a given article of journalistic content, based on a user's posted content?*

- **RQ1.2:** *How does hybridisation of tweet and user biography embeddings contribute to recommendation quality in the retrieval of social media accounts based on full-text querying?*

- **RQ1.3:** *Can forward time-decay functions be applied to recommendation ranking to improve the temporal relevance of results, and how does this affect the quality of recommendations?*

The main contributions of this work are as follows. First, a demonstration into how the hybridisation of posted content and user biographies using weighted averaging of word embedded features provides an advantage compared to baseline instances of applying individual features when used in a semantic search context, and define an optimal value for this weighting. Second, the work contributes to the evaluation of forward time-decay [117] as a ranking procedure in the information retrieval task, and demonstrates the impact this has on the quality of recommendations. The resulting architecture is provided as a commercially implemented tool, which is available at Reveela[1], a platform providing tools for journalists, content writers, and publishing houses.

4.1.3 Research Objectives

Based on the above motivation and research questions, the research objectives for this Chapter are:

- **RO1.1:** Investigate the effects of the construction of meta-embeddings from multiple text sources to enhance semantic search. (Addressing RQ1.2).
- **RO1.2:** Investigate the adaptation of forward time-decay techniques for the temporal ranking of information retrieval results. (Addressing RQ1.3).
- **RO1.3:** Provision of an information retrieval-based framework for the accurate retrieval of relevant social media accounts based on the semantic contents of social media posts and profiles, given a full-text query. (Addressing RQ1.1).

[1] `http://www.reveela.com`

4.2 Related Work

4.2.1 Content Recommendation

One of the most prevalent approaches for content recommendation is through collaborative recommendation [118], where data about users is used along with content ratings, to identify content a user could like, based on ratings provided by users with similar profile features. This presents advantages in that an expected rating for a user can be extrapolated from existing data on other users who have rated an item of content. However, limitations may arise due to the reliance on prior user-generated data. A "*cold-start*" problem occurs when a recommender system is configured, without any prior evaluation data, such that the recommendations are of poor quality initially [15]. Furthermore, any new user, or the addition of any new entries to the recommender system, would experience the same poor quality in recommendations, until sufficient information on the user or entry was collected. In addition to this, adversarial attacks may affect the quality of recommendations through spam ratings generated from malicious attacks [119–121]. For this aspect of research, the mitigation of these limitations will be demonstrated, by eliminating the need for user-collected ratings and demographics through a semantic search approach to recommendation.

4.2.2 Information Retrieval

In contrast, information retrieval methods, such as semantic search, may operate without priors, such as content ratings. In information retrieval, a common approach to semantic search is to identify query results using cosine similarity [122]. Such an approach may be used across a range of domains including video, images, or text through measurement of the similarity between a vectorised query and vectors of all possible results. The method used to obtain the vectors is independent of the ranking metric, and approaches may include TF-IDF, BOW, or Neural Embedding Methods such as transformer embeddings [5]. These exhaustive methods of nearest-neighbour search using cosine similarity may provide the closest-neighbour vectors given a query vector. However, a brute-force nearest-neighbour search in this manner presents a time complexity of $O(m * n * d)$ with m being the dataset size, n the number of queries to perform, and d being the dimensionality of the query data. Scaling complexity in terms of vector dimensionality, the curse of dimensionality [123–125], may be addressed to some extent via dimensionality reduction algorithms such as UMAP [126], and these steps have been demonstrated in collaborative recommendation engines. However, in neural embedding methods such as those provided by transformer networks, the contextual value of high-dimensional embeddings may be lost when dimensionality reduction is applied. In collaborative recommendation experiments, it was found that a larger dimensionality was found to also lead to a reduction in false positive results [125]. An alternative method to address scaling issues in a brute-force cosine similarity search is through the use of Approximate Nearest Neighbour algorithms (ANNs) [127,128], which trade a degree of accuracy for significantly improved inference times. Of these, Hierarchical Navigable Small Worlds algorithm (HNSW) has demonstrated effective performance with high-dimensional vectors [127]. Therefore, this aspect of the research employs this algorithm as the search strategy when utilising transformer-generated embeddings. A technical discussion HNSW is provided in Section 3.2.2.

As this work focuses upon the retrieval of Twitter users, it is necessary to evaluate current research related to this domain. Specific approaches to the identification and recommendation of Twitter users have been demonstrated in a semantic search context. Works such as [129] have provided semantically matching Twitter influencer recommendations, based on the content posted by the respective Twitter profiles and the audiences of such profiles. In contrast, [130] applied facet-driven search through enrichment of the semantics of tweets, through the extraction of facets and the linking of these, to provide additional metadata defined during search. Identification of named facets, such as locations or named entities, may additionally allow linking of external data sources, such as DBPedia [131], through the construction of RDF queries [130]. By leveraging such external data sources, prior work achieved a significant improvement in Mean Reciprocal Rank (MRR) [132] compared to faceted search strategies, which use hashtags or keywords only, when defining facets. This demonstrates that the presence of whole tweet content may provide semantic value when employed in search strategies; our research, therefore, seeks to ensure to maintain the full context of any tweet within our approach. Contrasting approaches also exist, which seek to identify content or relevant advertising, based upon user profiling [133, 134].

Additionally, hashtag retrieval has been presented as a method for the retrieval of tweets. Based on this approach, works have proposed the retrieval of hashtags based on a user query, through semantic enrichment of hashtag content [135], or through pattern mining, with [136] applying a pattern mining algorithm which facilitates the trivial and temporal retrieval of hashtags based on a user query. In [136], a query is defined as a group of hashtags, and does not facilitate full-text querying; however, the authors claim the interoperability of the approach in alternative search strategies.

Temporal features have been demonstrated in information retrieval and recommender systems [137], with temporally-aware approaches performing at a competitive level compared to collaborative recommenders. In such systems, a time-decay function may be applied to provide a weighting during the ranking of results; however, these may rely upon a 'backwards' function, so that weightings are applied based upon the age of a given item in relation to the current time. As time progresses, such as in a data streaming environment, it becomes necessary to recompute the weightings of items, as the age of items changes. In comparison, a forward time decay model was proposed by [117], which approached the task through the measurement from a defined point in time. This has been demonstrated to model existing decay functions in a manner that is scalable. As the industrial deployment of the outcomes of this aspect of work take place in a data streaming environment, with the number of Twitter users present within the data growing continually, a monomial forward decay function defined by [117] is employed as a weighting strategy in the ranking of Twitter users, based upon the recency of their tweets.

4.3 Materials and Methods

As discussed in our introduction, this work has been implemented online and is available for commercial access at Reveela[2]. The Reveela platform is targeted at journalists, publishers and content writers and provides tools to assist in mitigating the novel issues encountered during the transition from print to online media [138]. The proposed architecture is implemented as part of a larger document analysis tool, which allows producers of journalistic content to analyse their article and identify relevant journalists and influencers.

4.3.1 Curation of the Influencer Dataset

A dataset of $390,750$ tweets, with 9322 unique users was collected through a real-time data stream via the Twitter streaming API [3], providing a sampled 1% stream of new tweets posted on the platform. This sampled stream is defined within geographical parameters, based on the intention of ensuring that only tweets posted within the UK are collected. This collection method was determined due to the large volume of tweets, which may be collected and recomputed by a HNSW index in batches in the industrial deployment of the framework. Keyword parameters were not specified for collection at this stage. This is due to the necessity to collect a wide sample of topics to ensure the model is robust enough to handle all journalistic domains. This data was collected from the period of January 2021 to January 2022.

Implications of Sampled Stream Collection on Data Bias

While the collection of tweets through a sampled stream ensures that a large volume of tweets can be obtained, there are several caveats to consider. Firstly, it can be assumed to be difficult to ensure that multiple tweets can be captured from the same user, as there is only a 1% chance that a user's tweet will be captured when posted, this could lead to sparsity within the data and an under-representation of the topics posted by the collected influencers. In contrast, by focusing on tweets

[2] https://reveela.com

[3] https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction

that are determined within geographical parameters, a bias could be introduced within the data. Research published since this study was conducted has found that the geo-tagging of tweets, which is performed by users, is generally representative of exhibiting significantly more positivity and typically relates to special events in a user's life [139]. The authors of this study also reported that a dataset consisting of solely geo-tagged tweets represents significantly fewer unique users when compared to a dataset which was collected randomly without geo-tagging parameters, which is argued to be because randomly collected tweets are sampled from a larger pool of users. This has significant implications for the research focus of this chapter, where only tweets with geo-tags were collected with the aim of ensuring a UK-only subset of data. As denoted in Table 4.1, the number of unique users is significantly smaller than the number of tweets present within the dataset, which is further exaccerbated by the filtration process. Additionally, the collection of geo-tagged tweets does not guarantee that the posting account is based within the defined geographical parameters, as, given that research [139] has highlighted the prevalence of geo-tagging in special events, there is the possibility that the geo-tagged tweets are posted during travel or holidays.

Concerning the time frame for the collection of the data, entailing a full year from January 2021 to January 2022, it is also worth considering the impact that this collection window will have upon recommendations from the perspective of current, and future influencer recommendations. A social media account is not a static entity, with account users being able to both post new content and also delete any historical posts. Thus, the subjects that an account discusses can undergo conceptual drift if they post on different topics from those that were identified during initial data collection. This can be further exacerbated by the issues raised in relation to the 1% sampled stream collection method, such that the limited number of tweets obtained for each unique account implies a limited spread of the topics discussed by an account, this is raised further in Section 4.5. Furthermore, from the perspective of post deletion, the tweets captured during data collection may no longer be available after collection for several reasons including the deletion of the tweet by the user, the deletion of the account, or the suspension/removal of the account by moderation. This introduces bias in the data, and the subsequent recommender system, due to a mismatch between the data accounted for during recommendation, and the data available at the present time. The problem of data persistence [140] arises from this phenomenon, which is particularly prevalent in topics which discuss politically right-leaning topics and harmful content, where the suspension of accounts plays a significant role. Relating to this, *Misleading Repurposing* [141], is a practice where a social media account is modified for a new purpose while retaining the original followers of the account. While this can have benign intentions, such as users changing their interests and topics of discussion organically, they can also be through malicious purposes, including the sale of accounts with high followers, or the intentional switching of account themes for political motives. Methods for the identification of misleading accounts include the monitoring of follower accounts, as accounts with a high number of followers have a higher chance of being misleadingly repurposed, monitoring any periods of dormancy of the account if the owners are no longer using them and intend to sell them, and the monitoring of any deleted tweets which may no longer fit the narrative of the account.

Within the scope of the research in this chapter, both *data persistence*, and *misleading repurposing* should be considered both from the perspective of research and industrial deployment and continued use. While the recommendation of influencers does not focus on political themes but instead aims to capture a general range of topics, the problem of data persistence should still be considered both in terms of accounts which were captured but have since been deleted by the users or suspended by moderation, and the general conceptual drift which could arise from the long window of data collection where users may have changed their biographies or general topic of discussion. More prevalently, given that accounts with high follower counts have been found to be more susceptible to misleading repurposing, and given that the criteria defined for an influencer in this chapter is an account with more than 10,000 followers, consideration must be made in monitoring accounts in the dataset for any drastic changes in thematic focus, renaming, and account dormancy. To address this in the industrial deployment of this research, data is frequently updated in a batched manner via official X API resources to maintain up-to-date records of user account details such as profile names, biographies, and deletions. This also enables the identification of dormant accounts through the API by the collection of the most recent tweets posted by the account. While being valuable from an industrial and research perspective in mitigating instances of bias, this is also of paramount importance in adhering to both the terms of service for the X API[4], and adhering to General Data Protection Regulations in the United Kingdom in keeping data up-to-date.

Data Processing

For the processing of the influencers dataset, several stages of filtering were conducted, with the number of tweets and unique accounts in the dataset being detailed for each step in Table 4.1. The following sections elaborate on the methodology of this process, as well as the underlying rationale.

Initially, it is necessary to determine the criteria for an influencer. For this, a minimum follower count was defined at $10,000$ followers, with any accounts captured from the filtered stream that have fewer followers than this being discounted from the data set. While the resulting industry application of this research is an engine for the recommendation of social media influencers, it is necessary to make clear that the research focus is specifically the curation of meta-embeddings derived from textual data sources for enhancing vector retrieval purposes. Influencer detection on social networks is an independent area of study, with a steady growth in volume of publications investigating the field being observed [142]. However, the detection of influencers is not the focus of this study, hence the definition of an arbitrary follower count limit of $10,000$. The implementation of this filtration step has a considerable impact on the number of unique accounts, reducing from $36,869$ to $9,919$.

Next, tweets and bios are filtered to remove any profiles which do not have user biographies, and the removal of any tweets which are blank. Since the formation of meta-embeddings depends on both tweets and biographies, it is crucial to eliminate any empty features. These features provide no semantic value during embedding

[4] `https://developer.x.com/en/more/developer-terms/agreement-and-policy`

and could skew the meta-embeddings when averaging an embedding from an empty source with one from a valid source, such as a complete tweet alongside an empty bio. Similarly, the removal of duplicate tweets and bios is performed, with the objective of eliminating duplicate tweets as a priority given that these would also skew meta-embeddings. In terms of impact on dataset characteristics, the removal of blank tweets and bios has a small impact on the volume of tweets and unique accounts. However, the removal of duplicate tweets leads to a reduction in the dataset size from 592,350 to 516,316. It is worth noting that the filtration of user bios will have an influence on data bias as only accounts with a valid bio are accounted for in recommendations.

Tweet length filtering is applied with a minimum requirement of 30 characters. This criterion is derived from literature and anecdotal evidence, indicating that the median tweet length in the UK ranges from 30 to 57.5 characters, depending on the administrative region, based on data from 2009 to 2012 [143]. It is important to note, however, that the median tweet length from 2009 to 2012 may differ from more recent data. Non-peer-reviewed case studies have suggested that longer tweets may garner more engagement [144,145], which is a sought after criteria for influencers, but an analysis of five companies showed no significant correlation regarding the variable of tweet length [146]. Nevertheless, assuming that longer tweets may contain more valuable semantic content and considering the reported median tweet length for the UK, a minimum tweet length threshold of 30 characters has been established as, being the lower bound of the median length reported, this should ensure to continue to capture an adequate volume of data while also eliminating any tweets shorter than the norm. The implementation of this filtration stage confers a significant reduction in the number of tweets in the dataset, with 122,685 being omitted based on tweet length.

The final stage of the filtration procedure involves the elimination of any potential sensitive material from the dataset. The recommendation of accounts that promote violence, hatespeech, or pornographic material would understandably have a negative impact upon any organisation which deploys an industrial version of the research in this chapter, and from an ethical perspective, it is paramount that the domain experts tasked with labelling of recommendation results are not exposed to any unsafe material. The filtration process is broken down into two main stages. First, a lexical approach compares tweet and biography content against a pre-defined lexicon of banned terminology. The lexicon was constructed by amalgamation of the open-source datasets Banned Word List[5] and David Sojevic's Profanity List[6]. Thus, any tweets or biographies that contain any of the pre-defined banned words are omitted from the dataset. Finally, accounts on the X platform are required to label their content as sensitive if they contain or promote any nudity, violence, or other sensitive topics. Thus, any data which contains the sensitive tag, which is returned by the API results is removed from the dataset. It is important to acknowledge that both of these filtration processes have limitations. First, from the perspective of lexical filtration, a rule-based approach requires an exact match within the content to the banned lexical terms, and is therefore vulnerable to any adversarial changes in spelling in order to bypass rule-based filtering. Furthermore, given that language is constantly evolving in nature, slang, and what is deemed to be offensive, it becomes intractable to maintain an up-to-date exhaustive list of all potentially harmful terminology. The implementation of rule-based offensive language filtering is also susceptible to the Scunthorpe problem [147], wherein false-positives arise during filtering due to a banned string residing within a genuine word. To address this, a whitelist of words was curated based on input from the industry sponsor, including words such as *scunthrope* and *torpedo*, to name a few. Concerning the removal of tweets labelled as sensitive, relying on self-reporting by authors is flawed, as accounts with malicious intent could potentially avoid labeling their content as sensitive. Still, by adopting a two-pronged approach to filtering of sensitive material, a total of 2,881 tweets, and 61 accounts were removed.

4.3.2 Evaluation Dataset

A dataset of journalistic content for evaluation can be accessed through the Microsoft News Dataset (MIND) [148]; however, this dataset focuses on news content. In

[5] `http://www.bannedwordlist.com/`

[6] `https://github.com/dsojevic/profanity-list`

Table 4.1: Total number of unique tweets, and unique Twitter accounts in influencer dataset at different stages of filtering.

Filtering Stage	**Tweets**	**Unique Accounts**	**Avg Tweets per Account**	**Avg Tweet Length**	**Avg Bio Length**
No Filtering	687,600	36,869	18.392	101.265	112.438
Minimum Follower Count	594,504	9,919	59.714	95.907	113.935
Blank User Bios	592,350	9,827	59.875	96.095	113.934
Duplicate Tweets	516,316	9,825	52.180	30.342	113.932
Tweet Length	393,631	9,383	41.637	108.742	114.604
Sensitive Tweets (Lexical)	392,781	9,379	41.566	108.719	114.620
Sensitive Bios (Lexical)	391,222	9,348	41.495	108.755	114.616
Sensitive Tweets (Twitter Tagged)	390,750	9,322	41.533	108.721	114.670

comparison, the industry problem domain applies to both the news and magazines industries, respectively. Therefore, a novel dataset was defined. This was curated through the application of knowledge of three domain experts within the industry and existing commercial data, to construct and curate an evaluation dataset of magazine and news articles articles. This entails a total of 4676 UK magazines and was curated by domain experts to include descriptions of magazines and a label of their industry sector, which is not used in this analysis. Given the limitations of offline evaluation of recommendations at a large scale, a sample of 150 articles was randomly selected from the original set for model evaluation and labelling by domain experts.

4.3.3 Computation of Embeddings

Given that the proposed architecture will provide Twitter influencer recommendations based on a given text query, it is necessary to select a suitable method to generate a vector representation of the query article and candidate social media accounts. Embedding models for text have observed a significant shift towards deep learning and, more specifically, the transformer architecture [5] in recent years. Methods such as BERT [26] and its subsequent revisions consistently perform well in benchmark tasks. However, modifications to methods of pre-training have provided improvements to these models. One such improvement is presented by MPNet [41], which contributes a unification of the Masked Language Modelling (MLM) pre-training procedure of BERT, with Permuted Language Modelling (PLM) of XLNet [149]. In MLM, pre-training may fail to model the complex dependencies of language due to assumptions made about the independence of tokens [149]. PLM, on the contrary, may model dependencies between tokens. The unified pre-training procedure presented by MPNet presents improvements in benchmarked results compared to MLM and PLM architectures [41, 149]. Furthermore, experiments facilitated by the Sentence-Bert library indicate the overall best performance of MPNet in semantic similarity tasks [33]. As the core goal of this study is essentially a semantic similarity search to identify the most relevant Twitter user embeddings based upon a given query embedding, the MPNet architecture was selected as an embedding technique.

4.3.4 Semantic Searching via Meta-Embedding

Retrieval of Twitter accounts is performed using a HNSW [3] approximate vector search, which identifies Twitter user accounts based on the semantic similarity between the candidate embeddings and the query embedding. HNSW is selected as it provides an efficient and scalable method of nearest-neighbour searching, which is suitable for the large volume of data used in this experiment. While the number of unique influencer accounts is relatively low, it is essential to note that retrieval is performed upon meta-embeddings constructed from a weighted-average of tweet and biography embeddings. This means that a different meta-embedding must be computed for every unique tweet produced by a user. Thus, the size of the index will comprise $390,750$ unique meta-embeddings. This is an essential consideration given that the outcomes of this aspect of research are finalised into a commercially available solution, where the wait-time for users must be as low as possible. It is necessary to compute embeddings for queries and tweets/user biographies for use in the information retrieval approach facilitated by HNSW. For commercial implementation, these are generated using the MPNet transformer architecture [41], which provides advantages in minimising parameters and limiting memory constraints, thus ensuring minimising operating costs. For the current work, this approach is compared to the SGPT model, which provided the current competitive benchmark for zero-shot and entity retrieval tasks[7] [150] at the time of publication. Embeddings are computed from the text content of an individual tweet or user biography, following pre-processing steps. This consists of the removal of any special characters or hyperlinks. However, any usernames or hashtags present in a tweet are not removed, instead only having only their "#" and "@" symbols removed. This is due to the potential for hashtags, in particular, to be used within a sentence and therefore conveying semantic context e.g. *"Can't believe the #football last night!"*. As pre-trained transformer networks are used for the generation of tweet vectors, it is necessary to ensure to maintain contextual semantics in the text and thus do not perform any stopword removal, stemming, or lemmatisation.

As the focus of this work is on the evaluation of meta-embeddings based on tweet content and user biographies, experiments were defined to compare the performance across a range of weighting parameters. Hereafter, tweet-biography embeddings refer to the computation of a single meta-embedding based upon a weighted average of tweet and biography embeddings, respectively. The weighting parameter w represents the weighting of the tweet vector relative to the biography vector in the calculation, such that when $w = 0$ the resulting vector is identical to an embedding calculated from a user biography only. The resulting tweet-biography meta-embeddings are then used in the construction of a HNSW index. At the inference stage, an embedding of a given query is then provided as a query to HNSW in order to identify semantically similar tweet-biography embeddings.

Given the resulting set of candidates identified by HNSW search, candidate ranking is calculated through the Cosine similarity distance between the query vector and identified index vector, with a higher degree of similarity indicating a higher ranking of results. Cosine similarity is selected as a distance measure in final candidate ranking, as the distance ranking of HNSW is approximated. Furthermore, cosine similarity is widely used in existing information retrieval systems [151, 152].

[7] https://paperswithcode.com/paper/sgpt-gpt-sentence-embeddings-for-semantic

4.3.5 Time-Decay

A relative forward time-decay function for the ranking of results, based upon [117] is subsequently used in the ranking of the results identified by the HNSW algorithm. Given a tweet T of a suggested user and the predefined monomial function $g(n) = n^2$, the temporal relevance value of TR is such that:

$$TR\left(i,t\right) = \frac{g\left(t_i - L\right)}{g\left(t - L\right)}$$

Here L is the time of the point of reference, defined as the smallest timestamp within the dataset. ti represents the time of the respective tweet, and t is the current datetime. Temporal relevance is calculated in UNIX epoch time. As temporal relevance is independent of the distance metric for calculating semantic similarity, a final calculation is performed, taking the weighted average of the semantic relevance, and temporal relevance when calculating the overall relevance score for a recommended item. In this case, the determination of semantic relevance is calculated as the cosine-similarity measure between the query embedding and the candidate influencer meta-embedding. While the approximate distance from the query to the candidate embedding yielded by HNSW may be applied, this approximation warrants preference for an exact cosine-similarity calculation for enhanced accuracy. The formula for cosine-similarity computation is presented in Section 3.2.1. For evaluation, this is performed across a weighting range of 0 to 1, such that a weighting of 0 correlates to discounting the temporal relevance scoring in the overall relevance calculation, and a weighting of 1 implies an equal weighting in the calculation.

4.3.6 Evaluation of Meta-Embedding Weightings and Time-Decay Influence

Evaluation is performed on the top five scoring Twitter user accounts, given a query article. At the inference stage, a given article of journalistic content is vectorised, using the same embedding method applied to tweets, via the MPNET transformer [41]. The resulting embedding is then applied as a query vector for semantic search using HNSW [3]. The impact of including biography-tweet embeddings upon perceived model accuracy is evaluated across a range of weightings, incrementing by the weighting parameter in units of 0.25.

To evaluate the performance of the proposed experiments in terms of Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) measures, domain experts are provided with 150 articles randomly selected from an internal database of magazine and news articles. Domain experts were selected internally within the industry sponsor organisation based on their expertise and experience within the publishing domain. The top-5 scoring recommendations provided by the model were evaluated by 3 domain experts. Labellers were required to fully read all articles and then examine the respective Twitter profiles for all recommendations for that article. The criteria for each evaluation was for labellers to visit the profile of the recommended accounts via a hyperlink embedded in an evaluation spreadsheet which included the query article text, and hyperlinks to the top-5 recommended influencer accounts. After visiting the profile page, labellers were required to read the profile biography, and the first ten tweets posted by the account. This was specified to be original tweets, and the reviewers were instructed not to take into account retweets in their evaluation. Based on the user biography and tweet history, reviewers were then required to place the recommendation in one of three label categories. Label 0, which corresponds to no perceived relevance between the article and the Twitter user, label 1 corresponding to a weak degree of relevance or tweet posting history, and label 2 representing a strong link between the recommended article and the recommender Twitter account. On the basis of the labels applied by each domain expert, a majority vote was then used to assign the final label for the recommendation. In instances where votes were evenly split, where each label had one vote, reviewers were required to discuss and agree on the final label. On average, the total number of words per article was 726, with the upper and lower bounds for the length of the article being 388 and 887 words, respectively. Given the requirements to read each article fully and examine multiple Twitter profiles, the evaluation was limited to 150 articles. Evaluation is performed on a weighting w range from 0.25 to 2 representing the ratio of biography to tweet weighting used when computing meta-embeddings, and additionally includes experiments applying only a user biography in calculating embeddings, as well as only using tweets when calculating embeddings.

While evaluating in terms of MAP and MRR is useful in determining the precision of recommendations, which factors the proportion of relevant recommended items out of all recommendations, it remains necessary to also consider recall in recommendations. Recall, contrastingly, is a measure of the proportion of relevant recommended items compared to the total number of relevant items.

In a recommender system, this presents a challenge, due to requiring an existing set of labelled data, due to the necessity of calculating the number of relevant recommendations based on the proportion of actually relevant items. Thus, to evaluate recall, we perform a prior labelling of a sample of influencers, which is time-costly. Thus, we opt to approximate the recall, by randomly selecting a publication from the internal set, and then randomly selecting 100 influencers from the dataset and assigning a label to each influencer, to determine their relevance label, following the same criteria used in the prior evaluation of MAP/MRR with 3 domain experts. This influencer sample is then used to construct an index, and evaluate this across the same weighting parameters used in the evaluation of precision. Although this approach is beneficial for assessing the impact of weighting parameters on the quantity of relevant items within a sample of pre-identified relevant influencers, the limited sample size of merely 1 is suboptimal. This constraint arises from the labor-intensive process of evaluating a substantial number of influencers, necessitating that each of the 3 domain experts label 100 influencers. Therefore, it is imperative to acknowledge that although recall is a valuable metric, the results obtained for recall should be approached with caution, and an expanded sample size would contribute significantly to a more comprehensive understanding of the outcomes of this research.

4.4 Results

Table 4.2: Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) of top-5 for biography only, and tweet Only recommendations.

Model	MAP	MRR
MPNet Biography	**0.78**	**0.66**
MPNet Tweet	0.46	0.28
SGPT Biography	0.68	0.54
SGPT Tweet	0.51	0.3

Table 4.3: MAP, MRR and Recall of top-5 recommendations when comparing weighting parameters.

Weighting	Strong Relevance			Strong and Weak Relevance		
	MAP	MRR	Recall	MAP	MRR	Recall
Bio Only	0.78	0.66	**0.2**	0.9	0.83	0.2
0.25	0.8	0.71	0	0.91	**0.86**	**0.6**
0.5	**0.81**	**0.72**	0	**0.93**	**0.86**	**0.6**
0.75	0.79	0.68	0	0.92	0.83	**0.6**
Equal Weighting	0.77	0.65	0	0.9	0.79	0.4
1.25	0.75	0.6	0	0.86	0.75	0.4
1.5	0.72	0.55	0	0.83	0.7	0.4
1.75	0.7	0.51	0	0.8	0.67	0.4
2	0.68	0.48	0	0.77	0.67	0.4
Tweets Only	0.46	0.28	0	0.59	0.4	0.2

4.5 Discussion

Table 4.2 presents the performance of recommendations for baseline models constructed from an HNSW index constructed using embeddings based on user biographies, or user tweets only, using the MPNet or SGPT transformers. For both embedding models, an index constructed based on user biographies performs best, with the MPNet biography only model achieving a Mean Average Precision of 0.78, and Mean Reciprocal Rank of 0.66 when accounting for the top-5 recommendations. In comparison, basing recommendations on tweets demonstrates poor performance for both embedding methods.

Although the presence of embeddings generated from user tweets indicates poor performance when applied alone, Table 4.3 demonstrates an improvement in both MAP and MRR in the presence of hybrid weighted embeddings. However, with respect to recall, which was assessed on a substantially reduced test sample due to

Table 4.4: Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) of top-5 recommendations when comparing temporal weighting parameter, accounting for strongly relevant labels only.

Weighting w	**MAP**	**MRR**	**Recall**	**Average Recommendation Age (Days)**
No Temporal Ranking	0.81	0.72	0	207.03
0.1	0.79	0.7	0	144.50
0.2	0.79	0.69	0	120.93
0.3	0.78	0.69	0	111.26
0.4	0.77	0.69	0	103.63
0.5	0.77	0.69	0	97.68
0.6	0.77	0.67	0	91.56
0.7	0.77	0.66	0	86.53
0.8	0.77	0.65	0	82.15
0.9	0.77	0.64	0	76.45
1	0.77	0.63	0	73.96

the labor-intensive process of labeling the recall test set, only the bio-only retrieval index effectively identifies relevant outcomes from the expert-labeled recall sample for strongly relevant results. When strong and weakly labeled results are considered, recall scores suggest that a weighting configuration between 0.25 and 0.75 achieves the highest recall of 0.6, indicating that 4 out of the 5 recommendations were deemed as being either strongly, or weakly relevant to the query article. Based on these scores, an optimal threshold for hybrid embeddings is identified with a weighting w of 0.5, achieving a MAP of 0.8, MRR of 0.72, and Recall of 0.6.

The optimal weighting of 0.5 was then applied to an evaluation of the temporal ranking approach. This evaluated the performance of the model at different weightings during the ranking calculation. A general small trend is identified in a decrease in performance when increasing the weighting of temporal ranking in the overall ranking score, with results presented in Table 4.4, compared to the baseline MAP and MRR of 0.8 and 0.72 respectively, which is achieved by the optimally weighted hybrid embedding model when not accounting for temporal ranking (See Table 4.3). Notably, the recall score for all evaluations is 0, due to the fact that only strongly relevant labels were accounted for during the temporal analysis, where comparably, Table 4.3 indicates that recall is 0 for strongly relevant results, and does not appear to be influenced by temporal weighting. Table 4.4 demonstrates the average age of the recommendations, when increasing the weighting of temporal relevance in the ranking calculation. At a weighting of 0, temporal relevance is not taken into account in the overall relevance calculation. Hence, this represents a baseline with an average recommendation age of 207 days relative to the most recent entry in the data set. Increasing the weighting of temporal ranking in the overall relevance ranking calculations demonstrates a decrease in the average age of recommendations, which may be preferable in recommender systems. In comparison, increasing the influence of temporal ranking leads to a decreasing trend in MAP and MRR, however this small decrease is admissible in a domain where temporally relevant results are desired.

4.5.1 Study Limitations

While this study provides valuable insights into the effectiveness of meta-embeddings in the retrieval of social media accounts, it is important to recognise certain limitations that may impact the interpretation and generalisability of the findings, as well as any implications on future industrial impact.

Notably, the X API has discontinued free access to the sampled stream, with a filtered and sampled stream costing \$5000 monthly[8] in September 2024, which provides up to $1,000,000$ tweets per month. Furthermore, a monthly quota is now in place on all paid plans, with no free access to the API for tweet retrieval. This considerable cost of access to data incurs difficulties in the reproducibility of this study, particularly given the high cost of the sampled stream service. One alternative, albeit still costly method, can be facilitated through the X API Search Tweets Endpoint[9], which is accessible through the first paid tier of the platform for \$100 per month, enabling the retrieval of $10,000$ tweets per month. This endpoint provides a means for the retrieval of up to 100, chronologically ordered tweets based on a given query search parameter, and also permits the specifying of geographical parameters.

[8] `https://developer.x.com/en/products/x-api`

[9] `https://developer.x.com/en/docs/x-api/tweets/search/api-reference/get-tweets-search-recent`

However, the need for the specification of search terms limits the collection of a general feed of tweets within the geographical parameters which could be achieved via a filtered stream, and also relies upon the curation of a suitable lexicon of search terms. This may, retrospectively, enable the curation of a better quality dataset which has a higher degree of semantically-valuable content through careful selection of suitable search terms.

Regarding the evaluation of meta-embeddings by domain experts, the laborious nature of the labelling process, entailing 3,150 labelling tasks (150 Articles $*$ 21 Experiments) for each domain expert, constrained the feasibility of performing a more extensive investigation, with the domain-experts being required to conduct the labelling within their normal working hours and busy schedules. Thus, it is evident that an expanded and more comprehensive investigation on a larger evaluation set could substantiate the findings. The methodology followed during labelling was also vulnerable to the individual biases of each labeller, as well as the vagueness of the labelling criteria where label 1 corresponds to a weak degree of relevance or tweet posting history, and label 2 representing a strong degree of relevance. In both of these instances, the concept of relevance is open to interpretation by the labeller, with their prior experiences within the domain of each article influencing their perceptions on the relevance.

Moreover, it is necessary to acknowledge that profiles within the dataset may have been altered, either through the posting of additional tweets or through modifications to biographies since the time of data collection. Such changes may affect the labelling results owing to the conceptual drift of profiles, which may no longer be centred on the semantic domain of the profile at the point of collection. Thus, the findings that a priority of bio-weighting to tweet content in meta-emebddings as identified in Table 4.3 is particularly vulnerable to any future changes in the bios of users. To address this, the industrial deployment of the influencer recommender system adopts a process of batch updating of user biographies on a daily basis, with any modified bios being adopted during the recomputation of meta-embeddings to ensure that retrieval results remain robust to conceptual drift in user profiles.

While a nearest-neighbour approach to recommendation ensures mitigation of the cold-start problem of traditional recommender systems and avoidance of spam ratings from influencing recommendations, there remain particular vulnerabilities to consider from the perspective of the accounts collected. Trend-stuffing is a recognised phenomenon on the X platform where trending hashtags included in tweets that are unrelated to the hashtag's topic, thereby ensuring increased visibility for the posted tweet [153]. It is not unreasonable to believe that the proposed influencer recommender system would be vulnerable to a similar manner of attack. For instance, if an account owner knows their account is in the recommendation database (having been contacted previously), they could stuff tweets to increase the likelihood of being recommended. If the data collection procedure and its timeframes were known, accounts could adjust their posting schedules to be included in the sampled stream. Once in the internal influencer dataset, they could then alter their biographies or content to increase their recommendation chances.

The average recommendation age of 207 days, in the absence of temporal ranking (refer to Table 4.4), warrants further discussion, as it may signify problems inherent in the data collection methodology that require consideration. The use of a sampled stream, where only 1% of twitter posts within the geographical parameters may be an influencing factor. Considering the previously outlined limitations of this data collection methodology, it is probable that once a user and their associated tweets have been included in the database, the likelihood of subsequent data updates is minimal, given the reduced probability of encountering the same user again during the collection process which is influenced not only by the 1% sampled stream, but also the requirement for tweets to be geo-tagged. This raises a significant consideration regarding whether these influencers, characterised by a relatively high average recommendation age, may have altered their thematic focus through general conceptual drift. Such shifts could potentially impact the accuracy of recommendations if the embeddings employed during retrieval fail to correspond closely with the topics currently discussed by an account.

4.6 Epilogue

In this Chapter, a novel industrial problem domain is investigated, wherein the identification of social media accounts based upon full-text queries may provide value to the publishing and media industries. The work has contributed to two concepts within this domain. Firstly, demonstrating how the adoption of meta-embeddings based on posted tweet content, and user biographies provides advantages to using Tweets or user biographies independently, when applied in a semantic search context. Secondly, how forward time-decay reduces the average age of recommendations, however confers a negative influence upon recommendation quality.

Reflecting the research questions defined, a commercially implemented framework is demonstrated with the industry partner that allows the continued collection of data for evaluation and future adaptation. By applying neural-embeddings and approximate nearest-neighbour algorithms, a framework ensures to account for the contextual semantics in any given query, thus providing a full-text capable approach when compared to conventional keyword-based approaches and addressing the RQ1.1. This can be argued as a flexible approach, in which the method for generating embeddings may be easily adapted within the proposed architecture as new embedding techniques become available. Such an approach provides a benefit to the publishing industry through the removal of time-constraint barriers involved when searching for relevant accounts on online social networks.

In addition to this, RQ1.2 is addressed through the evaluation of meta-embeddings based on tweet content and user biography. This provides improvements to model performance when compared to using tweet or biography embeddings individually, with a weighting of 0.5 indicating the best performance. By the implementation of time-decay, our framework ensures a weighting towards recency of tweet postings when performing recommendations, with a minimal impact upon recommendation quality. This achieves the objective proposed by our RQ1.3.

CHAPTER 5

Exploring Topic Modelling as a Method for the Analysis of Large Volumes of Academic Literature

This Chapter responds to GRQ2, consisting of two items of work which were produced when addressing the research question. These works are titled "Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis", available on the arXiv repository, and "Wide-scale automatic analysis of 20 years of ITS research", published in the International Conference on Intelligent Tutoring Systems.

5.1 Prologue

As outlined in the Introduction chapter, Section 1.5, there is a need within the publishing industry to address the limitations of the time-consuming nature of analysing unstructured and unlabelled data. In the dynamic landscape of the publishing industry, the production of textual media is an ongoing process. It is imperative for the industry's major stakeholders to consistently update their comprehension of the evolving trends and interconnections among various topics. This understanding is crucial to maintain their relevance and effectiveness within the industry.

This was identified as something of value to the industry, where the investigation of topic modelling algorithms was proposed to allow the comprehension of vast sources of textual media, such as social media streams and news aggregation sources, to provide a better understanding to journalists of the current scope of the media landscape. For the industry sponsor, Intellectual Property (IP) limitations through patent applications required that any academic publications produced as part of this investigation do not include the analysis or release of internal data sources. Instead, it was proposed to use academic literature in place of the social media and news aggregation sources.

In this chapter, the works produced as part of an introductory analysis of topic modelling, as a means of comprehending large volumes of (*academic*) literature, are discussed in detail. In the academic process, the comprehension and analysis of literature is essential, however, time-consuming. Reviewers may find it difficult to identify relevant literature, given the considerable volume of available texts. It is arduous not only for starting Ph.D. students, but also for any researcher learning about a new field (in Chapter 6 , these are referred to as *"domain learners"*). To address this issue, it is necessary to investigate an *automated framework to assist in the literature review process.* Therefore, this chapter and the latter outline a dual-faceted methodology, consisting of a preliminary exploratory analysis of two recognised topic modelling algorithms in Chapter 5, followed by an experimental study in Chapter 6, informed by expert feedback. This approach has culminated in the development of a comprehensive framework for analysing sources of academic literature. The research outcomes of these two chapters concentrate on academic literature as a source of data. Nevertheless, this research indicates an applicability to any source of textual data. Currently, this methodology is being further explored in a subsequent post-doctoral research project, which is conducted through the Durham Impact Acceleration Account (IAA). The aim of this research project is to extend the application of this research to social media data, a development that could have a direct influence on the publishing industry. This ongoing research underscores the value of topic modelling in transforming data analysis across various domains.

This chapter is based on work that has produced two research papers published during the thesis: "*Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis*" [154] and a "*Wide-scale automatic analysis of 20 years of ITS research*" [155]. The former initiates an exploratory study into topic modelling, applying the Latent Dirichlet Allocation (LDA) [13] topic modelling algorithm to the analysis of literature gathered from various academic databases. This is conducted in conjunction with a manual analysis, as part of a traditional systematic literature review. This dual approach facilitates a comparison between automated and manual methods, enabling the identification of potential advantages and limitations of the automated approach. The main outcomes of this begin with the addressing of RQ2 through an exploration of the feasibility of topic modelling algorithms for the analysis of large volumes of academic text. More specifically, the use of the LDA algorithm presents an analysis using a well-known and widely used topic modelling algorithm, and identification of the caveats of this technique. In the latter work, "*Wide-scale automatic analysis of 20 years of ITS research*", an alternative method for topic modelling is explored, through the cluster-based topic modelling algorithm Top2Vec [16]. The selection of Top2Vec for assisting the literature review process has had a considerable impact on the general direction of the overall thesis, where subsequent chapters will investigate the algorithm in more detail, through the provision of an automated, end-to-end framework for literature analysis. Furthermore, in the final chapter, the concept of dimensionality reduction [83], which is a vital step in cluster-based topic modelling algorithms like Top2Vec [6] is investigated, with contributions to the domain indicating the benefit of this avenue of research in topic modelling. Hence, the outcomes of this early area of the Ph.D project have led to the shaping of the general direction of research, considerably.

5.1.1 Chapter-Specific Research Questions

The following research questions outline aspects specific to the current chapter. Given Generic Research Question GRQ2: *What AI techniques can be defined to assist readers in the comprehension of journalistic literature?*, the following specific research questions are outlined:

- **RQ2.1:** *How can Existing Topic Modelling Technologies be adapted to facili-*

tate a semi-automated analysis of academic literature?

- **RQ2.2:** *How can the semantic relationships between topics in literature contribute to understanding of topics during exploratory literature analysis?*

- **RQ2.3:** *How can analysing the temporal trends of identified topics in literature contribute to understanding of topics during exploratory literature analysis?*

The first RQ outlined for this section of the project investigates the feasibility of using existing topic modelling algorithms for the analysis of large volumes of academic literature. As discussed above, two algorithms are investigated as part of this, with the traditional and widely adopted LDA [13] algorithm being investigated in "*Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis*". In the second work of this chapter, "*Wide-scale automatic analysis of 20 years of ITS research*", the more recent Top2Vec [16] is considered, based on the identification of several benefits which make it a more robust algorithm for the problem domain. The outcomes of both items of work presented in this chapter show the efficacy of topic modelling algorithms for analysing large-scale literature collections. The next chapter will extend this approach by applying experimental methods to evaluate an expanded framework for literature analysis.

In terms of **RQ2.2**, this was formulated based on the assumption that the topics identified within the literature are semantically distinct. This, in turn, would allow the use of semantic similarity measures as detailed in Section 3.2.1 to illustrate the semantic connections between topics. Such an approach can aid in detecting connections within the literature, thereby enhancing the literature review process. This is addressed in "*Wide-scale automatic analysis of 20 years of ITS research*", through an analysis of the semantic similarity of embeddings generated based on identified topics within the literature. In the next chapter, this is further expanded upon, through experiments conducted in conjunction with domain-experts specific to the biomedical field.

Finally, **RQ2.3** is focused on the hypothesis that temporal metadata from publications can facilitate analysis of temporal trends in topics identified within the academic literature. This is initially explored in "*Arabic Text Sentiment Analy-*

sis: Reinforcing Human-Performed Surveys with Wider Topic Analysis", through a rule-based analysis of the occurrence of individual words, based on their occurrence within the data. In the subsequent work titled " *Wide-scale automatic analysis of 20 years of ITS research*", this concept was expanded to facilitate the temporal analysis of entire topics within the literature.

5.1.2 Research Objectives

Based on the above motivation and research questions, the research objectives for this Chapter are:

- **RO2.1:** To investigate the traditional, classical topic modelling technique of Latent Dirichlet Allocation as a method for the identification of topics within academic literature. (Addressing RQ2.1).
- **RO2.2:** To investigate cluster-based topic modelling techniques and compare these with classical topic modelling algorithms. (Addressing RQ2.1).

5.1.3 Chapter Contributions

The main contributions of this Chapter can be summarised as follows. (1) A comparative investigation of two algorithmically different approaches to topic modelling, namely, LDA and Top2Vec, from a quantitative and qualitative perspective. In evaluating quantitatively, established measures of topic quality provide an initial overview of how both models compare with regard to these metrics. Following this, an in-depth, qualitative assessment of the topics based on human-judgement compares topic terms with regard to similarities and differences between both models. Considering the primary objective of this study, which is to explore the automated analysis of unstructured texts, which in a publishing context could involve diverse data sources depending on the investigation domain, these analyses demonstrate the advantages of Top2Vec. (2) The demonstration of downstream visual analyses of topic modelling results with regard to the semantic relationships between topics, which could be adapted to a publishing context to provide journalists with an overview of news subjects they may be investigating, or provide an understanding of the semantic relationships between various topics being discussed on social media platforms. (3) The visualisation of temporal trends in identified topics, which from an academic perspective enables researchers to understand whether certain domains have observed an increase, or decrease in publication volume. Analogously, within a publishing framework, this approach could prove beneficial to journalists when applied to social media or news datasets. It allows for the identification of extensively discussed topics, potentially increasing readership should content be created on these subjects. Conversely, it facilitates the detection of over-saturated news topics, thereby advising journalists against producing further content on such subjects.

5.2 Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis

As an exploration of existing topic modelling techniques and their value to literature analysis, an automatic collection of literature in the Arabic Sentiment Analysis (ASA) domain was conducted, using the CrossRef API[1], Elsevier API[2], Springer API[3], Wiley API[4], Core API[5] and ArXiv preprints repository[6], for an in-breadth study of the ASA field. The search terms used during this collection process were: *"Arabic sentiment analysis"*, *"Arabic semantic analysis"*, *"Arabic subjective analysis"*, *"Arabic emotion detection"*, *"Arabic text categorisation"*, *"Arabic opinion mining"*, *"Arabic lexicon"*, *"Arabic corpora"*, *"Arabic sentiment analysis"*, *"Arabic sentiment classification"* and *"Arabic Opinion Mining"*. From these searches, a total of 53,405 potentially relevant scientific articles were identified prior to filtering. Given the range of sources used for data collection, there was a large degree of varia-

[1] https://www.crossref.org/education/retrieve-metadata/rest-api/text-and-data-mining-for-researchers/

[2] https://dev.elsevier.com/

[3] https://dev.springernature.com/

[4] https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining

[5] https://core.ac.uk/services/api/

[6] https://pypi.org/project/arxiv/

tion in the formatting of the full-text publications. Filtering was performed on the identified publications at the title and abstract level, to ensure that only relevant documents were used. This was facilitated through conditional Boolean searches of the identified texts, using the search terms used at the API search level. Following filtering, the total number of relevant publications was reduced to 2297. While some platforms provided results in the standard JATS XML format, many returned full-text results in the form of PDF documents which required further conversion into machine readable XML. Following the collection of the wider literature, the Latent Dirichlet Allocation [13] algorithm was applied in the analysis of the literature, in the same manner as the framework presented by [156]. With regards to LDA, it is essential to identify a suitable hyper-parameter configuration in order to ensure a suitable topic modelling solution can be obtained. In their study on smart literature review using LDA [156], the authors employed a comprehensive search approach by first defining the number of topics to be identified and then assessing this over various topic numbers by comparing a topic coherence score. For the work proposed in this section, this process was carried out in the same way as the approach described in [156], known as the "elbow" method, where the coherence score is plotted against different topic numbers to determine the point at which the curve shows the sharpest turn, indicating the optimal configuration for the number of topics present in the data. For this study, the optimal number of topics identified based on this 'elbow method' was 201 topics, which resulted in a C_V coherence [157] score of 0.39. Following the identification of topics within the corpus, a temporal analysis of the changes in topics throughout 2010-2020 is conducted, and evaluated with regard to the findings of the manual survey review.

5.2.1 Topic Modelling of Wider Literature

Given the high number of topics identified by the optimal model identified by exhaustive search and the elbow method, the ten topics deemed most relevant to the study of the ASA literature are presented in Table 5.1. These demonstrate several distinct areas related to model architecture, including transformer learning **(T1)**, Bayesian learning **(T5, T3)** and recurrent and convolutional networks **(T6)**. Additionally, the identified distributions place Twitter and Facebook terms into separate topics. There is still, however, a degree of noise present in the topic-word distributions, with some topics containing vague terms (i.e., challenge, multi, support, sad).

Table 5.1: Ten relevant topics to the ASA literature analysis identified by LDA topic model.

Topic No.	Topic-Words of Relevant Topics
1	arabert, challenge, wordnet, bert, transformer
2	facebook_search, reactions_love, users_trace, sad
3	named_entity, classifier, svm, naive_bayes
4	subjectivity, twitter_feed, supervised, classify
5	gram, multi, mwt_extraction, bayesian, naive_bayes
6	cnn, convolutional, algorithm, lstm, word_embedding
7	code_switching, bayesian, naive_bayes, classifier, complex
8	stemmer, stem, compound, lexicon, text_mine
9	query, sparql, support, rdf, owl
10	finite_state, tree, decision, topic, lsi

Before performing a temporal analysis of the identified word distributions, it was necessary to filter out unrelated or vague terms and, in some cases, provide new terms. Topics that showed a high degree of noise or were not particularly relevant were removed at this stage. The final selected topic words are presented below along with the manual deductions of their ideal label in Table 5.2.

Table 5.2: Topic Words present after filtration of unrelated and vague terms used in temporal analysis.

Topic No.	Topic-Words	Human Labeled Topic
1	Arabert, bert, transformer, gigabert transfer_learning	Transformer Networks
2	CNN, convolutional	Convolutional Networks
3	Lstm, long short term memory, rnn, recurrent	Recurrent Networks
4	Bayesian, naive_bayes, bayes	Bayesian Learning
5	Sparql, rdf, owl	Semantic Web
6	Twitter, tweet, retweet, twitter_feed	Twitter
7	Facebook, facebook_post, status_update	Facebook

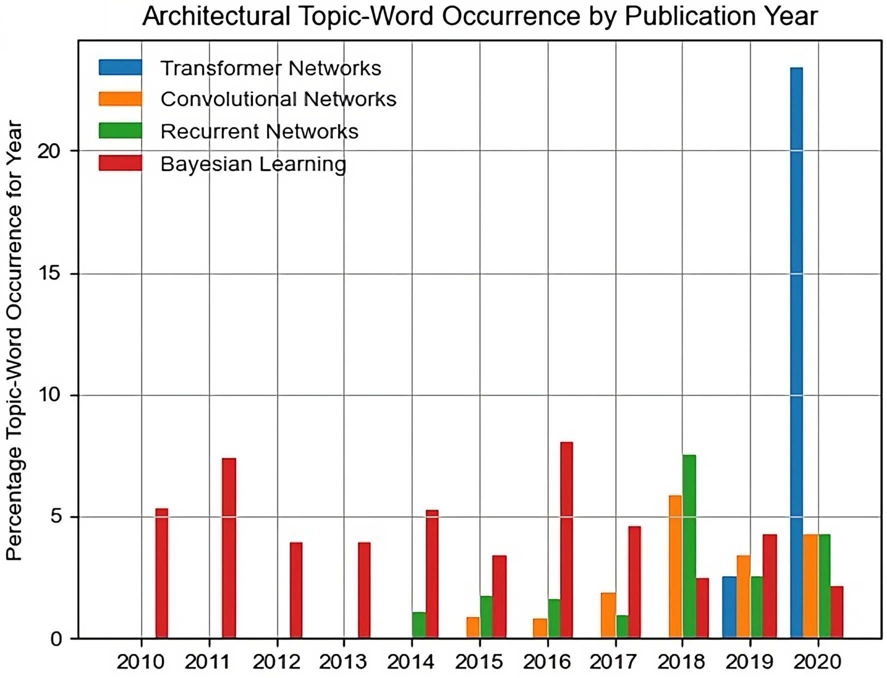


Figure 5.1: Proportion of architecture topic-word occurrences per year in automatically collected literature

5.2.2 Temporal Topic Analysis

Following the distillation of suitable terms and filtering of relevant topics a temporal analysis of the automatically identified literature was performed. Figure 5.1 presents a temporal analysis of model architectures in the ASA domain in the period from 2010 to 2020.

The temporal topic analysis indicates that Bayesian learning is frequently mentioned in publications throughout our sample, which reflects our identification of several Bayesian classifiers in Appendix IV of the human-performed survey [154] (this was not included, as the survey falls outside the scope of the thesis). Furthermore, the proportion of Bayesian learning in the sample is reduced in years from 2018-2020 in conjunction with the increase in both recurrent and convolutional network occurrences, reflecting the identification of deep learning approaches in the human performed survey [154]. By acknowledging both convolutional and recurrent network approaches as separate topics we can observe the differences in these. However, there appears no conclusive trend between these in our data. Our anal-

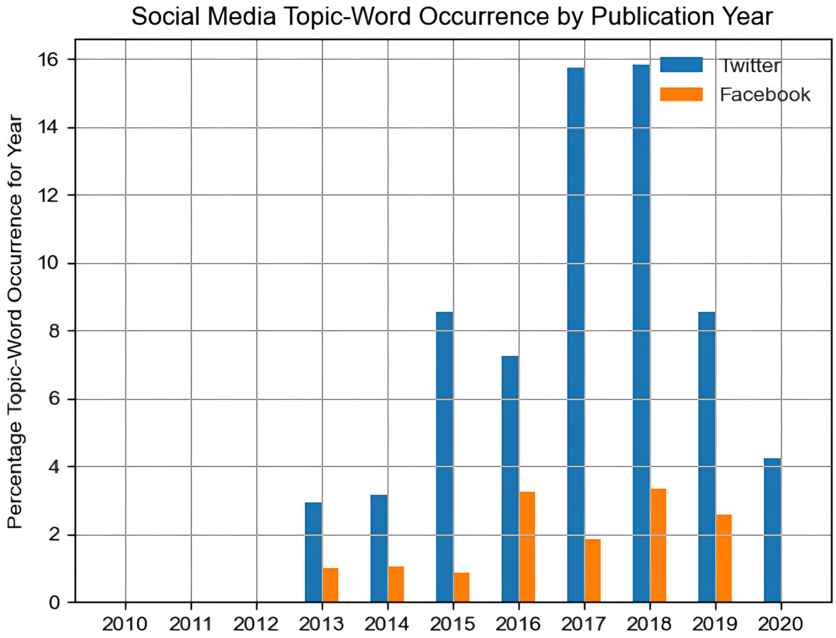


Figure 5.2: Proportion of social media topic-word occurrences per year in automatically collected literature

ysis of the automatically collected literature identifies transformer networks being present in ASA from 2019, again reinforcing trends identified in the human performed survey [154]. However, we identified a significant increase in the proportion of literature discussing transformer networks, with over 20% of the corpus mentioning transformer networks in 2020 for our automatically collected data.

Furthermore, it was apparent that there is a considerable difference between the occurrences of Facebook and Twitter topic within the automatically collected literature (Figure 5.2), with Twitter topic-words being present in more than 15% of the publications for 2017 and 2018. This is reflected in the manual study, where a significant proportion of corpora are tweet-based [154]. The prevalence of Twitter within the corpus may be due to the availability of access to real-time streaming data using the Twitter Developer API to allow for collection of tweet data (*Note: The availability of data collection for researchers on Twitter has changed considerably since the period when this study was conducted*[7]). Pre-processing steps taken to ensure fine-grained topic distributions required the removal of frequent terms from

[7] `https://developer.twitter.com/en/use-cases/do-research`

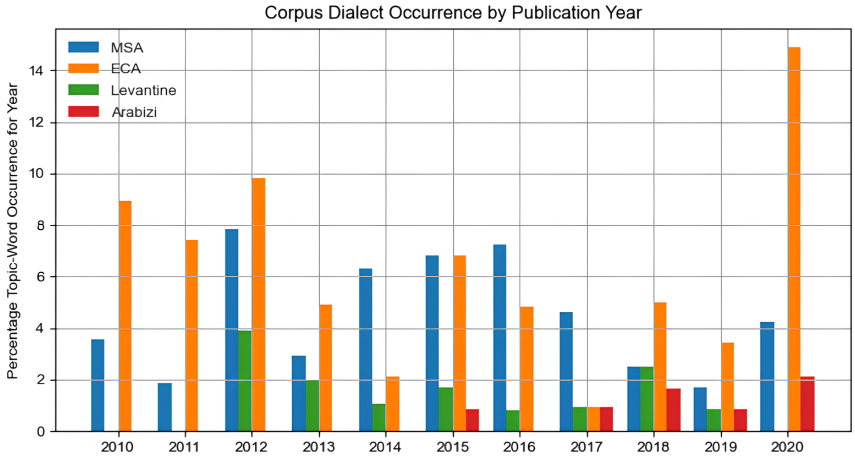


Figure 5.3: Proportion of dialectal topic-word occurrences per year in automatically collected literature

the corpus prior to applying LDA [156]. The process of removal of overly frequent terms however, led to the removal of some useful terms, particularly with dialectal terms and abbreviations (e.g., MSA, ESA). Furthermore, we performed a temporal analysis of the presence of dialectal corpora within the literature, presented in Figure 5.3.

Analysis indicates that MSA (Modern Standard Arabic) and ECA (Egyptian Colloquial Arabic) are the most common dialects used in corpora. The trends identified in our analysis of the wider literature reflects the prevalence of MSA and Egyptian dialectal corpora in the ASA domain, as found in our study. Additionally, the indication that the Arabizi dialect is present in years 2015-2020 may demonstrate a novel area of ASA research.

5.2.3 Discussion of Findings

The main objective of this study, carried out in the early stages of the Ph.D project, was to explore the feasibility of applying existing topic modelling techniques to the task of contributing to the literature review process. This was formed with a focus on the traditional, or classical, technique of Latent Dirichlet Allocation (LDA). By conducting this automated analysis of the literature, we have reinforced the findings of the initial review (this content was not incorporated into this thesis because it is beyond the scope of the thesis, but is available for reference [154]), through a second analysis of Open Access publications, on a larger scale than manually feasible, through topic modelling and subsequent temporal analysis. The findings of this study demonstrate the potential value of topic modelling and subsequent topic analysis to enhancing the literature review process, however, limitations were prevalent based on the algorithmic choice of LDA as a topic model. In particular, the exhaustive evaluation of multiple LDA models with different hyperparameters means that the technique is not suitable when trying to address the overarching goal of creating an automated literature analysis framework.

5.3 Case-Study Comparison of Cluster-Based Topic Modelling with Latent Dirichlet Allocation for Assisted Literature Analysis

Building on the findings of the previous work, which served as a surface-level exploratory investigation into using the classical LDA topic modelling technique for literature analysis, a subsequent study is proposed, aimed at evaluating recently proposed cluster-based topic modelling techniques and how they perform when applied to the literature analysis task. Within the literature, an existing work presented LDA as a topic model for the analysis of academic literature [156]. However, there are limitations to this algorithm which make the analysis difficult. Most notably, when using LDA, there is a requirement for prior specification of the number of topics to extract. In tasks such as identifying topics in literature, the number of topics is not known a priori. Therefore, to determine the correct number of topics for optimal results, an exhaustive search must be performed using the ‘elbow method’. For this, exhaustive evaluation of a range of predefined topic numbers needs to be performed and evaluated using relevant evaluation methods such as topic coherence [158, 159], where the optimal number of topics is identified by evaluating the point at which a line graph of the evaluation metric relevant to the topic number observes a sharp turn, analogous to the curve in an elbow. Such an exhaustive search to identify the optimal configuration for a topic model such as LDA is not suitable in cases where the results must be presented in a reasonable amount of time, such as when deployed in industry, and it is therefore necessary to consider alternative methods for topic modelling. Based on these criteria, we adopt a subsequent study into the use of the Top2Vec algorithm in a more in-depth manner is presented in "*Wide-scale automatic analysis of 20 years of ITS research*" [155], which was published in the Intelligent Tutoring Systems (ITS) conference in 2021. This is presented as a case-study analysis of the breadth of ITS research using a further development of the process explored in Section 5.2. To ensure direct comparison with LDA and reinforce the hypothesis that cluster-based topic models can mitigate the aforementioned issues with LDA, we perform a comparison upon the same corpus of intelligent tutoring systems research, from the perspective of both qualitative and quantitative measures.

The considerable volume of research within Intelligent Tutoring Systems (ITS) presents challenges to the quantification of the various fields present within the domain. Conventional literature surveys are typically performed using manual analysis and filtering of available literature and, as such, are limited in the volume of publications. Moreover, researchers may inadvertently overlook studies that hold significant importance. Surveyors are furthermore likely to include main-stream research only, or research assisting in their argument. Thus, it was proposed to leverage the novel topic modelling algorithm Top2Vec [16] for the analysis of a large volume of ITS research. Advantages to such an analysis include the volume of ITS research processed, which exceeds that which may be feasible even by a large team of contributors. Additionally, the speed of topic analysis ensures ample time for the further analysis of temporal factors within the corpus and presentation of relationships between any identified topics.

The major contributions of this work are: 1) automatically identifying, for the first time, significant trends in ITS (e.g., temporally, ITS has observed a significant shift in research popularity from Adaptive Hypermedia towards online MOOC platforms; applied architectures and algorithms have shifted significantly towards Deep Learning and applications of Neural Networks); 2) automatic extraction of relationships between several ITS topics indicates potential for novel areas of research; 3) the demonstration of the potential of the recent (at the time of publication) Top2Vec [16] algorithm to assist, for the first time, in large-scale literature analysis, without limitations presented by conventional probabilistic topic models. Compared to existing studies within ITS, this research provides a unique overview of the last 20 years of research, without the bias presented by human reviewers, who may arguably 'cherry pick' studies to argue their point.

5.3.1 Introducing Cluster-Based Topic Modelling

Traditional literature surveys in ITS follow a manual process, in which identified publications are filtered, resulting in a significantly smaller batch of publications used in the final review. ITS reviews, such as [160], apply a Systematic Literature Review process, with the study mentioned analysing a total of 33 publications, filtered from an initial corpus of 4,622 articles. These resulting papers are analysed in-depth; however, the exclusion of such a large volume of papers clearly indicates missed opportunities for obtaining insight from the excluded publications. Outside of manual literature surveys, we identified [161], which performed analysis of a larger volume of publications on a quantitative level. The scope of the publication addressed the barriers and trends of ITS adoption rather than the trends and relations of the overarching field. Topic modelling of unstructured data has been performed through Bayesian models including Probabilistic Latent Semantic Analysis (PLSA) [162], which performs probabilistic modelling of latent variables, and the subsequent Latent Dirichlet Allocation (LDA) [85], which is included in this study for direct comparison. These approaches require a definition of the number of topics prior to generating topic distributions, which requires a prior understanding of the ideal topic number. When the number of topics is not known, a search process is involved. Methods to determine the number of topics include exhaustive searches, which evaluate coherence [157] or perplexity scores on a range of topic numbers, optimisation using genetic algorithms [163] or individual intuition. Hierarchical Dirichlet Processes (HDP) [164] may mitigate this through automatic identification of topic number; however, pre-processing steps are required, as with other probabilistic bag-of-words models, to filter stop words to improve model performance. Top2Vec [16], in contrast, provides a very recent alternative to Bayesian topic models such as PLSA and LDA [13, 162], eliminating the need for pre-defining the topics number and filtering stop words. The algorithm leverages language embedding methods and enables pre-trained language models to be applied via Doc2Vec [2], BERT [26] or Universal Sentence Encoder [165]. As defined in Section 3.2.3, a semantic embedding of joint document-word vectors is generated where the distance between document and word vectors represents semantic association. This ensures that semantically similar documents achieve a smaller distance between each other when compared to dissimilar documents. Resulting embeddings are clustered using the HDBSCAN [166] algorithm into topic clusters, with the hierarchical nature of HDBSCAN ensures automatic identification of the topic number. Given the high dimensionality of document embeddings, clustering requires prior reduction of dimensionality of embeddings through the UMAP [126] algorithm. Top2Vec has been demonstrated to outperform LDA and PLSA when applied to the benchmark 20NewsGroups [167] dataset [16].

5.3.2 Corpus Generation

For collection of publication data from the ITS domain over the past 20 years, several API resources were used. These consist of the arXiv Preprint Repository[8], Springer API[9], SAGE API[10], Elsevier API[11] and CORE API[12]. Query terms to be used in the API search were taken from a sample of key phrases presented in the 2000-2020 ITS Conference Proceedings; however, inclusion of low-level specific terms was avoided to ensure the resulting document distributions were unbiased. To be comprehensive, the collected literature was not limited to journals and conferences, but also included book chapters, pre-prints and academic theses. In total, 5018 documents were identified from 2000-2020. Research from 2000-2020 was selected to provide

[8] https://arxiv.org/help/api/index

[9] https://dev.springernature.com/

[10] https://journals.sagepub.com/page/policies/text-and-data-mining

[11] https://dev.elsevier.com/

[12] https://core.ac.uk/services/api/

a larger-scale analysis, beyond that of only the most recent literature, which could assist in evaluating the changes in long-term trends of ITS quantitatively. Following the collection of raw publication data, it was necessary to filter out results that were not relevant to ITS using Boolean word matching at the abstract level. Given that some terms used within ITS (e.g., adaptive learning) may be confused with general machine-learning terms by a partial matching system, it was deemed necessary to apply absolute string-matching during filtering. Documents were excluded if they failed to contain any instances of the terms within the original search query. This rule-based prefiltering of the corpus entailed the application of regex queries to ensure string-matching of the original search queries within collected literature. For an article to pass the filtering process, at least one of the original search terms had to be found within either the article title or abstract. For example, a publication titled "Gender issues in computer-supported learning" contains a case-insensitive match to the original query term "computer-supported learning", and thus was not discounted from the corpus. Regex filtering of abstracts is also valid, but an example is not provided in this Chapter, for brevity. All searches were performed in a case-insensitive measure. Before filtering, there were 5018 articles with 4721 remaining after filtering based on this method. Language checks were performed on the corpus, to remove any non-English publications, which could impact model performance. For this, the langdetect[13] Python library was applied, excluding documents from the corpus if they are not identified as belonging to the English language. Following filtering of nonrelevant results, the corpus size was of 3898 document abstracts and titles, which were combined for the subsequent analysis. Although not a large corpus when compared to benchmark datasets used in topic modelling such as 20 Newsgroups (18,846 documents) [167] or AG News[14] (127,600 documents), this still well exceeds the volume of research that can be manually analysed by a single researcher, where the median number of documents used in the software engineering field was reported to be 57 [168].

[13] https://github.com/Mimino666/langdetect

[14] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

5.3.3 Corpus Preprocessing and Data Preparation

Preprocessing for LDA

As stated prior, the bag-of-words technique of LDA requires extensive '*classical*' preprocessing of a corpus prior to analysis. This entails several distinct processes, which are detailed hereafter, and generally follow the practices of similar lines of research [156].

A general first step is the removal of non-value-adding words or phrases from the corpus text. This involves the conversion of all words to lower-case, and removal of punctuation, special characters, URLS, and email addresses, if applicable. After this, stopword removal performs the removal of non-semantic contributing terms such as "can", "the", and "were", to name a few. A pre-defined lexicon of stopwords is applied through the NLTK[15] Python library, which provides an extendible list of stopwords to use during this process. As LDA is a probabilistic bag-of-words algorithm, and therefore does not account for contextual of how words are used in a document, the failure to remove stopwords will lead to these non-semantic contributing terms being accounted for in topics which may lead to sub-optimal word distributions in topics.

It is also necessary to conduct lemmatisation of words to their base terms [169], as otherwise words with the same meaning used in different phrasiology, such as "analyse" and "analysed", or "flood" and "flooding" would have individual topic probability distributions during analysis with LDA. For this process, the Spacy[16] Python library is applied, focusing on nouns, adjectives, verbs, and adverbs.

In some cases, it may also be beneficial to the analysis to calculate topics based on n-gram phrases, where an n-gram is a valuable method in ensuring to capture the ordering of multiple words in a corpus. In natural language processing (NLP), an n-gram is a contiguous sequence of n items from a given sample of text or speech [170]. These items can be phonemes, syllables, letters or words. For this research, we focus on contiguous sequences of words of length one (unigrams), two (bigrams),

[15] `https://www.nltk.org/`

[16] `https://spacy.io/`

and three (trigrams). N-gram sequences for the corpus are generated using the Natural Language Tool Kit (NLTK) [171]. Following these preprocessing steps and the identification of n-grams, the processed corpus is then applied to LDA across a range of predefined number of topics evaluated based on C_v coherence, as detailed in Section 5.3.4.

Preprocessing for Top2Vec

Unlike traditional topic models, there is no need for extensive pre-processing steps when performing topic modelling with the Top2Vec algorithm. In the original work that introduced the Top2Vec algorithm, contextual information encoded within the sentence structure is outlined as valuable to the embedding models present within the Top2Vec algorithm. Stop word removal, lemmatisation, or stemming is therefore not necessary, as detailed in [16], in contrast to LDA, where stop word removal and additional filtering of highly frequent terms is commonly performed [172] with the aims of improving model performance. Given the limitations of access to publications, it is only possible to perform analysis upon publication abstracts, as access to full-text publications is not provided by some of the platforms from which the data was collected.

5.3.4 Analysis

The methodology for the literature analysis task involves the modelling of topics within the corpus of academic publications using the Top2Vec algorithm [16] and Latent Dirichlet Allocation [13], and the subsequent analysis of the resulting topics in relation to the temporal range and the relationships between topics.

Analysis through Latent Dirichlet Allocation

A significant caveat when using LDA is the necessity to specify the number of topics. Determining the optimal number of topics can be challenging, with existing techniques involving the evaluation of measures of topic '*goodness*' through coherence scores, similar to that performed in the existing smart literature review research [156]. To determine the optimal configuration of topic numbers, we apply an evaluation of C_v coherence measure across a range of 1 to 301 topics, reported based on the average score for 10 individual iterations of each topic number experiment. The upper-bound for this evaluation was established based on the prior study of ASA literature, where it was found that 201 topics was the optimal number of topics. The results of this evaluation are presented in Figure 5.4.

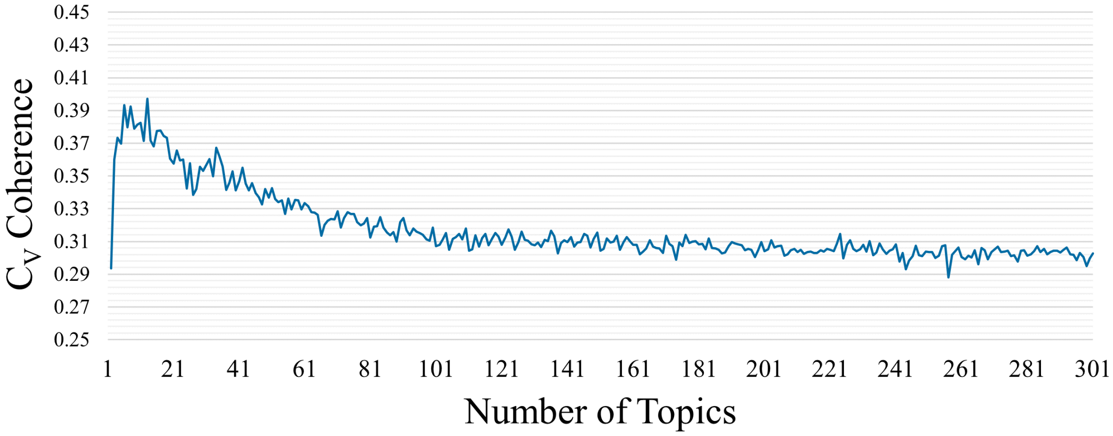


Figure 5.4: Average coherence score based on 10 individual iterations for LDA ranging from 1 to 301 topics.

Based on this evaluation of coherence measure, there is a general improvement in topic coherence up to a topic size of 12, beyond which there is a general negative trend as the number of topics increases. This indicates that the optimal topic number configuration for the corpus based on C_v coherence measure is 12 topics, attaining a score of 0.397.

Analysis through Top2Vec

Top2Vec identifies semantic relationships by learning a distributed representation via the Doc2Vec algorithm [173]. Alternatively, pre-trained models can be applied including Universal Sentence Encoder [165] or the BERT [26] transformer network. However, for this work, it was identified that applying Doc2Vec embeddings achieved the optimal human-readable topics, when compared to the topic modelling solution produced when leveraging the BERT pre-trained language model. This is for a number of reasons, which can be summarised as the problems of variable length text and out of vocabulary terminology, which are discussed below.

In the first case, Doc2Vec provides the unique advantage of allowing the vectorisation of variable-length texts, ranging from individual words up to entire documents [173]. In comparison, at the time at which the subject work of this aspect of the Ph.D project was produced, the recently proposed BERT algorithm was limited to a contextual window of 512 tokens. It is worth noting that in this case, tokens may not always represent a full word, but instead full words can be broken into multiple tokens, hence the word *'sleeping'* may be tokenised as *'sleep'* and *'##ing'* [40]. Thus in cases where a tokenised representation of a document is greater than 512, any semantic and contextual information which falls after the 512th token will not be accounted for when generating the embedding. Thus, it was deemed necessary to consider the eventuality that some documents within the analysis may be longer than this 512 token limit, and it is therefore necessary to adopt an embedding technique that may handle variable-length documents.

Secondly, it was necessary to take into consideration the domain-specific language that may be present in the niche domain of research on intelligent tutoring systems. In the case of Doc2Vec, the training objective is loosely based on the prediction of the next word in the sequence, given the surrounding language [173], and is discussed in detail in Section 2.1. For this study, the Doc2Vec model is trained upon the corpus of ITS research and will therefore be able to account for the domain-specific terminology present within the corpus. In comparison, the BERT embedding models are pre-trained, with the corpus for training being conducted based upon the BooksCorpus [174] and English Wikipedia datasets. Thus, it is possible that the domain-specific language present in academic literature, especially in niche domains, may not have been encountered or are under-represented in the pre-training corpus for BERT, and therefore the embeddings produced may fail to account for the semantics of domain-specific language. The details of this pre-training procedure is presented in Section 2.1.

In order to mitigate the limitation of domain-specific language, several options are available, which could contribute to enhancing the quality of the topic modelling solution when using pre-trained language models, which are beyond the current remit of this chapter. One such method is the fine-tuning of pre-trained language models on a domain-specific dataset, in order to better capture the semantics of domain-specific language and thus produce embeddings of a higher quality, which can then be clustered to identify topics. One method of achieving this is through the fine-tuning of pre-trained language models on sentence-similarity tasks, which is best represented by the work of [33]. With this approach, a siamese transformer network is defined, consisting of two identical transformer models. Pairs of sentences can then be passed as inputs to the respective transformer model, with a prior label or semantic similarity score denoting the degree of similarity of the sentences being used in calculating the loss using categorical crossentropy if labels are provided, or mean squared error if scores are provided. This permits the proposed siamese network to update weights during training in order to best produce emebddings that are semantically meaningful. It is worth noting that this fine-tuning procedure requires the provision of both sentence pairs, and a predefined label or semantic similarity score, and thus implies the necessity for a curated dataset, which may not exist for niche domains such as the intelligent tutoring systems corpus. Alternatively, existing research has already contributed to domain-specific language models such as SciBERT [175] and BioBERT [80] wherein a model is pre-trained upon a multi-domain corpus of scientific publications for the former, or biomedical corpora for the latter. In both cases, the pre-training of these models on a domain-specific corpus demonstrates advantages in downstream analyses such as classification, indicating that the semantic comprehension of these models is improved in their respective domains. Unlike the fine-tuning of existing models, the pre-training procedure is an unsupervised learning technique, discussed in Section 4.3.3, and therefore does not require labelled data, which was necessary in the simese-network training procedure of [33], instead needing only a sufficiently large corpus. Thus, it is evident that the adoption of either fine-tuning of existing models, or the pre-training on specific corpuses could enhance the quality of the topic modelling solution on niche corpora by improving the quality of embeddings.

The consideration of accounting for domain-specific language is useful not only in the context of academic literature, but also the wider publishing industry. Journalistic content can cover an innumerable range of topics based on the general subject

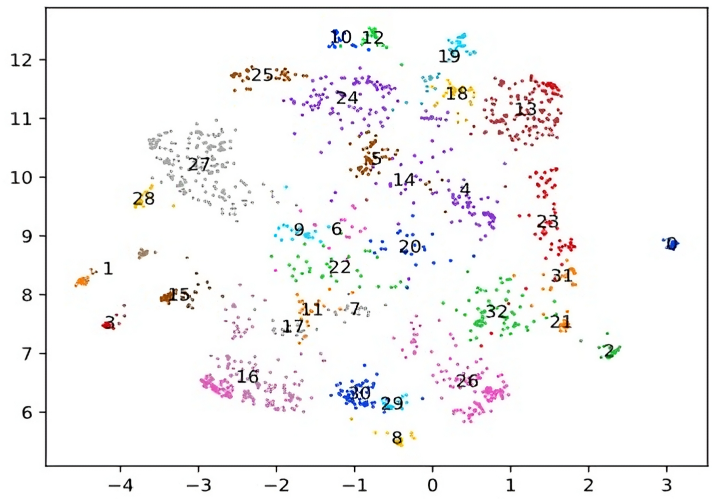


Figure 5.5: Clustering results following noise removal with topic labels assigned.

it is addressing. One example of this could be journalism focusing on a specific scientific domain and, therefore, will be prone to frequent use of niche language, jargon, and acronyms. In these instances, the adoption of embedding methods that can account for, or learn, relationships between domain-specific language based on their context is valuable in ensuring to capture a sufficient amount of information for subsequent clustering during topic modelling. These represent interesting avenues for further research.

After performing topic modelling with Top2Vec, the resulting clusters can be visualised to provide an understanding of the clustering results as presented in Figure 5.5. HDBSCAN [79] automatically labels sparse areas of embeddings within the corpus as noise, which were removed before presentation in Figure 5.5, leaving all topics without any noise present. Given that ground truth labels were not available for the evaluation of clustering, we applied the Silhouette score [176] using the Euclidean distance and achieved a score of 0.37 when accounting for all 33 topics. When reducing this to only account for topics relevant to ITS, this increased to 0.42. In total, HDBSCAN identified 33 separate topics.

For higher accuracy, we further performed manual evaluation of the resulting topics at a qualitative level and filtered nonrelevant topics which may have arisen, to ensure that only those relevant to ITS remain. Filtering consisted of analysis of the topic-word distributions for each topic, and exclusion was performed when a significant level of noise or noninformative terms were identified. A label was manually assigned to relevant topics, based on the word distributions they entailed. These are detailed in Table 5.5. The results indicate 11 highly coherent and relevant topics out of the 33 topics identified by Top2Vec. Topics were excluded where word-distributions contained unrelated terms and could not be clearly labelled, or when the distributions were related, however, contributed less to our analysis. The removal of nonrelevant documents and noise reduced the total corpus size for our analysis to 1223 documents. Topic 13 indicates the types of architectures and models present within research and, as such, does not serve as a useful topic label for the corpus. We investigate this separately through a temporal analysis in Section 5.4.2.

5.4 Results

5.4.1 Quantitative Comparison of LDA and Top2Vec

Although the algorithmic approaches to topic identification in text differ significantly, it is essential to conduct a quantitative evaluation of both techniques against a shared benchmark. Accordingly, a comparison is undertaken using the C_V coherence measure, normalised pointwise mutual information (NPMI), and the number of topics identified. These measures are introduced and discussed in Section 7.6.3, where they are used in other analyses of topic models. It is important to note that while these measures are valuable in gauging the 'goodness' of a topic model, they do not directly correlate to human judgement and it is essential to take this into consideration during the analysis [177]. The analysis is performed using the same corpus of ITS literature for both experiments while adhering to the different preprocessing procedures detailed in Section 5.3.3. All reported values are taken on the basis of an average of 100 individual iterations of each experiment.

Based on this analysis, it is clear that from a quantitative perspective, Top2Vec attains greater C_V and NPMI scores compared to LDA, indicating better coherence and topic quality. In the case of C_V, an increase of 0.194, and for NPMI 0.102. Furthermore, the identification of an average of 32.853 topics could indicate

Table 5.3: Reported evaluation metrics for Latent Dirichlet Allocation and Top2Vec based on an average of 100 iterations.

Topic Model	**C$_V$**	**NPMI**	**Number of Topics**
LDA	0.397	-0.123	12
Top2Vec	**0.591**	**-0.021**	31.853

a finer-grained topic modelling solution compared to 12 topics for LDA. This holds significant importance for literature analysis, where conducting subsequent analyses, such as those specified in Sections 5.4.2 and 5.4.2, on niche and specific literature topics would yield more meaningful insights into the trends and relationships of these subjects.

5.4.2 Qualitative Comparison of LDA and Top2Vec

Although quantitative coherence metrics provide valuable insights for model selection, specifically indicating the advantage of Top2Vec, it is important to recognise that these metrics are statistical in nature and do not necessarily align with human judgment. Consequently, it is worth conducting a qualitative review of the topic modeling solutions. Therefore, an analysis of the topic quality is undertaken by examining the topic-word distributions identified by both models.

Topics Identified by LDA

Topics identified by LDA analysis at the optimally configured number of topics of 12 are presented in Table 5.4. At a glance, it is apparent that the terms extracted from the corpus have some relevance to the overall intelligent tutoring domain. For each of the 12 topics, we now present a qualitative assessment of these topics based on the provided topic-terms.

Topic terms for the two largest topics (topic ID 0 and 1), indicate a description of the general ITS topic, with the terms ‘system’, ‘tutoring’, and ‘intelligent’ being the three terms reported with the highest probability of belonging to both topics. These are ommited from the table, where we present only topics deemed valuable to the analysis. Given the original search and filtering criteria for the corpus detailed in Section 5.3.2, where search criteria was specified based on keyphrase terminology

Table 5.4: Identified topic-word distributions by LDA. Topics deemed relevant to ITS by qualitative analysis.

Topic Terms	Topic ID	Topic Label
education, technology, emerge, school, distance, professional, organization, technological, trend, university	3	Educational Institutions and Organisations
elearner, social, collaborative, discussion, influence, collaboration, team, motivation, computer_supported, complexity	4	Computer Supported Collaborative Learning
adaptive, web, system, hypermedia, content, educational, style, learner, game, adaptation	5	Adaptive Educational Hypermedia
course, online, engagement, participant, video, instructor, moocs, mooc, promise, early	6	MOOCs
agent, pedagocical, affective, interaction, conversational, animated_pedagocical, body, vocal, simulation, movement	7	Pedagogical Agents
assessment, answer, dialogue, text, natural_language, question, essay, automatic, writing, scoring	8	Automated Assessment
emotion, emotional, description, modelling, state, hence, emotion_recognition, personality, computational, facial	9	Emotion Recognition
personalise, peer, recommend, portal, usuality, recommender, failure, partner, augmented_reality, tutee	10	Recommender Systems

used in conference proceedings for the 2000-2020 ITS Conference Proceedings, it is understandable that a significant proportion of the corpus will discuss this subject. Due to their generality, these topics are deemed to have limited value in literature analysis. In contrast, the subsequent, more narrowly defined topics offer substantial value for analytical purposes.

Topic 3 – Educational Institutions and Organisations: This topic, the largest topic that is deemed to be of value in literature analysis presents terms which could be interpreted as belonging to a topic of educational technologies in educational institutions. The presence of terms such as ‘school’, ‘organization’, and ‘university’ suggests a connection to institutional organizations. Additionally, the term ‘distance’ may indicate a focus on distance learning within educational institutions. However, the precise nature of this topic remains open to interpretation.

Topic 4 – Computer Supported Collaborative Learning These topic terms can be inferred as relating to Computer Supported Collaborative Learning (CSCL). This is best highlighted by the bigram ‘computer supported’, and the terms ‘collaborative’ and ‘collaboration’.

Topic 5 – Adaptive Educational Hypermedia (AEH) These terms seem to pertain to adaptive educational hypermedia, generally represented by ‘adaptive’, ‘adaptation’ and ‘hypermedia’. However, some terms are more vague, but may be interpreted in the context of AEH such as ‘web’ and ‘content’, where online media may be adapted based on the educational needs of users.

Topic 6 – Massive Open Online Courses These terms likely relate to Massive Open Online Courses (MOOCs).

Topic 7 – Pedagogical Agents These terms appear to pertain to characteristics and features related to pedagogical agents, affective interaction, and simulations in an educational context.

Topic 8 – Automated Assessment The presence of terms such as ‘assessment’, ‘answer’, ‘question’, ‘essay’, ‘automatic’, and ‘scoring’ justifies the classification of this topic under Automated Assessment. However, it is noteworthy that terms like ‘dialogue’, ‘natural language’, and ‘question’ may also align with the domain of Intelligent Dialogue Systems. This dual alignment is particularly significant considering the term ‘question’ may pertain to either a question within an assessment context or one posed to a dialogue system.

Topic 9 – Emotion Recognition The terms assigned to this topic appear to represent the domain of emotion recognition.

Topic 10 – Recommender Systems In the context of intelligent tutoring systems research, these terms can be interpreted as Recommender Systems.

Topics Identified by Top2Vec

Unlike LDA, results from topic modelling with Top2Vec identified larger number of ITS relevant topics within the corpus which we present ordered by the size of each identified topic in Table 5.5. Of the identified areas, high-level topics correlate to a larger volume of documents with specific topics containing a lower volume of documents. The clustering of publications using the HDBSCAN algorithm leads to the assigning of single topic labels to each document, meaning that unlike probabilistic models like LDA, documents may not belong to multiple topics, and therefore more specific or low-level topics typically contain fewer documents. Within this section, we ensure that all references are made using publications present within our corpus. Given the criteria applied during data collection, particularly through using ArXiv API tools, research discussed may include pre-print or thesis research which has not been peer-reviewed.

Topic 0 – Adaptive Educational Hypermedia This topic represents the high-level area of Adaptive Educational Hypermedia Systems (AEHS). These may be defined as adapting content to meet the goals and needs of a user or student [178]. Documents within our corpus labelled under this topic typically investigate web-based approaches for adaptive tutoring [179] and relate closely to other ITS areas including Learning Management Systems (LMS) and MOOCs. We identified several approaches involving neural networks of which [180] [181] [182] are a sample, as well as framework proposals for the building of e-learning platforms [183], [184].

Topic 2 – Intelligent Dialogue Systems Intelligent Dialogue Systems (IDS) typically investigate the application of conversational agents, applied to assisting in the pedagogical process. Sample documents involve the application of conversational agents to address tutoring of concepts and principles with students for both physics and programming [185], [186].

Topic 5 – Pedagogical Agents We identified considerable interest in research related to the impact of pedagogical agents [187], [188], [189], [190] and how the presentation of these agents may impact success within ITS. Other documents more closely correlate to emotion recognition through the assessment of learner feedback [191]. Adaptation of pedagogical agents in response to emotional queues are frequent within this topic [192], however research may alternatively investigate the impact of perceived emotions of pedagogical agents [193].

Topic 7 – Learning Management Systems This topic entails the high-level area of Learning Management Systems. Within this topic, a significant volume of research relates to e-learning platforms such as Moodle [194] and includes proposals for the modification of such platforms to adapt to user learning styles and requirements. Interestingly, the majority of publications present within this topic avoid architectural specifications or computing-based terminology, and instead typically provide case studies of the implementation of existing LMS.

Topic 8 – Computer Supported Collaborative Learning Documents assigned to this topic generally relate to Computer Supported Collaborative Learning (CSCL). Relations within this area include pedagogical agents [195], although generally there were fewer instances of bridging between the identified topics.

Topic 12 - MOOCs Massive Open Online Courses (MOOCs) are a relatively recent aspect of ITS research, and we identify our earliest instance of this within our corpus in 2013 [196]. We identify 38% of research in this topic entailing learning analytics [197–199], which involves the wealth of data provided by MOOC platforms. This data may be applied to dropout prediction and forecasting of MOOC platforms [200, 201], and we identify 11% of documents involving dropout prediction.

Topic 13 – Architectures and Algorithms Documents assigned to this topic are more closely associated with implementation and architectures of models than ITS processes. We identify applications of fuzzy logic [202–204] comprising 25% of the topic, with 11% discussing or applying neural networks [205, 206] and 8% through clustering [207]. Given that algorithms and architectures will be likely present in the wider corpus we perform a temporal analysis of the entire corpus in section 5.4.2.

Topic 14 - Simulations This topic represents research involving the simulation of learning environments and simulated agents. We identify articles relating to pedagogical agents [208], CSCL [209] and adaptive hypermedia [210] within this topic. While documents relate to other ITS areas the majority discuss the simulation of environments to assist with learning. Instances of simulation include resource allocation training for police forces [211] and the use of virtual reality simulated environments [212, 213].

Topic 17 - Gamification Publications applying gamification within ITS fall within this topic, with research contributing to the use of game mechanics for positive educational outcomes. Games may be applied to assisting learning in STEM subjects [214, 215] within virtual learning environments or in the tutoring of programming [216].

Topic 19 – Grading and Essay Scoring This topic relates to the grading of work with ITS and is most directly associated with the area of Automatic Essay Scoring (AES), however other areas of grading exist within the topic. We identify 42% of documents discuss essay scoring directly, with research investigating short question grading [217] and essay scoring [218, 219]. The grading of text is not the only research within this topic however, and we identify unpublished research in the automated grading of map sketches [220] within our corpus sample. Within this topic we identify 18% of publications applying neural net-works, while 6% apply ontologies and 5% applying Bayesian learning.

Topic 28 – Recommender Systems Documents relating to recommender systems comprise this topic, which is the smallest relevant topic identified by our analysis. Research within this area present systems for the recommendation of courses in MOOC platforms [221] and adaption of learning environments using recommender systems [222] amongst others. This topic can be closely linked to several of our identified topics including adaptive educational hypermedia, computer-supported collaborative learning and MOOC systems.

Discussion and Comparison of Qualitative Assessments

In comparing the topics identified by both LDA, and Top2Vec, it is worth first establishing the similarities between both approaches. Both LDA and Top2Vec have identified several overlapping topics, indicating common themes within the ITS corpus. These are Adaptive Educational Hypermedia, Computer-Supported Collaborative learning, Pedgogical Agents, MOOCs, Automated Assessment, and Recommender Systems. With regard to Adaptive Educational Hypermedia, we argue that the terms identified by Top2Vec for this topic more closely align with specific terms used to describe AEH, such as the presence of 'personalization'. It is noteworthy that the acronyms 'aeh' and 'aehs' remain unrecognised within the LDA-derived topic. For Computer-supported collaborative learning, both present the terms 'collaborative' and 'collaboration'. Once again, only the Top2Vec topic includes the associated acronym, however, LDA does achieve a representation of the 'computer supported' bigram, which is not possible through Top2Vec. Both models identify topics related to Automated Assessment or Automated Essay Scoring, however the Top2Vec results appear, generally, to be more closely related given that the contrasting terms produced by LDA include terminology from other domains such as 'dialogue', 'natural_language' and the homonym 'question' which can be interpreted differently depending on the context it is used.

Despite these similarities, there are notable differences between the two models. With regard to other topics identified, Learning Management Systems, Intelligent Dialogue Systems, Machine Learning Models and Algorithms, Simulations and Gamification are unique to Top2Vec, while LDA contrastingly identifies a topic related to Educational Institutions and Organisations, and Emotion Recognition.

The identification of a topic entailing terms specific to machine learning models and algorithms is unique to Top2Vec for this corpus. Considering the procedure for learning embeddings with Doc2Vec, employed as an embedding model in Top2Vec analysis, it is likely that algorithmic terms are represented as closely positioned in the embedding space. This is attributable to their tendency to encode words with similar meanings or contexts as vectors that are proximal in the embedding space, as discussed in Section 2.1.1. If publication documents therefore discuss algorithmic terminology, it is likely that they will lie close together in semantic space, which would enable them to be assigned to a cluster by HDBSCAN. In contrast, the probabilistic per-term modelling of LDA which does not account for the context in which terms are used may fail to group algorithmic terms as the model has no prior understanding of the 'meaning' of the words, only their occurence probabilities within a document or corpus. Considering that the optimally configured LDA model comprises 12 topics, it is plausible that this particular topic might not have been detected unless the number of topics was expanded. This is noteworthy, given that in the prior analysis of a corpus of literature on Arabic Sentiment Analysis (Section 5.2, Table 5.2), algorithmic topics were identified and were even separated into distinct topics, however the number of topics specified in the optimal model was also significantly greater than those identified by evaluating coherence in the latter study of ITS literature. This scenario necessitates a more human-guided and time-intensive analysis across various topic numbers to accurately discern the optimal topic model, while disregarding quantitative analysis using coherence measures.

Generally, it is clear that both LDA and Top2Vec provide valuable insights into the topics relevant to Intelligent Tutoring Systems. LDA tends to offer broader categories, capturing overarching themes within the ITS domain, whereas Top2Vec provides more specific and granular topics, highlighting detailed aspects and applications of ITS technologies.

In terms of selecting a suitable method for the end objective of assisted literature analysis, it is worth considering the strengths and caveats of both models from a qualitative, quantitative, and algorithmic perspective. For a high-level overview, LDA's broader categories might be more suitable. However, for a detailed and specific understanding, Top2Vec's granular topics offer deeper insights. According to the quantitative analysis presented in Section 5.4.2, coherence metrics suggest that Top2Vec is the preferable model. Furthermore, it generates a higher average number of topics from the corpus in comparison to LDA. It is also worth considering that a higher number of relevant topics were then identified from the Top2Vec results. The evaluation based on coherence metric also appears, generally, to be reflected in the qualitative analysis of the topics, where not only a larger number of topics were found to be relevant, but they also appear to be generally of a finer-grained nature, with fewer instances of overlap between topics. This finer-grained topic modelling solution is preferable to literature analysis given that subsequent investigations will seek to identify inter-topic relationships and temporal trends, and thus having a clearer separation enables a clearer comparison between topics.

From an algorithmic perspective, the underlying methodologies for both techniques differ considerably and this directly affects the possibilities of subsequent analyses. LDA works under the assumption that documents within a corpus can be represented as a probabilistic distribution over a mixed number of topics. This means that a single document could belong to multiple topics. In real-world datasets, particularly those employed in literature analysis, it is plausible to assume that documents encompass a multitude of distinct underlying topics, such as methodologies, algorithms, or domain-specific terminology. This multi-topic nature is advantageous for conducting a more comprehensive literary analysis, facilitating researchers' comprehension of the varying degrees of relevance that each academic work holds across different topics, thus informing their analysis. In contrast, cluster-based topic models such as Top2Vec assign each document as either belonging to one topic, or as being an outlier, which is facilitated by HDBSCAN clustering. Thus, it is relatively simple to summarise all documents belonging to a topic, however the specification of outliers during clustering does mean that a significant portion of the corpus is discounted from the analysis. Still, it is apparent through the quantitative and qualitative analysis that Top2Vec affords a more useful topic modeling approach that can be conveyed to users of the subsequent analytical framework presented in Chapter 6, where the intended audience may not have the time to dedicate to the interpretation of topic probability distributions. Moreover, considering that LDA necessitates a comprehensive assessment of topic coherence to determine the optimal number of topics, which is both time-consuming and computationally demanding, the implementation of an alternative solution that facilitates the automated discovery of an optimal topic model configuration, as exemplified by Top2Vec, is advantageous. For the subsequent analyses involving topic modelling in the rest of this chapter, and the following chapter investigating a comprehensive literature analysis framework (Chapter 6), we select Top2Vec, due to the advantages identified and discussion in the prior quantitative and qualitative analysis.

Temporal Analysis of Top2Vec identified Topics

We visualise the changes in resulting topics from our analysis in Figure 5.6. This process involves gathering all publications categorised under a specific topic and then counting the frequency of publications according to their publication year. These are normalised per-year to eliminate influence by changes in yearly publication volume. Resulting temporal distributions generally correlate to the sizes of our identified topics, with documents assigned to Adaptive Educational Hypermedia forming the largest portion of research from 2001-2015. Other topics outside of these

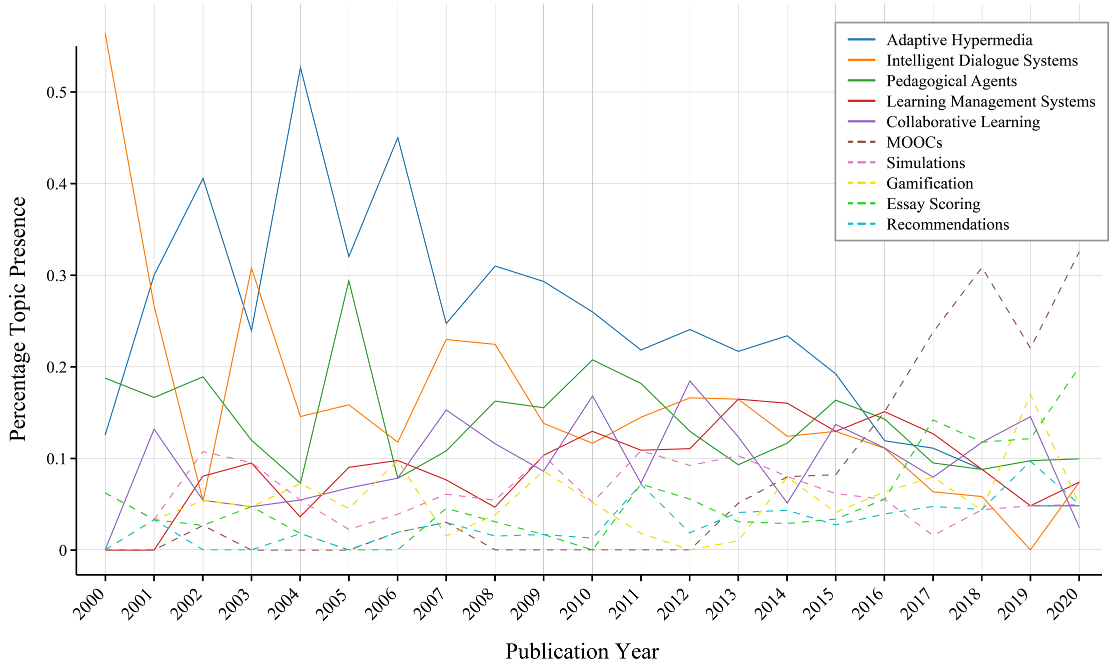


Figure 5.6: Temporal changes in topics from 2000-2020. Normalised by number of publications each year.

generally fail to form more than 20% of research interest prior to 2016, where research into MOOCs overtakes other topics to become the most dominant topic within our corpus. This considerable change in interest towards MOOC platforms may be influenced by the wealth of data obtained and provided by such platforms, with public datasets such as [223] allowing researchers easy access to data to contribute to the field, and the general trends towards 'Big Data' application and research. A decrease in popularity of AEH, IDS, and several other topic types may further contribute to the adoption of MOOC type research, which may provide more easily accessible datasets and feature ranges.

We present the occurrences of different algorithms and architectures in Figure 5.7. The results indicate the consistent presence of ontologies within research throughout 2000-2020, while the applications of other algorithms identified by Top2Vec analysis fluctuate in popularity. Most notably, an increase in the presence of both clustering and neural networks is observed from 2010-2020 within the entire corpus. In recent years (2019-2020), the volume of research discussing neural networks has increased considerably. This may indicate general trends in the larger computing field and may be attributable to recent novel algorithms such as the transformer

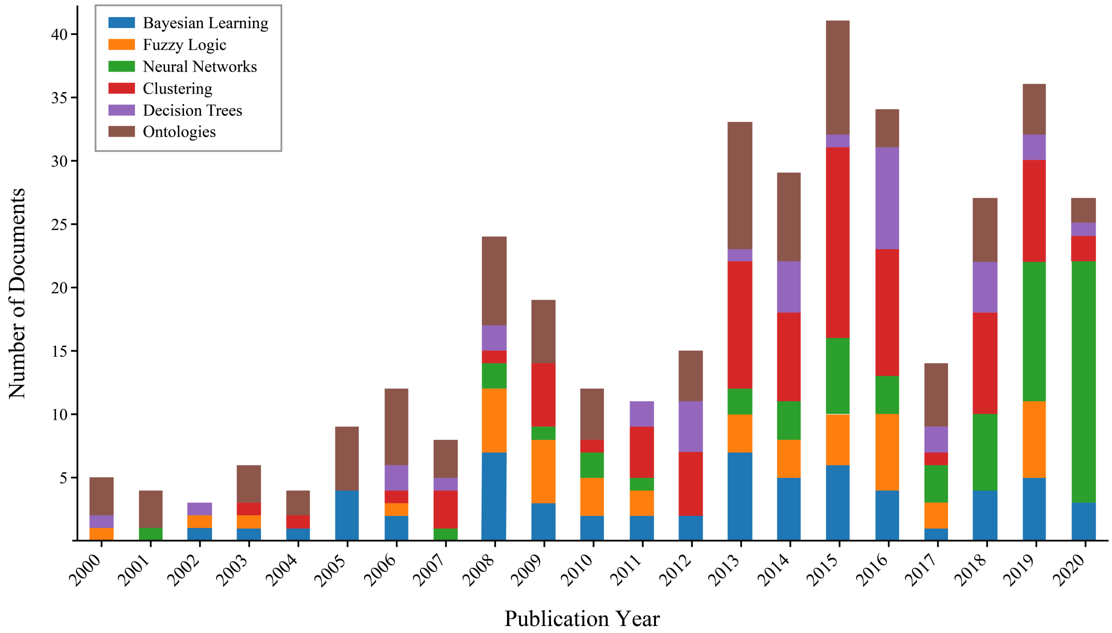


Figure 5.7: Algorithmic and architectural presence based upon publication year.

network and BERT [40], which have had a considerable impact upon research in the field.

Topic Graph Relationships

For further analysis, and addressing RQ2.2, we construct a network of relationships between the topics of ITS, as shown in Figure 5.8. These are constructed using the cosine similarity between the average embeddings of the documents of each topic. The average document embeddings were generated using the Doc2Vec document embeddings of all documents assigned to a topic. We assign connections between topic nodes using the three highest-scoring similarity relationships for each topic, based on an exhaustive search of all combinations of topics.

Relations between Pedagogical Agents, Simulations, IDS and Gamification reflect the links present within the corpus, where agents may be presented to users in a graphical manner. The cosine similarity scores between these four topics are the highest within the network and demonstrate the linking themes and interoperability that these areas present. In the case of AEH and simulation, research investigating the adaptation of simulated agents in response to user or student is present within both topic corpuses. Further high-scoring similarities are observed between

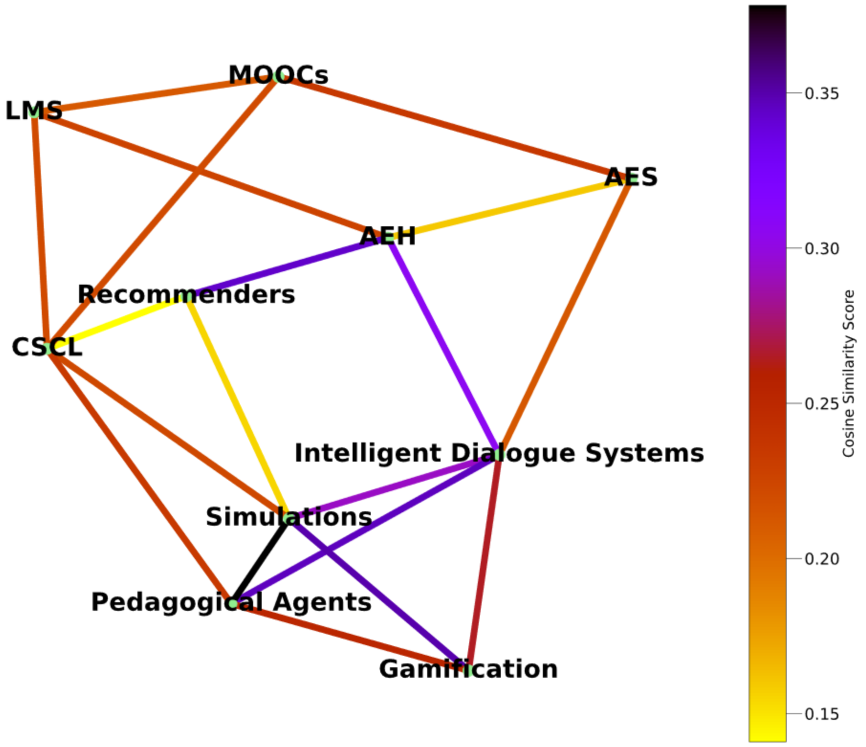


Figure 5.8: Relationships of relevant topics based on cosine similarity between average of topic vectors.

AEH and Recommender Systems, wherein publications may discuss the adaptation of recommender systems dynamically, based on user responses and performance. Given that recommender systems may be closely attributed to adaptation of systems to user input, we argue that this is represented through the association with Adaptive Hypermedia (AH) and Learning Management Systems (LMS) through the recommendation of course content.

LMS is additionally associated with the CSCL and MOOC systems. We argue that LMS and MOOC systems are by nature closely linked (with MOOCs forming a subset of LMS), and therefore documents within these topics may share semantic terms. Both LMS and MOOC system research incorporate aspects of collaborative learning within our corpus. A link between MOOCs and Automated Essay Scoring research is present, being the strongest link for AES, which is one of the weakest scoring topics in terms of cosine similarity with other topics. This reflects how many of the topics present within the corpus offer a degree of interoperability, which is less so in the case of AES.

5.4.3 Discussion of Findings

The application of the novel Top2Vec [6] algorithm to topic analysis of the ITS literature allows an overview of the current research field within the case-study domain. Contrary to well-known approaches, such as LDA [13,156], which was applied for the analysis of Section 5.2 the algorithm requires fewer pre-processing steps and therefore demonstrates potential in application to a range of epistemological research without expert knowledge. Furthermore, this analysis approach ensures that a significantly higher volume of research can be processed and analysed compared to manual review types. Subsequent analysis of the identified resulting topics contributes to an understanding of the relationships between topics and the volume of research that various areas contain, which addresses both RQ2.2 and RQ2.3. General findings from our investigation indicate that Adaptive Hypermedia research constitutes the highest volume of research overall. This area presents a high level of interoperability with others, such as research applying Simulation and Recommender Systems in an adaptive manner, based on user input. Temporarily, adaptive hypermedia accounts for the largest portion of ITS research up until 2016, where it is overtaken by MOOC research. Given that our analysis accounts for all research from 2000-2020, there exists further opportunity for a dedicated analysis of the more recent years publications only, in order to form a better understanding of reasons for MOOC research popularity, and identification of potential new areas of research within. Topics such as Automatic Essay Scoring are clearly under-represented and may deliver promising avenues of future research, especially as some of this research seems yet unpublished. Temporally, we identify a shift in research in recent years (2016-2020) with a considerable increase in interest of MOOC systems and applications of neural network architectures to research within these years. This, we argue, is likely the result of the increase in availability of data generated by MOOC systems, which achieve a considerable throughput of users and therefore volume of data. In the case of applications of neural networks, we argue that the interest spike follows the considerable improvements made in recent years for transformer-based and pre-trained networks. As a final note on our methodology, we identify limitations in the applications of abstracts only within our corpus, whereas structured full-text data may have provided valuable insight into topics of separate sections (e.g., related works, methodologies). The key phrases for our search were based only on the ITS Conference. This may be a limitation, and other possible variations could be considered. However, given that the extraction was done over the last 20 years, so conforming exactly to our target time period, we can say with some confidence that these results clearly show the progress of ITS research during the past 20 years from an ITS conference perspective.

5.5 Epilogue

The main objective of the research presented in this chapter has been to evaluate two starkly different algorithmic approaches to topic modelling, with the objective of identifying a suitable candidate to be applied to assist the literature analysis process. Beginning with an exploration of LDA applied to a corpus of literature related to Arabic Sentiment Analysis, it was possible to uncover a number of relevant topics, with a topic modelling solution that was generally of a good quality, with topics even separating a multitude of different algorithmic terms into distinct topics. This begins to address RQ2.1 by establishing that LDA is a valid method for literature analysis, something that is already supported in existing studies [156]. However, when applied to a dataset derived from ITS research, it became evident that LDA analysis necessitates comprehensive evaluation to determine the appropriate number of topics as a hyperparameter for achieving an optimal solution. Exhaustive comparison of coherence measures in order to identify an optimal topic number configuration of LDA is not suited to the overarching objective of designing a framework for automatic literature analysis.

Despite achieving an optimal configuration, the solution derived from LDA was generally inferior in both quantitative and qualitative assessments compared to the results produced by Top2Vec when applied to the ITS corpus. This direct comparison of Top2Vec and LDA on the same corpus is valuable in further addressing RQ2.1, where based on the results of the analyses, and the benefits of automatic identification of an optimal number of topics and the possibility of leveraging neural-embedding models to ensure that the semantic and syntactic nature of the data can be accounted for, Top2Vec was selected for all subsequent topic modelling research presented in this thesis.

In answering RQ2.2, through the adoption of neural-embeddings in conjunction with Top2Vec, it became feasible to apply downstream analyses of the semantic relationships between topics, which were demonstrated in the case-study of ITS literature as supporting the analysis. For literature analysis, this is useful to researchers in helping to assist in quickly identifying inter-domain relationships in terms of semantics, which could be valuable when approaching a new domain. This is valuable not only in the context of assisting literature analysis, but also the wider publishing industry, where one application could be for journalists, the mapping of relationships in topics identified from social media streams.When supplemented with additional metadata, such as key terminology, named entities, or sentiment ratings, this methodology could serve as a valuable resource for identifying potential cues for the exploration of novel topics to produce content on or investigate further.

Finally, in answering RQ2.3, both studies included in this chapter demonstrated a downstream temporal analysis of topics within the literature. This contributes to the field of literature analysis by offering a graphical depiction of trends in literary subjects, aiding researchers in determining if specific topics are currently trending. However, both studies presented in this chapter have been focused upon niche domains, and have failed to demonstrate two key concepts which is essential for an automated literature analysis framework. Namely, a demonstration of the generalisability of the approach to any literature domain, and the demonstration of the analysis in a fully-automated manner. In order to address these limitations, in the next Chapter, a continued development of the concepts of automated literature analysis is conducted, in order to explore the contribution that automated literature analysis can provide during literature analysis.

Table 5.5: Identified topic-word distributions by Top2Vec. Topics deemed relevant to ITS by qualitative analysis.

Topic Terms (Manually-Determined)	Topic ID	Topic Label
Hypermedia, aeh, aehs, ims, adaptive, adaptation, adaptivity, navigation, personalization, links, specification	0	Adaptive Educational Hypermedia
Dialogue, tutoring, natural, intelligent, language, tutorial, automatically, conversational, apos, corpus, medical, quot	2	Intelligent Dialogue Systems
Agent, animated, emotion, affective, emotional, pedagogical, agents, emotions, facial, conversational, apos	5	Pedagogical Agents
Moodle, lms, source, management, open, centre, lectures, platforms, basic, dashboards, assignments	7	Learning Management Systems
Peer, assistance, collaborative, conditions, learned, tutor, collaboration, dialogue, actions, cscl	8	Computer-Supported Collaborative Learning
Moocs, massive, mooc, dropout, forum, open, engaging, courses, videos, rates	12	MOOCs
Bayesian, networks, fuzzy, logic, artificial, diagnosis, intelligence, intelligent, neural, tutoring	13	Machine Learning Model Types and Algorithms
Simulations, simulation, intelligent, animated, virtual, training, multimedia, agents, reality	14	Simulations
Games, game, serious, play, agent, interact, intelligent, bring, initiative, metrics, simulation	17	Gamification
Essay, scoring, essays, automatic, grading, writing, automated, English, language, neural	19	Grading and Assessment Scoring
Recommender, recommendation, personalization, links, personalized, java, hypermedia, experiments, lecture, adapting	28	Recommender Systems

CHAPTER 6

Provision of an End-to-End Framework for Assisting Literature Analysis

This Chapter continues in addressing GRQ2, through iterative research aimed at the provision of an end-to-end framework for automated literature analysis. This is achieved through the works titled "A Topic-Centric Crowdsourced Assisted Biomedical Literature Review Framework for Academics", published in the International Conference on Educational Data Mining, and "A Co-designed Framework for the Analysis of Large Volumes of Biomedical Literature", which is currently under review for the Journal of Biomedical Informatics.

6.1 Prologue

As discussed in Section 5.5, the findings of the previous chapter served as an exploratory investigation into the feasibility of using topic modelling techniques to help in understanding academic literature, more specifically, to contribute to assisting literature analysis.

Based on the outcomes of the previous chapter, it remains necessary to consider the barriers to using the aforementioned topic modelling methodology for individuals who do not possess the prior knowledge to design the functionality for literature review. Therefore, this Chapter extends the findings of the previous Chapter - specifically, the utility of topic modelling of academic texts using the Top2Vec algorithm [6] for the identification of literature topics - by exploring them from an applied perspective. The aim is to develop an end-to-end framework that facilitates the analysis of academic literature, with a focus on testing generalisability by transitioning to a different domain: the biomedical domain. This is formed of two works. Firstly, a case-study investigation into biomedical literature was conducted, in conjunction with experts from the domain, to form the basis of a visualisation framework, which allows the analysis of topics identified within the literature by domain experts; this work has led to a paper titled *A Topic-Centric Crowdsourced Assisted Biomedical Literature Review Framework for Academics* [224]. The outcomes of this study demonstrate the value that the proposed visualisation functionality contributes to the specific domain of biomedical literature related to the human kinome. However, the framework remains focused on a specific domain and requires the prior collection of suitable data from literature sources. Consequently, further research was undertaken, culminating in the publication *A Co-designed Framework for the Analysis of Large Volumes of Biomedical Literature.* This work advances the findings, to contribute to a **full-fledged framework for literature analysis**, which is evaluated based on feedback from a study of **medical experts**. The outcomes of this study highlight the value of the proposed framework in contributing to improving the analysis of the literature, while also indicating potential avenues of exploration in order to improve the value for the research community.

6.2 Chapter-Specific Research Questions

The following research questions are addressed in this chapter, which focus on the application of the findings of the previous Chapter 5 in an accessible format, which can benefit researchers:

- **RQ2.4:** *Can rule-based extraction of named biomedical resources contribute to the comprehension of automatically generated literature topics?*

- **RQ2.5:** *Can an end-to-end framework, built upon cluster-based topic modelling confer a benefit to researchers through assisting the literature review process?*

- **RQ2.6:** *What are the effects on cognitive load, technology acceptance, and general acceptance of users when using a software implementation based on our framework?*

- **RQ2.7:** *How does the proposed literature analysis framework and software implementation influence the accuracy of the identified results in practical use cases for medical researchers?*

6.2.1 Research Objectives

Based on the above motivation and research questions, the research objectives for this Chapter are:

- **RO2.3:** Conduct a small-scale case study of automated literature review, using expert knowledge to evaluate the outcomes of the identified topics and confirm the feasibility of applying topic modelling techniques to contribute to the literature review process. (Addressing RQ2.5).

- **RO2.4:** The provision of a framework for the assisted analysis of large volumes of academic literature, evaluated through feedback by medical professionals with respect to the concepts of cognitive load, technology acceptance, and general satisfaction. (Addressing RQ2.5, RQ2.6 and RQ2.7).

6.3 Related Work

The task of automatically identifying and creating knowledge from academic literature aims to take advantage of the large volumes [225] of research currently available. The analysis of academic literature within the medical domain has observed an increase in the volume of citations between 2011 and 2018, indicating a general increase in interest in research within the domain [226]. The synthesis of knowledge from this wealth of available literature presents challenges, in part due to the innate nature of scientific articles, where domain-specific language presents limitations to traditional, term-based approaches. Instead, recent advances in deep learning have allowed the use of contextually aware and semantically enabled techniques, which have demonstrated effectiveness in the retrieval, curation, organisation and interpretation of literature content [227].

As a data-intensive field, there has been an emergence of numerous biological data sources of various quality and provenance. However, due to the lack of a standard ontology with fine granularity, it is difficult to perform data mining on heterogeneous biology data. Despite efforts to normalise data formats and collect observations in databases, including the introduction of standardised Journal Article Tag Suite XML(JATS)[1] to ensure standardisation of structured formats of publications, and Medical Subject Headings (MeSH) categorisation of publications, a large amount of information relevant to biomedical research is still recorded as free text in journal articles and comment fields of databases [228]. Therefore, the retrieval of information from articles and the merging it with existing biological databases can provide a crucial foundation for data-mining in bioinformatics [229]. To address this, works have been proposed for the extraction of entities within biological texts, including RegulonDb [230], which applied rule-based identification of regulatory interactions in *Escherichia coli*, with results demonstrating a rule-based approach sufficient in identifying 45% of all named entities when compared with a manual extraction approach. The extraction of genetic and protein interactions was carried out by BioC-BioGRID, which released a corpus of 120 full-text articles with annotations of biological and molecular entities [231]. Based on these problems and subsequent approaches, [229] proposed the Curator Assistance Problem. This problem proposes four objectives, which were defined given a full-text biological research paper: 1. Recognition of evidence described in an article, compared to information in other articles. 2. Recognition of evidence which is supported by experimental data, compared to hypothetical or vague statements. 3. The distinction between

[1] `https://jats.nlm.nih.gov/`

statements on layout, compared to statements of results. 4. The recognition of negative statements.

For the biomedical field, text mining of the literature can be performed to achieve several different objectives. Raja et al. [232] divides biomedical text mining tasks into *Document Retrieval*, *Document Prioritisation*, *Information Extraction*, *Knowledge Discovery*, *Knowledge Summarisation* and *Hypothesis generation*. More specifically to a semantics-enabled domain, Kilicoglu et al. [227] defined four broad domains, including literature-based discovery, automated knowledge-base construction, knowledge-augmented biomedical NLP, and literature search and information retrieval. Literature discovery techniques are based on the discovery of literature from structured data within knowledge-bases. Examples include the characterisation of literature discovery as a link prediction task for the production and analysis of a knowledge graph, focusing on Alzheimer's disease [233]. Automated knowledge-base construction, on the other hand, focuses on the construction of knowledge-bases through the extraction of relations from biomedical literature. This could be through the construction of graphs encoding relationships between foods, medicines, and mental illnesses [234], or the visualisation of named entities within the literature [235]. For knowledge-augmented biomedical NLP, the authors categorise works which incorporate knowledge-base-derived information for the improvement of pre-trained language models, to improve performance on specific NLP tasks. From the perspective of pre-trained language models, it has been demonstrated that fine-tuning of language models on biomedical literature can contribute to state-of-the-art results upon biomedical NLP benchmarks, with language models such as PubMed-Bert [236], SciBERT [175] and BioBERT [80]. Finally, for literature search and information retrieval, examples include incorporation of semantic information to improve the retrieval of biomedical literature, by contributing to the queries used during literature retrieval. An example of such works is a method for fine-grained analysis of MeSH concepts, which are commonly used in the semantic indexing of biomedical concepts [237]. Alternatively, Khader and Ensan [238] investigated a Contextual Query Expansion framework, which involved a deep learning-based language model tasked with the generation of candidate query expansion terms, which may enhance an original query. Based on results from the TREC-COVID dataset [239], the work demonstrated a drastic improvement in search performance when expansion terms were used, compared to the original query. This is of interest, as it indicates that the definition of suitable search queries is vital in ensuring the retrieval of suitable results. We take this into account when designing our framework for assisted literature analysis, through the provision of aliases and rule-based logic in filtering search terms prior and during analysis (see Section 6.7.2).

Medical and academic literature is not intended to be machine-readable, with the main priority being readability for humans. This confers some difficulties in any tasks which seek to perform a computational analysis of the information present, due to a multitude of reasons. First, full-text versions of academic texts are not always freely available. While open access publication has observed a steady growth in adoption, there remain a significant portion of the literature that can only be accessed through paid services [240, 241]. Additionally, the automated retrieval of academic resources through APIs is still limited, with discrepancies between publishers as to what information can be extracted. Furthermore, the styling requirements of a myriad of academic journals, conferences, and other platforms confer difficulties in the extraction of information within publications from a machine-learning perspective. For example, a human-reader could easily identify useful information and named entities, such as the keywords provided by the author to assist in the categorisation of the article, the title, abstract, related works, methodology, results, and discussion and concluding remarks. This can be easily performed by a human, even when the styling of the publication and aliases of common terminology of sections is used, i.e. a "*Methodology*" section may be provided under the alias "*Materials and Methods*". In comparison, an automated system would experience difficulties in the accurate partitioning of sections of the text. This can be framed as a named-entity recognition task, and has been investigated through methods such as automated keyphrase extraction [242], rule-based extraction of methodological sections, and deep learning methods for the extraction of scientific methods, using pre-trained language models [243].

Rule-based extraction requires the definition of rules for the identification of sections of the text by experts. Such methods include dictionary matching, wherein a dictionary of entity names is compiled, and can be matched against the sections of text. Such approaches, however, require the prior definition of the dictionary terms with which to match, which is time expensive [244]. On the contrary, deep learning methodologies have shown potential in tasks related to named entity recognition, which essentially aligns with the task of identifying sections within scholarly texts. It has been demonstrated that pre-trained language models, fine-tuned to a scientific corpus, can extract scientific resource-terms [245], and subsequently profile these based on citation information. These rely upon the provision of prior data in order to ensure that the model is generalisable upon the problem domain. This was factored into our methodology, where our analysis is concentrated on the abstracts of publications, thereby mitigating the uncertainties associated with parsing full-text publications. By taking only abstracts into account, with no clear labelling present within each abstract, we frame the problem of making sense of large volumes of publication abstracts as an unsupervised learning task. Additionally, we ensure to eliminate uncertainties with the parsing of documents, allowing a focus instead upon a topic-modelling driven analysis of publications, and subsequent analysis based on the semantics of topics.

Of the concepts discussed, *knowledge discovery* is of interest to the research presented in this chapter, defined as the creation of knowledge from large volumes of structured or unstructured data [246]. To perform knowledge discovery within a large corpus of such texts, machine learning (ML) or artificial intelligence (AI) techniques may be applied to the corpus. Examples of these include classification approaches where a prior goal is known, and existing labelled data can be used to train an algorithmic model to subsequently classify new data. Examples of these include the classification of medical literature based on hallmarks of cancer [247] and cancer susceptibility genes [248]. This falls into the domain of supervised learning. In contrast, in the goal of knowledge discovery in academic literature, an investigator may have no prior knowledge on a topic and, therefore, seeks to identify patterns or trends within data without provision of any previously labelled data. In such tasks, as the one defined in our study, unsupervised learning techniques are better suited. Soldaini and Goharian [249] presented an unsupervised method for the extraction of named medical entities from the medical literature. Similarly, Ding and Luo [250] presented a neural network attention-based approach for the unsupervised extraction of keyphrases from the medical literature, which can then be provided as a visualisation of relevant keyphrases in a knowledge graph format, based on a given user query. This is of value to the research outlined in this chapter, as the work presents the objective of grouping and visualising knowledge extracted from literature in an unsupervised manner, based on the relevance of the extracted knowledge relevant to a user query. In the case of our proposed study, we adopt an alternative approach to knowledge graph generation. While the work presented in [250] determines relevance based on the attention weights of a neural attention network for given noun keyphrases, our work determines graph nodes as an averaging of all document vectors belonging to a topic, where document vectors entail embedded literature abstracts. To construct the semantic-relation graph, which may be considered akin to a knowledge graph, we compute the semantic similarities among all candidate nodes, as elaborated in Section 6.7.6. This provides the benefit of accounting for the contextual information encoded in abstracts, which are typically written in a style that concisely summarises the overall motivation and findings of the literature. Furthermore, rather than focusing on extraction of relevant keyphrases from literature based on a given query, the research presented in this chapter adopts presents researchers and domain learners with the opportunity for the wide-scale analysis and extraction of a significant volume of research, and the consolidation of the identified themes identified into an understandable semantically-aware graphical visualisation.

From the point of view of supervised learning, Portenoy and West [251] proposed a method to select relevant articles for a literature review using seed articles. Existing reference lists from review papers were used as labelled data to train supervised classifiers. However, this approach presents limitations due to the requirement of training data in the form of citation lists. This, in turn, implies a reliance upon data that is domain relevant, which can make the generalisability of such an approach difficult if the domain which is under review has not been sampled adequately in the original training corpus.

6.3.1 Topic Modelling and Assisting the Literature Review Process

In comparison to the aforementioned works identified within the literature, the technique of topic modelling, which has already been shown to be effective in assisting literature reviews [156], may be applied. Latent Dirichlet Allocation (LDA) [252], is probably the most well known variant of the topic modelling algorithm, and introduced a generative probabilistic model, based on the hypothesis that documents can be represented as random mixtures over latent topics, with each topic entailing a distribution of words. Since its inception, LDA has been demonstrated in several research domains, with the original paper garnering over 51,000 citations on Google Scholar at the time our work was published[2]. Examples of these include the topic modelling of scientific abstracts [252, 253], the detection of hate speech [254, 255] and opinion mining [256], to name a few. In our specific research area, the LDA algorithm and its variations have been showcased in several academic papers within the medical and biomedical fields. Bio-LDA [257] demonstrated that LDA can be applied to the identification of biological terminology within PubMed articles and to identify latent topics within the literature. From a more general perspective, [156] adopted an unsupervised topic modelling approach to knowledge discovery in the academic literature, through the proposed "Smart Literature Review" framework. For this, the study applied the Latent Dirichlet Allocation (LDA) algorithm [252] to identify topics within the literature. Following this, the framework applies post-processing steps for the labelling of the papers based on the highest probabilities assigned for each paper.

There are, however, limitations to using LDA for a task such as automated literature analysis [252]. LDA requires a prior specification of the number of topics to be extracted from the literature, and the correct specification of this hyperparameter

[2] https://scholar.google.co.uk/scholar?cites=17756175773309118945&as_sdt=2005&sciodt=0,5&hl=en

is essential in identifying meaningful topics from the corpus. Methods to mitigate this exist, such as evaluating topic coherence metrics across a range of topic numbers as used in [156]; however, this is an exhaustive and time-consuming approach that makes it unsuitable for a generalisable framework, where results must be provided in a practicable amount of time. Additionally, pre-processing steps are required, to remove frequently used words within the literature which do not contribute any semantic meaning, commonly referred to as stopwords. Consqeuently, LDA is a bag-of-words approach to topic modelling, and any literature analysis based on this algorithm cannot account for the contextual information encoded in documents. In comparison, the Top2Vec algorithm [16] provides an alternative to Bayesian topic models such as LDA [252], eliminating the need for pre-defining the topics number and filtering stop words. This is achieved by leveraging pre-trained language models such as Doc2Vec [258], BERT [40], or Universal Sentence Encoder [165], to produce a vector representation of each document within the corpus. A semantic embedding of joint document-word vectors is computed where the distance between document and word vectors represents semantic association. This ensures that semantically similar documents (semantic similarity between vectors is discussed in Section 3.2.1) achieve a smaller distance between each other, compared to dissimilar documents. This is best represented visually in Figure 6.1, representing a simplification of how semantically similar documents *(purple)* and words (*green*) lie closer together in the embedding space.

Given the high dimensionality of document embeddings, clustering requires a prior reduction of dimensionality through the UMAP [14] algorithm. The resulting lower-dimensional embeddings are clustered using the HDBSCAN [259] algorithm, with the resulting cluster labels representing the topic of each document within the corpus. The hierarchical nature of HDBSCAN ensures automatic identification of the topic number, which presents the opportunity to eliminate the exhaustive evaluation of topic numbers that was necessary in the framework presented by [156]. Furthermore, the capacity of Top2Vec for the leveraging of pre-trained language models introduces the ability to capture the semantics of documents within the corpus, with language models such as BERT [40] demonstrating a considerable im-

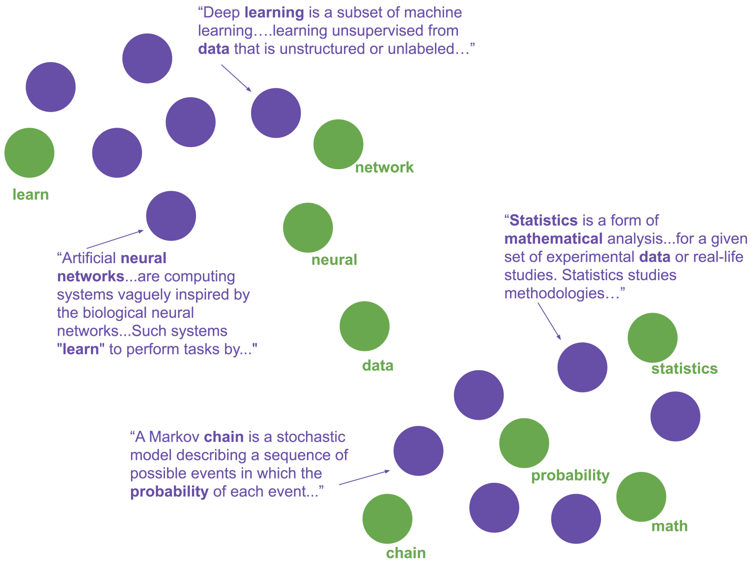

"Deep learning is a subset of machine learning….learning unsupervised from data that is unstructured or unlabeled…"
network
learn
"Artificial neural networks...are computing systems vaguely inspired by the biological neural networks...Such systems "learn" to perform tasks by..."
neural
"Statistics is a form of mathematical analysis...for a given set of experimental data or real-life studies. Statistics studies methodologies…"
data
statistics
"A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event..."
probability
math
chain

Figure 6.1: A Visual representation of document-word vectors in embedding space, with documents represented in purple and words represented in green, taken from Top2Vec [6].

provement across a number of NLP benchmarks. Top2Vec has been demonstrated to outperform LDA when applied to the 20NewsGroups [167] benchmark dataset [16]. In our previous research presented in both this chapter, and the former, we have explored the application of the Top2Vec algorithm for literature analysis in two studies. First, a case study analysis of research related to intelligent tutoring systems demonstrated the feasibility of the algorithm, by identifying trends within the literature and relationships between topics identified within the literature [260]. Subsequently, a small-scale study was conducted on biomedical research related to the human kinome, where we explored the introduction of a framework for the expert labelling of topics identified in the literature [224]. This advances the current research on smart-literature-review [156, 252, 253] by accounting for contextual information within texts, rather than relying on probabilistic measurements of individual term frequencies. Additionally, it facilitates the efficient development of a visual framework by eliminating the need for exhaustive hyperparameter tuning, specifically concerning the number of topics and the alpha and beta parameters essential to LDA methodologies [261, 262].

Building upon the findings of these studies, we formulate our proposed work with the aim of contributing to a general framework that can facilitate automated literature analysis from any domain selected by medical experts, without the need for prior intervention or programming, which would be alternatively necessary in the framework presented by [156]. Through the utilisation of Top2Vec's cluster-based topic modeling approach, we can eliminate the requirement of determining the number of topics to be retrieved, thus simplifying the overall system implementation for users.

6.3.2 Cognitive Load and Technology Acceptance

In the domain of educational psychology, it has been theorised that the processing of new information is profoundly influenced by the *working memory* [263–265], so that only a few elements of information may be processed at any time. In this context, working memory can be used interchangably with short-term memory [266], and refers to the cognitive ability to temporarily hold and manipulate information, which influences learning outcomes in an academic context [267]. Subsequently, attempting to process too much information at one time leads to an overload in the capacity for processing information. The concept of capacity for processing information, referred to as cognitive load, can be generally summarised as representing the load imposed by performing a particular task on the cognitive system [268]. This is characterised as both mental load and mental effort. In this context, mental load refers to the cognitive demands placed on learners, affecting their ability to process information [266], and mental effort refers to the amount of cognitive capacity required to meet the demands of the task [268].

Thus, within this research, we consider the effects that interaction with our framework will have with regard to cognitive load through gauging user perceptions of mental effort, and mental load. To measure such types of cognitive load, numerical rating scales have been found to be a suitable method [269] based on obtaining feedback from individual participants, where they have been applied to determine the self-reported cognitive load of system interactions in other software implementations [270, 271].

Similarly, the concept of technology acceptance refers to perceived usefulness, perceived ease of use, and user acceptance of information technology [272]. The idea that perceived usefulness influences system usage was proposed by Schultz and Slevin [273] and Robey [274]. Schultz and Slevin performed an exploratory factor analysis on 67 questionnaire items, resulting in a seven-point scale. We modify the questions presented by these works in designing the participant questionnaire to assess the perceived acceptance of the technology of users. Finally, to measure satisfaction when using the implemented system, we design questions based on user satisfaction questionnaires [271, 275], based upon a similar 7-point scale.

6.4 A Case Study of Topic Analysis Framework Upon Literature Related to the Human Kinome

Through the application of topic modelling of academic articles, the framework proposed in this section encourages senior researchers within a specific field to act as experts to contribute to the labelling of topics. In addition to this, domain learners (as introduced in Section 5.1) can benefit from visualisation tools intended to assist in the comprehension of vast amounts of academic texts. The approach allows reviewers to identify the *topics*, *trends* as well as *relations between topics* in a given research field. For illustration, this is applied to a case-study of biological texts, specifically texts related to human protein kinases. To further enhance the educational capabilities of the approach, triangulation of external biomedical databases is performed to illustrate how the multi-pronged approach can provide a comprehensive understanding of the research domain.

The process of understanding academic literature is a time-consuming process for both students and professionals. Identifying relevant literature may present difficulties, due to a researcher – especially a new Ph.D. student, a young researcher, or any researcher starting to learn about a new field – not fully knowing which papers, from the vast amount of available information, they should be investigating.

Thus, there is a necessity to enable the process of quickly comprehending such literature. Furthermore, limitations to the amount of work which can be analysed by an individual or team may be encountered, due to the time it takes to understand each item of literature. This is of particular relevance to the biomedical field, where estimates in 2016 proposed that the field observes 3 new publications per minute uploaded to the PubMed database alone [229]. Moreover, the identification and recognition of evidence and named entities within full-text biological literature - the Curator Assistance Problem [229], is limited by the requirement for manual analysis of text by experts within the field, using their own knowledge and intuition.

To address these issues, we propose *an automated approach to assist in the literature review process in **biomedical literature***. By leveraging the findings of Chapter 5, we seek to provide a novel benefit to the academic and research process when learning about a certain new research area. Further to this, to address the Curator Assistance Problem [229], we perform rule-based identification of human-related protein kinases within the literature, for automatic triangulation of multiple external data sources, to assist in the understanding of research. As part of this published work, we provide access to the case-study for reproducibility [3].

The main contributions of this paper are thus: **(1)** Provision of a novel framework for the analysis of literature topics by accounting for semantic relationships and temporal trends of identified topics, which reinforces the findings of Chapter 5 which were identified when addressing RQ2.2 and RQ2.3. **(2)** A capability for the crowdsourcing [276] of domain experts for the labelling and filtering of topics within the literature. Experts such as educational tutors or senior researchers may contribute to the labelling of identified topics based upon their own knowledge. Subsequently, domain learners may then benefit from the visualisation tools provided to assist in their learning on the subject. **(3)** Triangulation of multiple external data sources for the extraction and linking of named kinases present within literature, to assist in the expert labelling and comprehension of topics within the literature, which is aimed at addressing RQ2.4. Although this is a case-study based upon literature related to protein kinases, in collaboration with bioscientists, this robust approach

[3] https://github.com/ryanon4/topic_labelling_tool

begins to demonstrate the generalisable capabilities of such a technique, which is further expanded in the next section when proposing an overall general framework for literature analysis and visualisation in Section 6.5.

6.4.1 Design Methodology

The proposed framework (as summarised in Figure 6.2) entails three components. Firstly, topic modelling is performed on the academic literature that is identified as relevant to the case-study domain through the use of the API resource. Following this, senior researchers are presented with an interactive topic labelling and visualisation analysis tool, which allows them to determine whether a topic is relevant to their literature search, as well as assign a title for that topic. Finally, this visualisation analysis tool can facilitate the comprehension of relevant literature of domain learners. To illustrate the framework, here, on the basis of the full text of documents assigned to a topic, extraction of human protein-kinase terms (the topic of the literature review selected by bioscientist co-authors of this work) is performed by rule-based extraction of kinase names. Also, in terms of actual implementation, an internal relational database provides external links to external resources – here, those coming from UniProt Knowledgebase (UniProtKB) [277] and InterPro [278], which are well-known protein databases. The resulting interface may be applied as a framework for learning through the crowdsourcing of domain experts within a field for the labelling and filtering of literature topics. Expertly annotated topics may then be provided to learners or experts, who may apply the visualisation tools provided to assist in their learning and comprehension of the literature.

Data Sources

A full-text representation of publications is provided through the PubMed API, which ensures to capture the full semantic context of a text if it is analysed by a compatible language model for variable length document-embeddings. Two additional sources were selected for the triangulation of kinases identified within the literature, with these being the UniProtKB [277] and Interpro [278] databases. UniProtKB is a database of protein sequences and functional information, which includes infor-

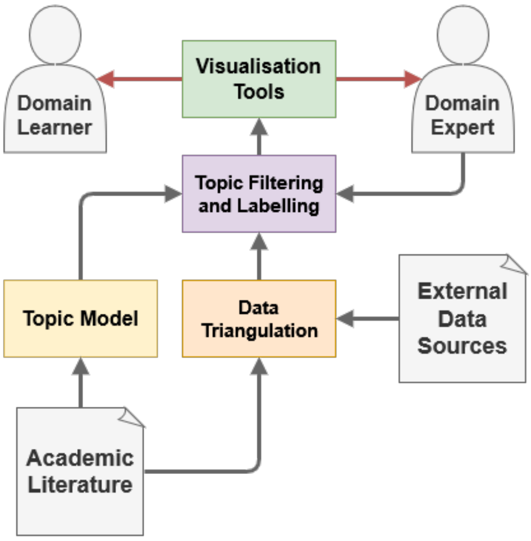
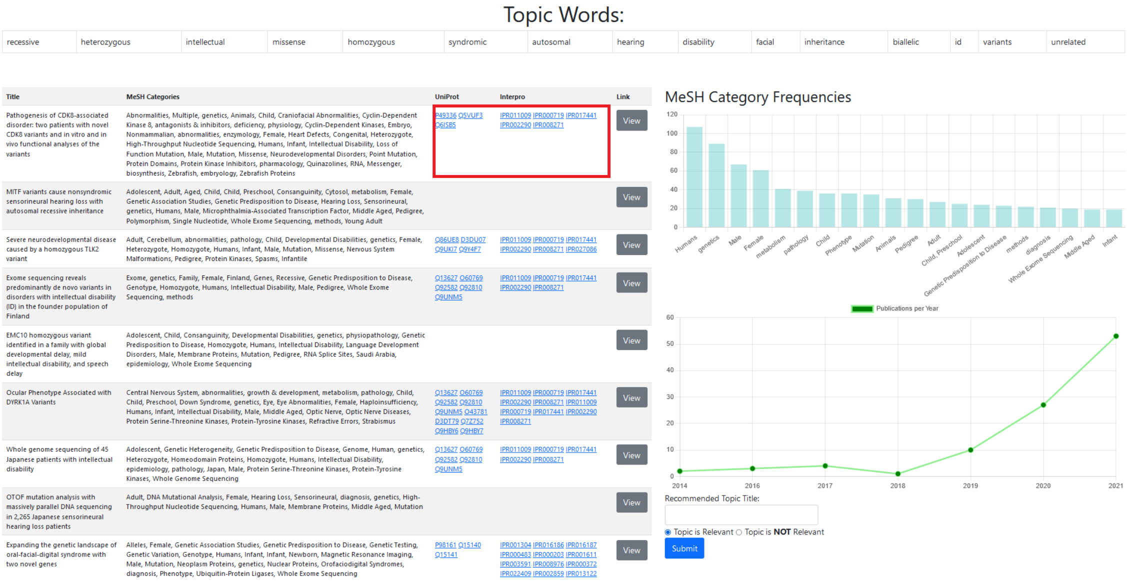


Figure 6.2: An automated framework based on topic modelling and crowdsourcing.

mation relevant to specific proteins, their function, organism, presence in biological processes, and relevant literature. InterPro serves as a similar platform, however, includes hierarchical family classifications of protein entries as well as domains and functional sites found in proteins.

For the case-study, a list of 624 human protein kinases was first obtained from a kinase database [111, 112]. A comprehensive list of human proteins with links to their respective InterPro IDs was obtained from UniProtKB (Swiss-Prot, Jan 2014 version) [277]. Matching of protein kinase names to gene names including synonyms of proteins from UniProtKB was performed. The resulting 357 protein kinases that successfully matched UniProtKB entries were used in this study. This entailed the names of relevant genes, kinase names, and kinase families. In addition to specific kinase names (e.g. *AATK1*), the data included full-named entities relevant to a kinase. Relevant to the given example, this would be *Apoptosis-associated tyrosine kinase 1*. This list of shortened and complete terms was applied in an automated literature search using the PubMed API [115]. A total of 19,258 records were returned in this manner. Following this, filtering was performed to ensure the elimination of any noise introduced into the dataset during the search stage. This was carried out by filtering out publications that failed to contain MeSH terms related to humans. MeSH terms are a standardised classification, similar in manner to a keyword, which are assigned to research published through PubMed to assist in



Topic Words:

recessive	heterozygous	intellectual	missense	homozygous	syndromic	autosomal	hearing	disability	facial	inheritance	biallelic	id	variants	unrelated

Title	MeSH Categories	UniProt	Interpro	Link
Pathogenesis of CDK8-associated disorder: two patients with novel CDK8 variants and in vitro and in vivo functional analyses of the variants	Abnormalities, Multiple, genetics, Animals, Child, Craniofacial Abnormalities, Cyclin-Dependent Kinase 8, antagonists & inhibitors, deficiency, physiology, Cyclin-Dependent Kinases, Embryo, Nonmammalian, abnormalities, enzymology, Female, Heart Defects, Congenital, Heterozygote, High-Throughput Nucleotide Sequencing, Humans, Infant, Intellectual Disability, Loss of Function Mutation, Male, Mutation, Missense, Neurodevelopmental Disorders, Point Mutation, Protein Domains, Protein Kinase Inhibitors, pharmacology, Quinazolines, RNA, Messenger, biosynthesis, Zebrafish, embryology, Zebrafish Proteins	P49336 Q5VUF3 Q6ISB5	IPR011009 IPR000719 IPR017441 IPR002290 IPR008271	View
MITF variants cause nonsyndromic sensorineural hearing loss with autosomal recessive inheritance	Adolescent, Adult, Aged, Child, Child, Preschool, Consanguinity, Cytosol, metabolism, Female, Genetic Association Studies, Genetic Predisposition to Disease, Hearing Loss, Sensorineural, genetics, Humans, Male, Microphthalmia-Associated Transcription Factor, Middle Aged, Pedigree, Polymorphism, Single Nucleotide, Whole Exome Sequencing, methods, Young Adult			View
Severe neurodevelopmental disease caused by a homozygous TLK2 variant	Adult, Cerebellum, abnormalities, pathology, Child, Developmental Disabilities, genetics, Female, Heterozygote, Homozygote, Humans, Infant, Male, Mutation, Missense, Nervous System Malformations, Pedigree, Protein Kinases, Spasms, Infantile	Q86UE8 D3DU07 Q9UKI7 Q9Y4F7	IPR011009 IPR000719 IPR017441 IPR002290 IPR008271 IPR027086	View
Exome sequencing reveals predominantly de novo variants in disorders with intellectual disability (ID) in the founder population of Finland	Exome, genetics, Family, Female, Finland, Genes, Recessive, Genetic Predisposition to Disease, Genotype, Homozygote, Humans, Intellectual Disability, Male, Pedigree, Whole Exome Sequencing, methods	Q13627 O60769 Q92582 Q92810 Q9UNM5	IPR011009 IPR000719 IPR017441 IPR002290 IPR008271	View
EMC10 homozygous variant identified in a family with global developmental delay, mild intellectual disability, and speech delay	Adolescent, Child, Consanguinity, Developmental Disabilities, genetics, physiopathology, Genetic Predisposition to Disease, Homozygote, Humans, Intellectual Disability, Language Development Disorders, Male, Membrane Proteins, Mutation, Pedigree, RNA Splice Sites, Saudi Arabia, epidemiology, Whole Exome Sequencing			View
Ocular Phenotype Associated with DYRK1A Variants	Central Nervous System, abnormalities, growth & development, metabolism, pathology, Child, Child, Preschool, Down Syndrome, genetics, Eye, Eye Abnormalities, Female, Haploinsufficiency, Humans, Infant, Intellectual Disability, Male, Middle Aged, Optic Nerve, Optic Nerve Diseases, Protein Serine-Threonine Kinases, Protein-Tyrosine Kinases, Refractive Errors, Strabismus	Q13627 O60769 Q92582 Q92810 Q9UNM5 O43781 D3DT79 Q7Z752 Q9HBY6 Q9HBY7	IPR011009 IPR000719 IPR017441 IPR002290 IPR008271 IPR011009 IPR000719 IPR017441 IPR002290 IPR008271	View
Whole genome sequencing of 45 Japanese patients with intellectual disability	Adolescent, Genetic Heterogeneity, Genetic Predisposition to Disease, Genome, Human, genetics, Heterozygote, Homeodomain Proteins, Homozygote, Humans, Intellectual Disability, epidemiology, pathology, Japan, Male, Protein Serine-Threonine Kinases, Protein-Tyrosine Kinases, Whole Genome Sequencing	Q13627 O60769 Q92582 Q92810 Q9UNM5	IPR011009 IPR000719 IPR017441 IPR002290 IPR008271	View
OTOF mutation analysis with massively parallel DNA sequencing in 2,265 Japanese sensorineural hearing loss patients	Adult, DNA Mutational Analysis, Female, Hearing Loss, Sensorineural, diagnosis, genetics, High-Throughput Nucleotide Sequencing, Humans, Male, Membrane Proteins, Middle Aged, Mutation			View
Expanding the genetic landscape of oral-facial-digital syndrome with two novel genes	Alleles, Female, Genetic Association Studies, Genetic Predisposition to Disease, Genetic Testing, Genetic Variation, Genotype, Humans, Infant, Infant, Newborn, Magnetic Resonance Imaging, Male, Mutation, Neoplasm Proteins, genetics, Nuclear Proteins, Orofaciodigital Syndromes, diagnosis, Phenotype, Ubiquitin-Protein Ligases, Whole Exome Sequencing	P98161 Q15140 Q15141	IPR001304 IPR016186 IPR016187 IPR000483 IPR000203 IPR001611 IPR003591 IPR008976 IPR000372 IPR022409 IPR002859 IPR013122	View

MeSH Category Frequencies

Recommended Topic Title:

● Topic is Relevant ○ Topic is **NOT** Relevant

Submit

Figure 6.3: Labelling interface for topic “Genetic Disorders”.

the categorisation of research. Following elimination of non-human-related subjects, 14,631 documents remained.

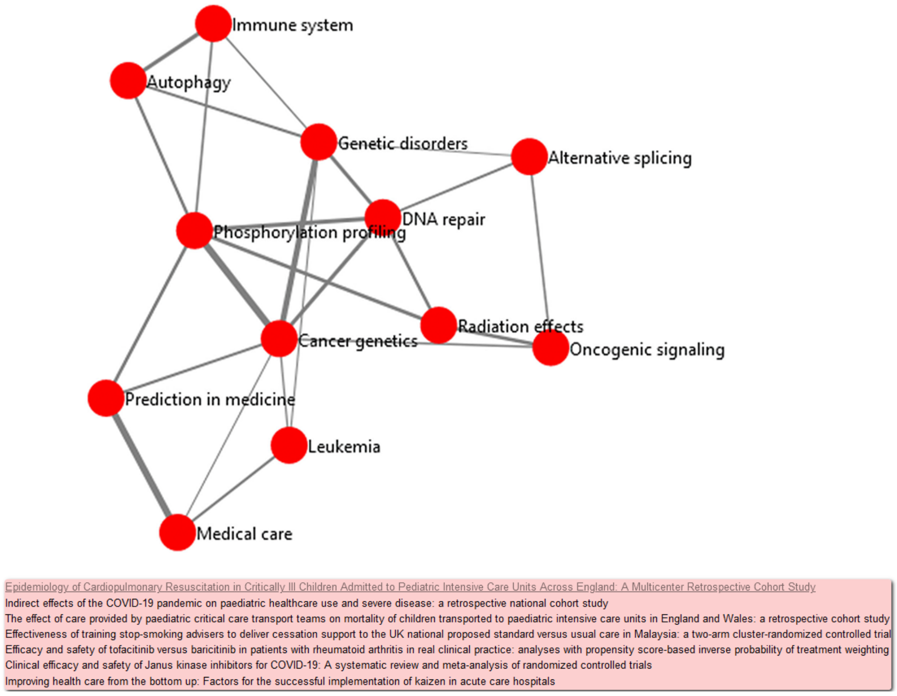
Topic Modelling

For the identification of topics within the literature, the Top2Vec [6] algorithm, which leverages the HDBSCAN [79] algorithm, is applied to identify dense clusters of document embeddings, and determine these to be topics, as detailed in Section 3.2.3. Human-readable labels for topics are generated based on the top-scoring topic words for each topic. To expand this functionality, a domain-expert may manually assign a title for each topic on the topic labelling screen. The use of HDBSCAN presents a unique opportunity when clustering, given that HDBSCAN labels sparse areas of documents during clustering as noise [79]. This provides the benefit of eliminating documents that do not fall directly within a topic.

Literature Analysis Toolkit

The main contribution of this aspect of the work centres around the development of a simple user interface to evaluate the results of the literature topic analysis and subsequent deeper analysis. Researchers may find difficulty in the application of algorithmic analysis approaches, as these require a degree of knowledge of programming and an understanding of the underlying algorithms applied. Therefore, by providing an interactive user interface, we ensure that researchers in the biomedical domain can analyse the literature without the need to spend time learning the underlying system. Results from the topic modelling of the literature are presented to the domain-expert in a simple landing page, which provides a table format of top-scoring words for each topic, the manually assigned label, and relevancy classification which can be manually defined. By clicking on the *Edit Label* button, domain-experts can view a detail page (as shown in Figure 6.3), featuring top-scoring publications assigned to a topic, the top-scoring words for the topic, and a bar graph visualisation, detailing the total frequency of MeSH terms for publications assigned to that topic. Providing these features presents domain-experts with an opportunity to decide the title of the topic through a number of means. For example, MeSH subject headings can provide the frequency with which certain diseases are mentioned in documents on a topic. Furthermore, by extracting mentions of human protein kinases within full-text articles, we facilitate the triangulation of the InterPro [278] and UniProtKB [277] databases, and can provide domain-experts with external links to related resources relative to a given publication (as shown in the highlighted box in Figure 6.3). Domain-experts may also view the full publication at the place of publishing through a hyperlink, so that individual publications they find useful through the above analysis methods may then be read further.

Further academic features are provided to users (both domain experts and learners) through an analysis of the semantic relationships within the topics, following the same methodology presented by [279] in Section 5.8. Semantic relationships are generated through the calculation of the cosine similarity between all combinations of topic vectors, where a topic vector is defined as the average of all document embeddings belonging to that topic. We define each topic as a node, with the top three highest scoring topic similarity pairs (edges) for each topic being presented in the graph visualisation. The nodes have at least three edge relationships. However, some nodes may have more than three edges. For instance, as shown in Figure 6.4, the node "*DNA repair*" displays the top 3 scoring edge relationships with "*Genetic disorders*," "*Cancer genetics*" and "*Phosphorylation profiling*". However, for the



Epidemiology of Cardiopulmonary Resuscitation in Critically Ill Children Admitted to Pediatric Intensive Care Units Across England: A Multicenter Retrospective Cohort Study
Indirect effects of the COVID-19 pandemic on paediatric healthcare use and severe disease: a retrospective national cohort study
The effect of care provided by paediatric critical care transport teams on mortality of children transported to paediatric intensive care units in England and Wales: a retrospective cohort study
Effectiveness of training stop-smoking advisers to deliver cessation support to the UK national proposed standard versus usual care in Malaysia: a two-arm cluster-randomized controlled trial
Efficacy and safety of tofacitinib versus baricitinib in patients with rheumatoid arthritis in real clinical practice: analyses with propensity score-based inverse probability of treatment weighting
Clinical efficacy and safety of Janus kinase inhibitors for COVID-19: A systematic review and meta-analysis of randomized controlled trials
Improving health care from the bottom up: Factors for the successful implementation of kaizen in acute care hospitals

Figure 6.4: Relationships of topics identified by the bioscientist Team and expanded link view

"*Radiation effects*" and "*Alternative splicing*" nodes, the "*DNA repair*" edge relationship is within the top-3 ranking for the nodes, respectively, hence "*DNA repair*" demonstrates 5 edge relationships.

The topic-relationship graph is computed when initialising the system, and stored in a serialised format. Edge thickness is a reflection of the cosine-similarity score between two nodes, for easy understanding of the strength of the relationship between those topics, with a tooltip presenting the similarity score when the user hovers their mouse over an edge. From this view, users may also further expand a topic, navigating to a view of the top-scoring academic texts that have been assigned to that topic based upon the distance of the publication to the topic centroid, by right-clicking on a node. For example, the pink panel on the bottom of Figure 6.4 is displayed when right-clicking the node "*Medical care*". This panel provides links of the paper directly to the external resources. Users may choose to view all identified topic-relationships from the topic modelling process, or only those that have been manually labelled as relevant.

6.4.2 Case Study in Biology - Analysis by a Domain Expert on the Subject of Human Protein Kinases

As a case study, the toolkit generated 279 topics from publications' texts related to the 357 human protein kinases. A representative list of associations is provided, for each topic, and contains the top 20 highest scoring topic words, the top ten individual publications, and MeSH category frequencies of publications assigned to the topic. The rest of this section presents a discussion provided by a domain expert in the biomedical field, evaluating their findings from the topic modelling and semantic mapping process with respect to their domain knowledge. To highlight aspects of the analysis and the justification by the domain-expert based on their knowledge, all responses by the expert are italicised:

The 10*th* *topic (shown in Figure 6.3) provides the following information: relevant topic words such as "recessive," "heterozygous," "inheritance," "missense," "homozygous," "syndromic," "autosomal," "disability," and "variants"; titles of individual publications with phrases such as "neurodevelopmental disease caused by a homozygous TLK2 variant," "variants in disorders with intellectual disability," and "phenotype associated with DYRK1A variants"; and MeSH categories such as "Humans" and "Genetics" with the highest frequency in the assigned publications. Based on this information, the topic can be labeled as "Genetic disorders" associated with human protein kinases. Moreover, links to UniProtKB and InterPro databases are listed for specific kinases included in each individual publication. For example, the publication titled "Pathogenesis of CDK8-associated disorder: two patients with novel CDK8 variants and in vitro and in vivo functional analyses of the variants" is provided with the following links: the entry of the kinase CDK8 in UniProtKB; and the entries of domains, binding-sites, etc. found in CDK8 such as "Protein kinase domain," "Protein kinase, ATP binding site," and "Serine/Threonine protein kinases active-site" in InterPro.* Such external databases provide detailed information such as biological functions, disease association, sequences, and domain architectures of protein kinases along with links to further explore other databases, should such a need arise for the user.

The toolkit also visualises relationships among topics based on their semantic similarity. To demonstrate its efficacy, the bio-experts taking part in this study selected 12 topics that have at least six out of ten individual publications with links to protein kinases in UniProtKB, and then manually labelled the topics as shown in the previous paragraph. The resulting graph network (shown in Figure 6.4) generated from these topics has 12 nodes with the assigned labels and edges with different thickness that reflects the similarity between the linked topics. The node "Genetic disorders" is strongly connected to the nodes of "Cancer genetics" and "DNA repair." This is consistent with the idea that both genetic disorders and cancer genetics involve genetic changes and genomic technologies. Similarly, DNA repair defects are associated with certain genetic disorders. For example, ATM and ATR are DNA repair-associated serine/threonine kinases and their genetic mutations can cause hereditary disorders, ataxia telangiectasia and Seckel syndrome, respectively. Thus, it is reasonable to include both ATM and ATR in topic-words for the topic "DNA repair." Another strong connection visualised by the network is between "Medical care" and "Prediction in medicine," which aligns with the clinical aspect of treatment. The visualisation enables the users to easily explore major topics and their relationships. Collectively, the toolkit allows users to navigate a new field through vast amounts of literature in an efficient way.

6.4.3 Section Summary

In this work, an automated literature review framework was proposed for the analysis of large volumes of literature. The resulting topics from topic modelling of academic literature within specific domain, are provided via an interactive UI tool, which may allow the manual analysis and filtering of resulting topics, and exploration of entailing literature. From an educational data mining perspective, the proposed framework achieves the goal of grouping large amounts of academic text into understandable topics, and allows for crowdsourcing of expert-knowledge for the labelling of topics identified in the literature. The resulting framework achieves two objectives by serving as a base literature review assistance tool, or as an e-learning utility to allow expert analysis of topics, before providing the results to learners. The integration of external databases such as UniProtKB and InterPro facilitates the demonstration of how the extraction of named kinases from individual pieces of literature enables users to further probe into the identified kinases, thereby addressing RQ2.4. Overall, this framework incorporates topic modelling with a crowd-sourcing approach, achieving the goal of expediting the task of analysing large volumes of literature.

The findings of this research have demonstrated a preliminary exploration of the provision of an automated literature analysis framework, which was based upon the findings of the previous chapter, where it was identified that topic modelling algorithms were a suitable technique for the analysis of academic literature. Although the visualisation functionality was able to be demonstrated as having value to researchers, open questions remain regarding the generalisability of the approach to any domain of literature, rather than a specific case-study. Therefore, in the next section, a complete, final framework is proposed, based on the research contained within this thesis, and evaluated with regard to a study of medical experts. This sets forth to address the final research questions of this section; RQ2.5, RQ2.6, and RQ2.7.

6.5 A co-designed Framework for the Analysis of Large Volumes of Biomedical Literature

From an educational data mining perspective, the initial exploratory work [224,260] presented in Sections 5.3 and 6.4 achieved the goal of grouping large amounts of academic text into understandable topics and allowed the crowd-sourcing of expert knowledge for the labelling of topics identified in the literature. This previous work served as a demonstration of the feasibility and practicability of the proposed approach, which relied on feedback from a small group of biomedical experts. Although promising, it is necessary to take into account that to be robust enough to be adopted by medical professionals, a literature analysis framework must demonstrate the potential for analysis of any literature search determined by a user, without the need for prior programming or setup. Moreover, the framework that is put into practice should contain sufficient information to facilitate knowledge discovery effectively, while at the same time avoiding overwhelming the user's working memory. This is crucial as an excess of information could hinder the usefulness of the framework if users are unable to mentally digest the information presented [263–265].

Thus, the research described in this section aims to address research questions 2.5, 2.6 and 2.7, by focusing on exploring an end-to-end framework which can assist in literature analysis:

- **RQ2.5:** *Can an end-to-end framework, built upon cluster-based topic modelling confer a benefit to researchers through assisting the literature review process?*
- **RQ2.6:** *What are the effects on cognitive load, technology acceptance, and general acceptance of users when using a software implementation based on our framework?*
- **RQ2.7:** *How does the proposed literature analysis framework and software implementation influence the accuracy of the identified results in practical use cases for medical researchers?*

Our work presents the following contributions:

Through discussion with medical experts, we *co-design essential functions with medical professionals*, to assist with finer-grained literature searches and analysis. Namely, we propose the *first comprehensive automatic review framework for the medical literature*, able to accommodate the specific needs of medical researchers through a *human-AI hybrid approach*, enabling the recording of self-labelled research topics at the fine-grained level of individual publications. We name this *open-source aSsisted MedicAl LiteRaTurE aNalysis* framework SMARTEN. This introduces several concepts that have not yet been utilised in the field of literature analysis in this manner. These ideas involve allowing the generation of user-defined queries for analysing different areas, providing search alias groups that allow medical practitioners to employ conditional logic in their searches, and categorising identified literature according to the semantic content of each piece, facilitating filtering by users.

In addition to the proposed SMARTEN framework, we implemented and evaluated in real-life settings a software implementation (the SMARTEN system). We conducted an experimental evaluation addressing SMARTEN system with 18 medical experts, to explore how the *large volume* of biomedical literature is consolidated, in a semantically-relevant way; the analysis results confirm the feasibility and practicality of the SMARTEN framework and contributes to highlighting future improvements of the literature assistant in this framework.

6.6 Methodology

Building on the preliminary findings of previous work [224, 260], in this paper, we propose *a framework for the analysis of large volumes of biomedical literature*, which addresses the needs of medical professionals.

To address the research questions set forth in this work, the following methodology was formulated, with the overall contribution addressing the research questions being the end-to-end literature analysis framework presented in this work. First, in tackling the RQ2.5, it is necessary to devise a general framework that permits the (semi-automatic) analysis of large volumes of academic literature, via a *human-AI hybrid* model. To address this, we set forth a co-design strategy with medical experts, the target users of the framework, to devise a suitable framework structure that will present the most benefit to their requirements. Building on previous work on the visualisation of topics in the literature identified by cluster-based topic modelling [224], we engage with medical experts to identify several aspects of usability and visualisation that were specified as beneficial to any downstream analyses required by medical experts, these are presented in Section 6.6.1. The general structure of the final framework is presented in Section 6.7.

In addressing RQ2.6, we focus on an experimental evaluation of the overall framework, through a software implementation of the proposed framework (SMARTEN) and a subsequent evaluation study involving 18 users from the medical domain. Participants were required to first complete a literature analysis task in their chosen domain and then complete a questionnaire related to their perception of cognitive load [263], technology acceptance, and general satisfaction when using the framework [271, 280]. An analysis of user feedback is detailed in Section 6.7.7, including the basis for the design of the questionnaire with regard to cognitive load and the assessment of acceptance of technology.

Finally, in evaluating RQ2.7, which focuses on the accuracy of the identified topics in relation to the use cases of medical researchers, we assess the accuracy of the framework, evaluating the labels assigned by the study participants, compared to the number of papers recommended by the framework. The method for this is detailed in Section 6.7.7.

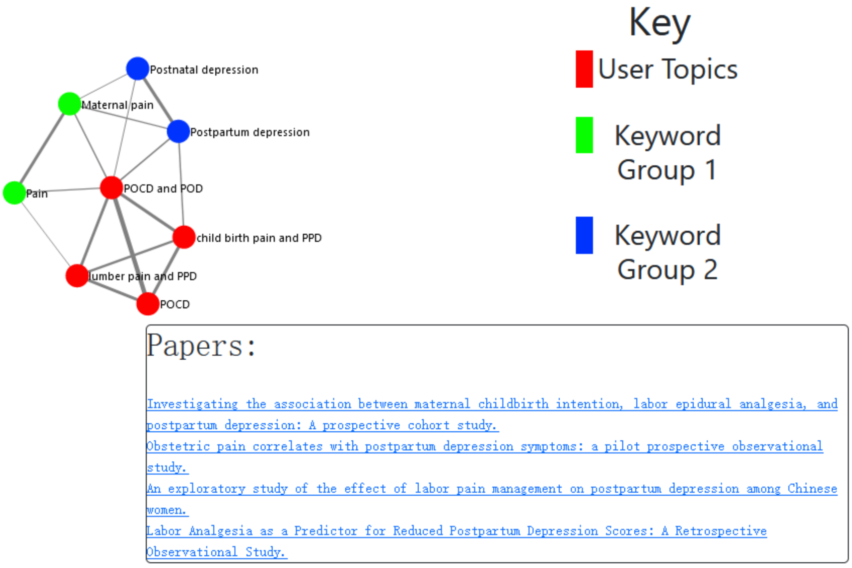
6.6.1 Co-design of the SMARTEN Framework

A co-design methodology was adopted in designing the SMARTEN framework, to ensure that direct benefit could be made to medical experts. Building upon the findings from previous studies [224], Section 6.4, we engaged in multiple rounds of discussions with a medical expert, to identify how to improve the process of searching the relevant literature. This co-design process consisted of several stages of discussion, with feedback provided by medical experts contributing to design changes. This was followed by a wider study of the effects of the design choices, based on user feedback. For the first part of this process, we conducted the co-design discussion with a medical expert, chair of the experimental focus group, and paper co-author. The medical expert would liaise with colleague doctors and medical students based on their evaluation of the system development, and identify functionality which they requested or modified. Based on this feedback, the requirements for the implementation were defined. In general, this co-design approach closely resembled a SCRUM software development method [281]. It involved sprint iterations when designing specific framework functionality, followed by a feedback session with the medical expert once they had access to the most recent version of the software framework. This allowed them to give feedback and pinpoint features that would enhance their analysis. These are defined in the follow-up sub-sections, leading to important features of the framework.

Accounting for Synonyms During Literature Search

Amongst outcomes, the medical expert underscored the potential advantages of implementing a *multi-keyword search capability*. In the medical domain, one keyword may have several synonyms; therefore, it is important to provide a function that allows the user to specify the aliases of each keyword. This was taken into consideration when designing the framework, to facilitate the extraction of relevant literature, ensuring that users may provide search terms in groups, so that within each group, multiple synonyms (or aliases) can be provided. For example, a user may wish to create a search group consisting of a keyword *"Maternal Pain"*, and its alias *"Pain during pregnancy"*. An additional outcome of the co-design methodology was the expressed desire of medical professionals to amalgamate these alias groups, facilitated through the use of logical operators. Based on the criteria provided by the user, a query can be constructed using logical operators that are used to query the PubMed literature. In the case above, assuming that the user creates another keyword *"Postpartum depression"* and its alias *"Postnatal depression"*, the final request could be extended to the following query: *"(Maternal Pain OR Pain during pregnancy) AND (Postpartum depression OR Postnatal depression)"*. Users can enter these query terms when using the SMARTEN system, as shown in Figure 6.6. Although this functionality is already provided by some online literature search platforms, such as the PubMed web portal[4], we extend this functionality through subsequent analyses that are facilitated by the SMARTEN framework, by automatically consolidating large amounts of literature into topics and allowing the analysis of semantic relationships of the aliases provided by users and comparison of these with topics identified within the literature, as shown in Figure 6.5. We provide a deeper explanation of this functionality in Section 6.7.6.

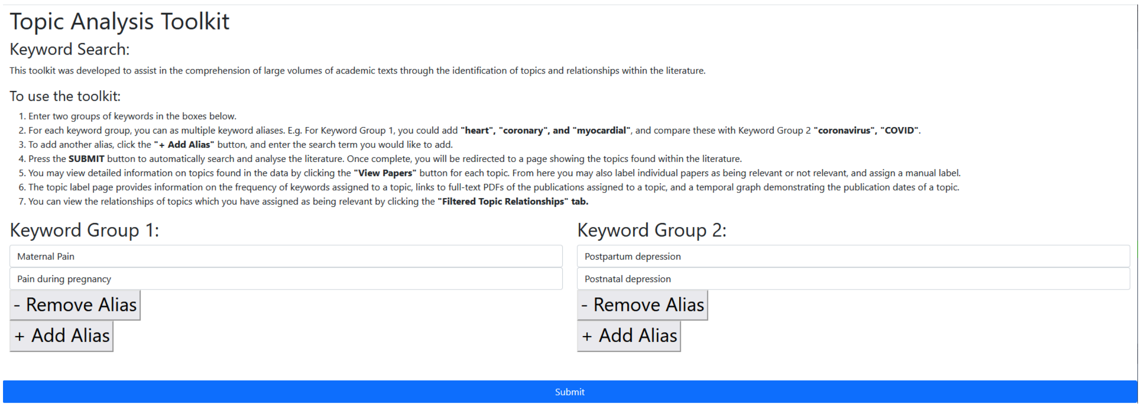
[4] `https://pubmed.ncbi.nlm.nih.gov/advanced/`



Papers:

Investigating the association between maternal childbirth intention, labor epidural analgesia, and postpartum depression: A prospective cohort study.
Obstetric pain correlates with postpartum depression symptoms: a pilot prospective observational study.
An exploratory study of the effect of labor pain management on postpartum depression among Chinese women.
Labor Analgesia as a Predictor for Reduced Postpartum Depression Scores: A Retrospective Observational Study.

Figure 6.5: The graph relationships view within SMARTEN, showing topic relationships based on semantic similarity.



Topic Analysis Toolkit

Keyword Search:

This toolkit was developed to assist in the comprehension of large volumes of academic texts through the identification of topics and relationships within the literature.

To use the toolkit:

1. Enter two groups of keywords in the boxes below.
2. For each keyword group, you can as multiple keyword aliases. E.g. For Keyword Group 1, you could add **"heart", "coronary", and "myocardial"**, and compare these with Keyword Group 2 **"coronavirus", "COVID"**.
3. To add another alias, click the **"+ Add Alias"** button, and enter the search term you would like to add.
4. Press the **SUBMIT** button to automatically search and analyse the literature. Once complete, you will be redirected to a page showing the topics found within the literature.
5. You may view detailed information on topics found in the data by clicking the **"View Papers"** button for each topic. From here you may also label individual papers as being relevant or not relevant, and assign a manual label.
6. The topic label page provides information on the frequency of keywords assigned to a topic, links to full-text PDFs of the publications assigned to a topic, and a temporal graph demonstrating the publication dates of a topic.
7. You can view the relationships of topics which you have assigned as being relevant by clicking the **"Filtered Topic Relationships" tab.**

Keyword Group 1:

Maternal Pain

Pain during pregnancy

Keyword Group 2:

Postpartum depression

Postnatal depression

Figure 6.6: The keyword search screen in SMARTEN

Identification of Meta-Information within Publications

During discussions with the medical expert, we emphasised the potential of extracting labelled information from publication *metadata*, which is associated with each entry within the PubMed database. Specifically, we introduced various types of information that could be gleaned from metadata, such as publication keywords, MeSH terms, year of publication, authors and named chemicals or medicines discussed in a publication. Subsequently, the medical expert identified which of these data points would enhance their analytical processes. For this, the experts identified that the presence of *MeSH terms, chemicals, publication keywords, and year of publication* as being beneficial in contributing to their understanding of topics identified within the literature.

Defining the Number of Publications to Return

In early versions of the literature analysis framework, it was highlighted by the medical expert during co-design that providing a high number of publications to the user of the system can overwhelm users, making it difficult to identify publications that would be useful in their search. Thus, one of the outcomes of the co-design process was the definition of providing only a *limited subset* - the *threshold* was set at 10, to retrieve the ten most relevant publications identified as belonging to each topic, where relevance is defined as the cosine-similarity of the publication embedding relative to the cluster centroid of the topic. The threshold of 10 publications per topic was established based on feedback obtained during the co-design process, reflecting the consideration that labeling publications within topics is a time-intensive task. While the time required to read and comprehend research publications differs by domain and user intentions, it is reasonable to argue that the evaluation of many publications would be time-intensive for participants. Moreover, requiring participants to label a large number of papers could potentially impact subsequent evaluations of cognitive load measurements captured in the questionnaire. While this threshold of 10 was established partially based on the considerations for subsequent evaluation questionnaires, it is worth noting that subsequent iterations of the SMARTEN framework could be configured to yield all publications associated with a given topic. This approach would ensure comprehensiveness in literature analysis, thereby mitigating the risk of omitting pertinent information.

Viewing Biomedical Publications and Associated Meta-Data

An earlier version of the automatic literature review proposal considered the option of viewing publications. The expert highlighted during co-design that for biomedical publications, not only the link to the full article is necessary, but also a display of the publication year and the full MeSH category information for each publication, as also mentioned above.

6.7 Instantiation of a Comprehensive Automatic Literature Review Framework for Biomedical Literature

Here, we define an overall structure for the literature analysis framework. This can be broken down into four distinct stages, which are numerically labelled in Figure 6.7. In the first stage, a user provides their search terms via the search page interface, which allows the provision of groups of keywords, as detailed in section 6.7.2. These search terms are then structured into an API request for the PubMed database API, which will be used to retrieve all publication data in a structured XML format. Secondly, the response for this API query is parsed, with the full-text and metadata for each publication being extracted. Publication texts are then analysed using the Top2Vec [16] algorithm, considered as most appropriate as justified in section 6.3.1; with an overview of the topics identified within the literature then presented to the user via the topic overview screen. At the third stage, a user may view the publications which have been assigned to each topic on the topic details page, along with topic metadata consisting of the frequencies of keywords, MeSH terms, chemicals, and volume of publications based on publication date, which were defined during the co-design process with the medical expert focus group. Based on this information, a user labels publications as relevant, which they feel are suitable to their analysis. Finally, the fourth stage entails the processing of the labelling provided by a user into a graph network-based visualisation, based upon the semantic similarity of the topics labelled by the user.

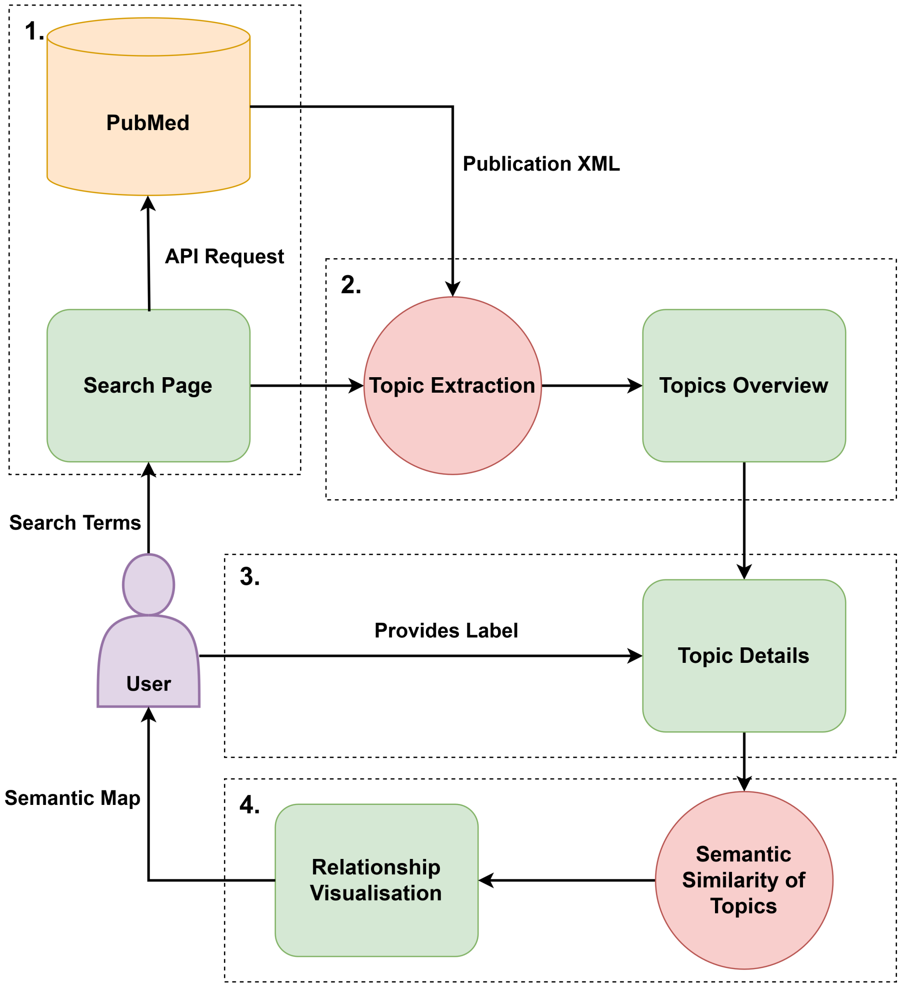


Figure 6.7: The stages of the comprehensive automatic biomedical literature review framework

6.7.1 A Brief Overview of Data-Persistence in SMARTEN

The SMARTEN framework is implemented as a web-based platform, which allows participants to access the system using a unique account via their browser. After providing their initial search query in stage one of the framework, queries are applied in a server-side API request to the PubMed database, to perform retrieval of relevant literature. These literature searches and the associated query are stored on the web server, to ensure permanence of results should the participant choose to end their analysis and come back at a later date. After the collection of relevant literature and the extraction of text from publication XML, the topic modelling process is started automatically, without the need for intervention by the participant. As with the collection of relevant literature, persistence of the topic modelling process is ensured.

This is particularly important, given that the topic extraction can take several minutes, and preparation should be taken in case the participant loses connection to the platform or ends their session. For stage three of the framework, any labels provided by users during their analysis is stored within an internal database using the unique user ID, which again allows them to revisit their labelling later. Finally, the fourth stage of the framework, which involves the generation of a relationship mapping, is performed in real-time, given that users may revisit their publication labels at any time, and thus the relationship mapping must be recomputed after each change to the labelling configuration.

Generally, throughout the implementation of the SMARTEN framework, consideration is made to ensure that data persistence is maintained to ensure that users may revisit the results of their analysis at any time. However, some aspects of permanence were not considered at this stage of investigation, which focuses on evaluating the efficacy of the framework. These include the caching of existing literature search results to ensure enhanced speed during the search stage for popular queries, and the maintenance of a history of analyses for users. The latter of would enable users to perform an analysis, save the results, and then perform another analysis with the ability to revisit any historical analyses. Both of these points will, however, be considered in future adaptations of this research, where it could be applied to either literature analysis, or, from a publishing context, the analysis of journalistic material and social media sources.

6.7.2 Search and Retrieval via the Biomedical Review Framework

As defined by the co-design process in section 6.6.1, implementing stage 1 of the framework constructed in section 6.7, search terms created by users using the interface depicted in Figure 6.6, are provided as a query to the PubMed repository through Entrez E-Utility [115], for the retrieval of relevant publications. Publication results are provided by the PubMed API in a structured eXtensible Markup Language (XML), which provides publication data in a structured format, which includes textual content and metadata associated with publications. Metadata includes information such as the keywords assigned to a publication by its authors, MeSH terms, publication dates, and citations, to name a few (a full summary of the available data can be found at [282]). The extraction of the publication XML is used to identify the title, text content, publication data, MeSH terms, keywords, and chemicals associated with each publication.

6.7.3 Topic Modelling of Identified Literature - Framework Design and Instantiation in SMARTEN

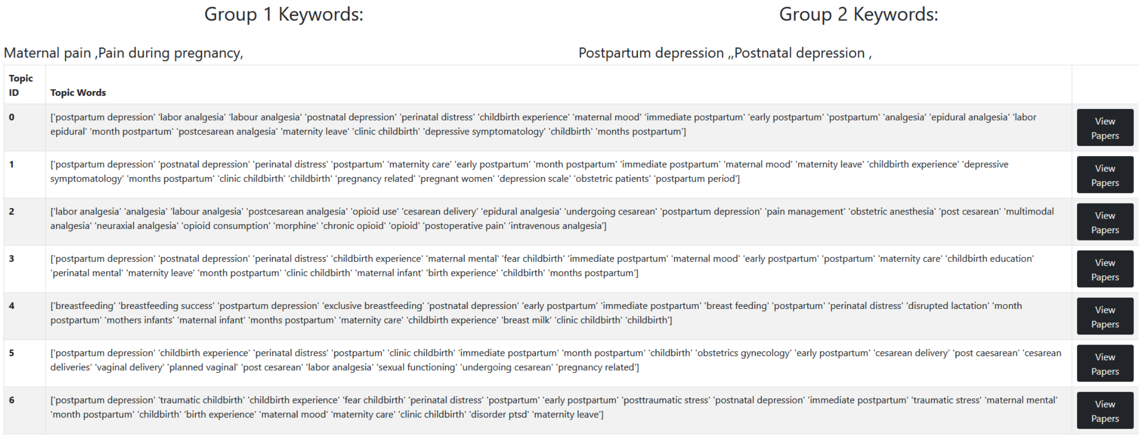
Next, as defined by the co-design process in section 6.6.1, reflecting stage 2 of the framework proposed in section 6.7, the identification of topics from within the literature is performed using hierarchical clustering of document embeddings, which is facilitated using the Top2Vec algorithm [16]. As discussed in section 6.3.1, publication title and text are first converted into an embedding (a numerical representation that encapsulates the semantic information of the text in a highly-dimensional vector) by a language model. These highly dimensional representations cannot easily be used in downstream tasks such as clustering, as downstream algorithms may be negatively affected by the "curse of dimensionality". Therefore, dimensionality reduction needs to be performed. This is achieved with the current state-of-the-art UMAP [14] algorithm, providing not only a method for the dimensionality reduction of highly dimensional document embeddings into low-dimensionality representations, but also preserving the global and local features of the embeddings. This allows for a significantly smaller dimensionality, to assist in the downstream task of clustering, which is performed upon these document embeddings. For clustering, the Top2Vec algorithm applies the HDBSCAN algorithm, which performs hierarchical clustering based on the density of the vectors produced by UMAP.

It is worth noting that HDBSCAN is a noise-aware clustering algorithm, which performs labelling of sparse areas of points as outliers and thus discounts these documents from being assigned to a cluster. In a corpus of academic literature, it can be reasonably assumed that the noise level within the data is relatively low. Consequently, automated assignment of documents within the corpus as outliers during clustering may result in the omission of a significant amount of valuable information.

This, in turn, could lead to the inadequate representation of certain topics present within the literature during topic modelling. However, several considerations underscore the value of noise-aware clustering when working with academic literature. First, the domain-specific language used in the literature can differ considerably between publications where terms are used interchangeably or have different meanings depending on the context in which they are used. Similarly, acronyms used in a different context can have drastically different meanings, with "LDA' within the computing domain being interpreted as either *Linear Discriminant Analysis*, or *Latent Dirichlet Allocation* as one example. While context-aware embedding models may address this issue, the adoption of Doc2Vec embeddings to tackle the problem of domain-specific language, as discussed in Section 5.3.4, fails to account for the full context of a text due to the inherent limitation of Doc2Vec in disregarding the ordering of terms within a document.

As a second consideration why noise-aware clustering remains valuable, the phrasiology of the term *noise* should be considered. Using the term "outlier" instead of "noise" when discussing outlier detection in HDBSCAN is more precise and contextually appropriate. "Outlier" specifically refers to data points that deviate significantly from the rest of the dataset, indicating they do not fit well within any cluster. This emphasises the statistical anomaly aspect of these points, making it clear that they are exceptions rather than random disturbances. On the other hand, "noise" implies random, irrelevant data that might not necessarily be statistically significant anomalies. As HDBSCAN identifies outliers based on their density and distance from core clusters [79], it is reasonable to argue that during topic modelling, the elimination of outlier points which cannot be clearly assigned to a cluster will ensure to produce topics which are closely related in terms of semantic similarity.

The resulting clusters represent topics identified in the literature. These topics can then be presented to the user as a topic overview page, as presented in Figure 6.8.



Group 1 Keywords:

Maternal pain ,Pain during pregnancy,

Group 2 Keywords:

Postpartum depression ,,Postnatal depression ,

Topic ID	Topic Words	
0	['postpartum depression' 'labor analgesia' 'labour analgesia' 'postnatal depression' 'perinatal distress' 'childbirth experience' 'maternal mood' 'immediate postpartum' 'early postpartum' 'postpartum' 'analgesia' 'epidural analgesia' 'labor epidural' 'month postpartum' 'postcesarean analgesia' 'maternity leave' 'clinic childbirth' 'depressive symptomatology' 'childbirth' 'months postpartum']	View Papers
1	['postpartum depression' 'postnatal depression' 'perinatal distress' 'postpartum' 'maternity care' 'early postpartum' 'month postpartum' 'immediate postpartum' 'maternal mood' 'maternity leave' 'childbirth experience' 'depressive symptomatology' 'months postpartum' 'clinic childbirth' 'childbirth' 'pregnancy related' 'pregnant women' 'depression scale' 'obstetric patients' 'postpartum period']	View Papers
2	['labor analgesia' 'analgesia' 'labour analgesia' 'postcesarean analgesia' 'opioid use' 'cesarean delivery' 'epidural analgesia' 'undergoing cesarean' 'postpartum depression' 'pain management' 'obstetric anesthesia' 'post cesarean' 'multimodal analgesia' 'neuraxial analgesia' 'opioid consumption' 'morphine' 'chronic opioid' 'opioid' 'postoperative pain' 'intravenous analgesia']	View Papers
3	['postpartum depression' 'postnatal depression' 'perinatal distress' 'childbirth experience' 'maternal mental' 'fear childbirth' 'immediate postpartum' 'maternal mood' 'early postpartum' 'postpartum' 'maternity care' 'childbirth education' 'perinatal mental' 'maternity leave' 'month postpartum' 'clinic childbirth' 'maternal infant' 'birth experience' 'childbirth' 'months postpartum']	View Papers
4	['breastfeeding' 'breastfeeding success' 'postpartum depression' 'exclusive breastfeeding' 'postnatal depression' 'early postpartum' 'immediate postpartum' 'breast feeding' 'postpartum' 'perinatal distress' 'disrupted lactation' 'month postpartum' 'mothers infants' 'maternal infant' 'months postpartum' 'maternity care' 'childbirth experience' 'breast milk' 'clinic childbirth' 'childbirth']	View Papers
5	['postpartum depression' 'childbirth experience' 'perinatal distress' 'postpartum' 'clinic childbirth' 'immediate postpartum' 'month postpartum' 'childbirth' 'obstetrics gynecology' 'early postpartum' 'cesarean delivery' 'post caesarean' 'cesarean deliveries' 'vaginal delivery' 'planned vaginal' 'post cesarean' 'labor analgesia' 'sexual functioning' 'undergoing cesarean' 'pregnancy related']	View Papers
6	['postpartum depression' 'traumatic childbirth' 'childbirth experience' 'fear childbirth' 'perinatal distress' 'postpartum' 'early postpartum' 'posttraumatic stress' 'postnatal depression' 'immediate postpartum' 'traumatic stress' 'maternal mental' 'month postpartum' 'childbirth' 'birth experience' 'maternal mood' 'maternity care' 'clinic childbirth' 'disorder ptsd' 'maternity leave']	View Papers

Figure 6.8: The Topic Overview Screen in SMARTEN

In the instantiation of the Comprehensive Automatic Literature Review Framework for Medical Literature, the SMARTEN system, users can then view detailed information on a particular topic identified within the literature by the Top2Vec algorithm, by clicking the *"View Papers"* button to the right of each individual topic (Figure 6.9). This view provides users with a number of metrics, to assist in their comprehension of the literature within the topic, defined based on feedback during the co-design process (section 6.6.1): 1. A breakdown of the general keywords that can be used to describe the topic. 2. A bar chart, representing the frequency of Medical Subject Headings. 3. The frequencies of individual chemicals or medicines. 4. The frequencies of publication keywords. 5. The frequency of publications within the topic based on the year of publication. These metrics were proposed with the hypothesis that they could aid in comprehending the overarching subject of the topic. Additionally, they enable users to assess whether specific medications, concerning chemical frequencies, or medical terms, in the context of MeSH, are frequently mentioned in the literature associated with the topic. This approach could assist not only in categorizing topic-related publications but also in identifying potentially common terms that might not have been previously considered in their literature analysis. Likewise, examining the frequency of publications within a topic according to the year of publication can provide researchers with an initial assessment of the topic's developmental trajectory in published literature. This temporal trend analysis could be valuable in determining whether further investigation into the subject is warranted.

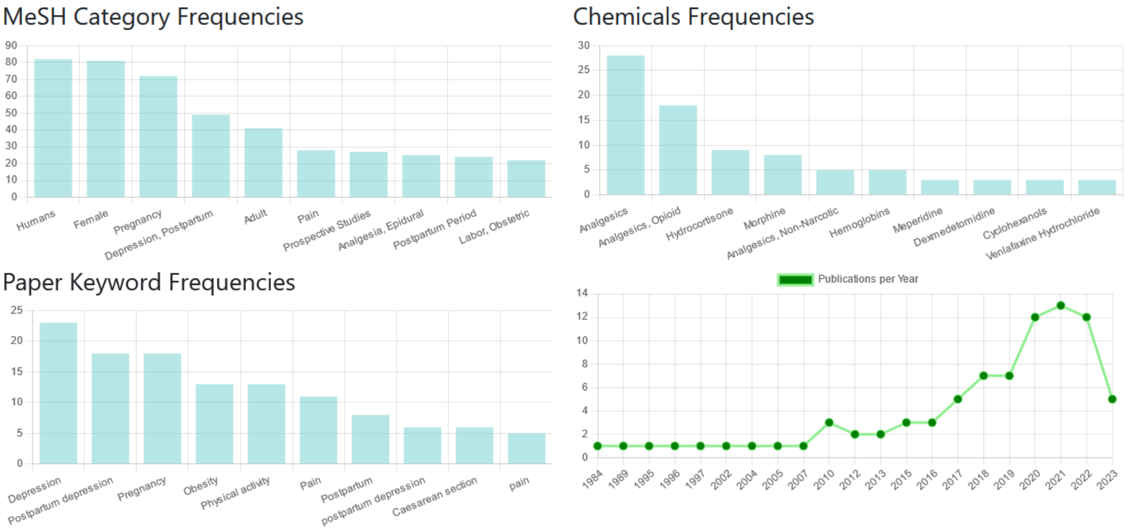
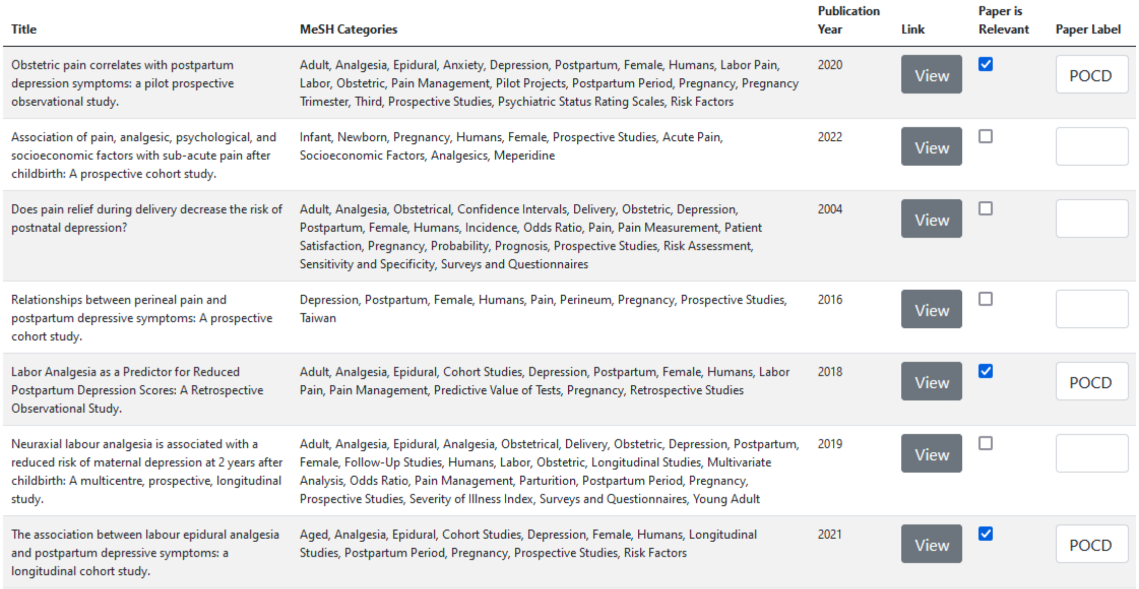


Figure 6.9: Feature graphs from the co-design process of the biomedical review framework, illustrated in the SMARTEN system.



Title	MeSH Categories	Publication Year	Link	Paper is Relevant	Paper Label
Obstetric pain correlates with postpartum depression symptoms: a pilot prospective observational study.	Adult, Analgesia, Epidural, Anxiety, Depression, Postpartum, Female, Humans, Labor Pain, Labor, Obstetric, Pain Management, Pilot Projects, Postpartum Period, Pregnancy, Pregnancy Trimester, Third, Prospective Studies, Psychiatric Status Rating Scales, Risk Factors	2020	View	☑	POCD
Association of pain, analgesic, psychological, and socioeconomic factors with sub-acute pain after childbirth: A prospective cohort study.	Infant, Newborn, Pregnancy, Humans, Female, Prospective Studies, Acute Pain, Socioeconomic Factors, Analgesics, Meperidine	2022	View	☐	
Does pain relief during delivery decrease the risk of postnatal depression?	Adult, Analgesia, Obstetrical, Confidence Intervals, Delivery, Obstetric, Depression, Postpartum, Female, Humans, Incidence, Odds Ratio, Pain, Pain Measurement, Patient Satisfaction, Pregnancy, Probability, Prognosis, Prospective Studies, Risk Assessment, Sensitivity and Specificity, Surveys and Questionnaires	2004	View	☐	
Relationships between perineal pain and postpartum depressive symptoms: A prospective cohort study.	Depression, Postpartum, Female, Humans, Pain, Perineum, Pregnancy, Prospective Studies, Taiwan	2016	View	☐	
Labor Analgesia as a Predictor for Reduced Postpartum Depression Scores: A Retrospective Observational Study.	Adult, Analgesia, Epidural, Cohort Studies, Depression, Postpartum, Female, Humans, Labor Pain, Pain Management, Predictive Value of Tests, Pregnancy, Retrospective Studies	2018	View	☑	POCD
Neuraxial labour analgesia is associated with a reduced risk of maternal depression at 2 years after childbirth: A multicentre, prospective, longitudinal study.	Adult, Analgesia, Epidural, Analgesia, Obstetrical, Delivery, Obstetric, Depression, Postpartum, Female, Follow-Up Studies, Humans, Labor, Obstetric, Longitudinal Studies, Multivariate Analysis, Odds Ratio, Pain Management, Parturition, Postpartum Period, Pregnancy, Prospective Studies, Severity of Illness Index, Surveys and Questionnaires, Young Adult	2019	View	☐	
The association between labour epidural analgesia and postpartum depressive symptoms: a longitudinal cohort study.	Aged, Analgesia, Epidural, Cohort Studies, Depression, Female, Humans, Longitudinal Studies, Postpartum Period, Pregnancy, Prospective Studies, Risk Factors	2021	View	☑	POCD

Figure 6.10: The table visualisation, displaying the 10 most relevant publications within a topic and the corresponding labelling form, in SMARTEN.

6.7.4 Limiting the Publication Search

Additionally, in the SMARTEN system, users can view the 10 most relevant publications that reside within a topic (as per requirements in Section 6.6.1), where the degree of relevance is determined based on the cosine similarity distance of the individual publication embedding, relative to the cluster centroid that determines the central embedding of the topic. In the original work proposing the Top2Vec algorithm, this is presented as a method to identify documents that are most relevant to a given topic [16].

6.7.5 Meta-information and Manual Labelling

Furthermore, following the requirements established during the co-design process (described in Section 6.6.1), as well as stage 3 of the framework (Section 6.7), as shown in Figure 6.10, users have the option to view the full-text version of the paper, as well as see the title and MeSH categories for individual publications. Based on the information provided, users can then choose to label individual publications as being relevant to their literature search and assign a label of their choosing to any relevant literature which is identified.

6.7.6 Semantic Similarity of Identified Topics

Finally, as defined in the requirements in section 6.6.1, and reflected in the framework stage 4 in section 6.7, to enhance the discovery of knowledge during literature search, the relevant literature identified by users can be subsequently analysed and visualised in terms of the semantic similarity of topics that have been labelled by users. Figure 6.5 presents an example of the general view of this visualisation in SMARTEN, while the expanded list of publications assigned to a topic "POCD" is displayed after its representing node is clicked. In this relationship graph of topics, nodes are colour-coded to represent the topics identified by users in red, the search terms from the first alias group in green and the second alias group in blue.

For this relationship graph, SMARTEN groups individual papers by the labels assigned by the user and then calculate an overall *topic vector* taken as an average of the individual embeddings for each paper in the topic group. All combinations of topic vectors are compared and a cosine similarity metric is used to calculate the distances of all combinations of topic vectors. The cosine similarity distance measure represents the distances between two topic vectors, with a higher similarity demonstrating that the two groups of documents are closely related. Once similarities have been calculated for all combinations of topic vectors, a *network graph* is created, by considering the top three highest-ranking connections for each topic. The search terms provided by the user when initiating the literature search are included as nodes on the graph, which are generated through the same embedding method by which the papers were embedded, such that they can be compared using cosine similarity. Once constructed, the graph demonstrates the degrees of similarity between topics and the user's search terms. Additionally, this graph is used in SMARTEN to allow users to view the papers they have assigned to each topic, by directly clicking on individual graph nodes.

This graph view was designed to demonstrate to the user how the topics that they have deemed relevant to their investigation relate to each other, and the search criteria that they adopted. This is intended to assist in the comprehension of topics that may be new knowledge to the user, and the effects of this are analysed as part of our study.

6.7.7 Experimental Design

For the evaluation of the Comprehensive Automatic Literature Review Framework for Biomedical Literature, we evaluate its software instantiation, SMARTEN. For this, it is important to ensure a suitable cross-section of the target user group, which is people with medical background. Thus, we conduct the software instantiation evaluation with a relatively even distribution of gender (10 female and 8 male participants). We report this distribution of participants for a number of reasons. Firstly, although significant disparities in results based on the reported gender of the participants are not anticipated - which was an optional response - it is important to demonstrate an effort to mitigate response bias while ensuring the acquisition of diverse perspectives. Moreover, documenting a relatively even gender distribution is intended to ensure a representative sample of participants, thereby enhancing the generalisability of the research findings. In terms of medical backgrounds, we have selected specialists at different levels in the medical profession, for a representative coverage (2 medical students, 5 junior resident doctors, 6 senior resident doctors, 4 staff doctors or consultant doctors, and 3 researchers in medicine). In terms of age distribution, we also opted for a wide, as representative as possible, range, thus our participants are aged between 18-50.

Experiment Briefing

Participants were briefed on an overview of the architecture of the framework, as well as the functionality of the software implementation and the goals the framework was aimed at addressing and ethical considerations. Details of the ethical procedure, for which consent was obtained and approved, are detailed in section 3.4.2. Following this briefing, participants received a login for the software, which would allow the tracking and storing of logs of user interactions, including their literature searches, topics identified within the literature found from their search, and the labels assigned by users to topics and publications. Participants were not given any objective or subject from which to base their literature search on, and were encouraged to use the software to investigate a domain of their choosing. Once the search terms are entered, the framework-inspired SMARTEN software would present participants with a summary of the topics found in their literature, generated through the topic modelling process. Users may then choose to analyse individual topics and provide labels to any publications identified within the wider literature search which they felt was relevant to them. Following this, participants could view the relationships of the topics they identified, along with their original search terminology, based on the semantic similarities between the topics and search terms, as presented in the general overview of the framework in Figure 6.7.

Participant Questionnaire

After completion of the literature analysis task using the SMARTEN system, participants were presented with a questionnaire designed to gauge their perceptions of the interaction. The questionnaire consisted of questions involving responses on a seven-point Likert scale (1-3: strongly to slightly disagree, 4: neutral, 5-7: slightly to strongly agree). The content of the questions was designed to address the second research question by evaluating participant perceptions of *cognitive load* [263], *tech-*

nology acceptance [271–274, 280], and *general satisfaction* [271, 275, 280] when using the SMARTEN system for literature analysis. The questions on cognitive load and technology acceptance are provided in Appendix B. The objective of this questionnaire based on cognitive load, technology acceptance, and general satisfaction is to address RQ2.6 in this study; *What are the effects on cognitive load, technology acceptance, and general acceptance of users when using the software implementation built on top of our framework?.* As a final stage of the questionnaire, participants were asked to provide a general, qualitative overview of their experience with respect to the positive and negative aspects of using the SMARTEN system. Participants were not required to provide feedback on both positive and negative aspects of the SMARTEN system, in order to avoid obligating them to give responses in either category if they did not perceive any. Feedback on both positive and negative aspects of the system was optional.

Assessment of Relevance of Identified Publications

Further to the responses recorded in the questionnaire, we conducted an analysis of the labels assigned by users for the literature identified by the system and compared this with the search terminology used for user queries. This allows for the calculation of the average accuracy of the framework in the identification of relevant publications and addresses RQ2.7. We calculate the percentage of papers that the user has assigned a “*relevant*” label, relative to the total number of papers suggested to the user. As the model may suggest a large number of topics to the user, we only count topics that have been assigned at least one “*relevant*” label by users, as we deduce that any topics which have not been accessed or labelled by the user are not relevant to their investigation. We present the findings of this analysis in Section 6.8.2.

The collection of user logs for this study was carried out in a fully anonymised manner, with only the original login being used to store the search terms applied by each user and the associated topics identified within the literature. Using the identified topics in the literature and the labels assigned by users, we can investigate RQ2.7. This involves quantifying the number of publications labelled as relevant by users among those recommended by the framework.

6.8 Results

To evaluate the Comprehensive Automatic Literature Review Framework for Biomedical Literature, the results of the experiment run on its instantiation, the SMARTEN system, are presented. To respond to our research questions, we perform analysis of experts' perceptions (including cognitive load, technology acceptance, and general satisfaction, etc.), of labelling, and of the experts' qualitative responses.

6.8.1 User perception analysis

Table 6.1: Effects upon cognitive load of the literature analysis activity in SMARTEN.

	Mental Effort		Mental Load	
	Purpose	Labelling	Distraction	Stress
Mean	3.50	4.56	3.72	3.61
S.D.	1.64	1.86	1.63	1.77

Table 6.1 displays the results of the evaluation of the cognitive load [263] of the 18 subjects while participating in the literature analysis task. The evaluation of cognitive load was carried out using a questionnaire developed from previous studies on cognitive load measurement [263–265, 269–271], as elaborated in Sections 6.3.2 and 6.7.7. In terms of *"mental effort"*, the average rating for *"effort required for understanding the purpose of the activity"*, is 3.5 ($S.D. = 1.64$). In comparison with other studies adopting a similar evaluation of cognitive load using a questionnaire [270, 271], the reported values are higher than the comparative studies, which report 2.90 or 2.97, and 3.02, respectively, for both studies. With regard to this specific measurement of the cognitive load questionnaire, which investigates how participants reported on their understanding of purpose of using the system (Appendix B, Q1), this indicates that most of the participants felt that they could understand the purpose of the activity, with a moderate effort. However, some participants found difficulties understanding the purpose of the literature analysis task. This could be for a number of reasons, such as a lack of sufficient information being provided prior to participation in the experiment, the user interface design, or the information presented to users on-screen. This is discussed in further detail in Section 6.8.3, after

analysing the qualitative feedback reported by participants.The mean score for the *"effort required for labelling the papers"* is 4.56 (*S.D.* = 1.86), indicating that the task of analysing the literature was somewhat challenging for the participants. With regards to this measure, which is not directly comparable to previous studies due the specific nature of the question in relation to the labelling task (Appendix B, Q2), it is worth noting that participants were required to read *any* publications that were assigned to topics, before assigning a label. This could induce a degree of cognitive load, due to the requirements of comprehending the publication content. In terms of *"mental load"*, the average ratings of the degree of distraction and stress while using the framework were both slightly less than 4; this implies that the participants felt only a moderate stress during the literature analysis activities, and most of them felt that the framework did not distract them from concentrating on the literature analysis.

Table 6.2: Technology acceptance of SMARTEN.

	Feature Graphs		Publication List Labelling		Relations Map	
	Easiness	Usefullness	Easiness	Usefullness	Easiness	Usefullness
Mean	5.83	6.17	5.89	6.11	5.78	6.39
S.D.	0.90	0.90	1.15	0.87	1.31	0.76

In terms of *"technology acceptance"* measures, as shown in Table 6.2, the average rating on the *"perceived ease of use"* item was 5.83 (*S.D.* = 0.90) when analysing the feature graphs as shown in Figure 6.9; 5.89 (*S.D.* = 1.15) when labelling related publications, as shown in Figure 6.10; and 5.78 (*S.D.* = 1.31), when checking the relation maps shown in Figure 6.5. This indicates that the majority of participants somewhat agreed that these three functions were straightforward to use and get accustomed to. The average ratings of *"perceived usefulness"* of these three functions are all greater than 6, which implies that most participants agreed that these three functions were useful for improving their literature analysis.

When asked about the feature graphs (*MeSH Category Frequencies*, *Chemicals Frequencies*, *Paper Keyword Frequencies*, and *Publication per year*) that may not have been beneficial for their analysis and could be excluded from the framework (where participant experts could select from multiple choice options), half of the participants opted for none, 7 participants selected MeSH Category Frequencies, and 4 participants opted for Chemicals Frequencies. This suggests that the MeSH Category Frequencies and Chemicals Frequencies may be less valuable, compared to the other categories. It is noteworthy that despite MeSH categories being standardised in PubMed[5], with which participants are presumably familiar, these categories were perceived as the least useful in their analysis. Conversely, conventional publication keywords were deemed more beneficial. It is worth considering whether this is attributable to the abstract character of certain high-frequency MeSH terms (e.g., '*Human*' emerges as one of the terms with the highest frequency in many topics), and whether the use of publication keywords, which are determined by authors with a more considerable degree of discretion, provides a more valuable understanding of topics to participants.

	Question	**Mean**	**S.D.**
Q1	The system helped me to have a deeper understanding of these topics.	6.17	0.69
Q2	Using the system can help me find new information within the literature.	6.39	0.49
Q3	Overall, the system was easy to use and navigate.	5.72	1.33
Q4	I like to use the system.	5.61	1.06
Q5	Overall, the right amount of information was given by the system.	5.56	1.30
Q6	I felt confident using the system to improve my understanding of new medical subjects.	5.94	0.52
Q7	Using this system would enhance the effectiveness of my literature analysis.	6.06	0.62
Q8	I would be able to use the system without assistance from others.	6.06	0.85
Q9	I would like to use the system in my future work.	6.33	0.75
Q10	I would recommend this system to my colleagues.	6.11	0.66

Table 6.3: General user satisfaction in SMARTEN.

Table 6.3 presents the average ratings for the 10 items in "General satisfaction" questionnaire. In summary, all ratings were above 5.5, with particular emphasis on Q1, Q2, and Q7-Q10, which exceeded 6. This suggests that the majority of respondents were content with the provided functionality. Notably, the responses to Q1 and Q2 indicate a high degree of satisfaction, reflecting the users' views on the system's effectiveness in aiding their literature analysis. The system was found to enable participants to gain a more profound insight into the topics, receiving an average rating of 6.17 with a standard deviation of 0.69. Additionally, users responded that utilising the system could help them discover new information in the literature, giving it a mean score of 6.39, with a very low standard deviation of

[5] `https://www.ncbi.nlm.nih.gov/mesh/`

0.49. The small degree of standard deviation for both questions suggests a consistent trend in responses. Regarding Q3 and Q4, a mean response of 5.61 ($SD = 1.06$) and 5.56 ($SD = 1.30$), respectively, indicates that,generally, users could navigate and use the software easily, and liked using the system. Overall, the general satisfaction questionnaire results suggest a high level of user satisfaction and a perception that the system can enhance the effectiveness of literature analysis and support learning in medical subjects.

6.8.2 Analysis of User Labelling

The participants were encouraged to label 1 or 2 groups of the topics that they considered to be most relevant to the associated key phrases during their literature search. Table 6.4 presents the percentage of papers that were labelled relevant by the participants, out of the total number of papers (10 per topic) suggested to them, after topic modelling with Top2Vec. It is important to note that the inclusion of a secondary keyword group is not mandatory, resulting in some users opting to exclude this when their research interest is confined to a specific research focus.

The lowest labelling accuracy was 15%, which suggested that at least 1 out of the 10 papers within a topic was relevant to the participant's search. Moreover, there were 9 groups of search terms (IDs 3,5,10, 11, 13-17) for which more than 50% of the papers suggested by SMARTEN were highly relevant to the search terms. This suggests that half of the participants believed more than half of the suggested papers were within their research interests.

6.8.3 Qualitative Responses from Participants

In addition to the quantitative analysis of user perceptions, qualitative feedback regarding the user experiences with the SMARTEN system, were also obtained, and presented in Tables 6.5, 6.6, and 6.7 involving *positive aspects*, *aspects that could be improved*, and *general suggestions*, respectively.

From a *positive perspective* of the SMARTEN system, most of the participants reported encountering serendipitous related publications while conducting their literature search task, especially one highlighting the system's efficacy in facilitating the discovery of new knowledge (see Table 6.5, Response 1). Additionally, four participants found the relationship mapping of semantic similarity (detailed in section 6.7.6) beneficial (see Table 6.5: responses 2, 4, 5 and 6). Furthermore, one participant expressed satisfaction with the publication year graphing feature, which was introduced during the co-design process as described in section 6.6.1 (see Table 6.5, Response 3). One participant mentioned that the labelling and graphs helped them focus their literature searches(see Table 6.5: responses 6); this response encompasses not only the relationship mappings graph (Figure 6.5) but also the topic feature graphs (Figure 6.9). As for response 7, the expert appreciated the various levels of visualisation that are available in the SMARTEN system; it is to be noted that this particular expert did not focus on one specific visualisation, but mentioned it as a generic positive feature.

With regards to *the aspects that could be improved* in the SMARTEN system and the *general feedback*, participants identified several areas, mainly concerned with language version, topic labelling, semantic relationship mapping and update notification. Language support functions were proposed by four participants to overcome the language barrier. Regarding topic labelling, five participants stated that they found some difficulties with using the labelling system (responses 5, 6, 8 and 9 in Table 6.6 and response 1 in Table 6.7). This was reported as being caused by the graphical interface when labelling publications, particularly due to the input text box (as seen in Figure 6.10) being too small when users input a long list of keywords. Some participants mentioned in their feedback to provide only the MeSH category keywords relevant to their search terms in order to create more space for this input text box. Additionally, one participant suggested that functionality to generate labelling recommendations automatically would be beneficial. One participant stated that they found difficulties with using the relationship mapping functionality (detailed in section 6.7.6, response 2 in Table 6.6), specifically due to them finding the moving of the graph inconvenient. One participant stated that they would like to have options to increase the number of keyword groups, instead of the current limit of 2 keyword groups during the experiment setting (Table 6.6, response 9).

Moreover, a participant suggested an additional feature of receiving email notifications for newly published papers in their specified area of interest (Table 6.7, response 3). Lastly, a participant expressed overall satisfaction with the SMARTEN implementation (Table 6.7, response 4) and expected its updated version.

6.9 Discussion and Responses to the Research Questions

In addressing **RQ2.5** regarding the SMARTEN framework, we have investigated the application of the Top2Vec [16] algorithm, for the exploratory analysis of medical literature. This algorithmic approach presents advantages, compared to existing techniques for the analysis of the literature using probabilistic models such as LDA [156, 252], as our *Comprehensive Automatic Biomedical Literature Review Framework* defines methods for the automatic detection of the optimal number of topics that can be extracted from the literature. Furthermore, it allows the leveraging of state-of-the-art embedding models, which can be easily changed or updated. This makes it potentially effective for the use case of assisting medical experts in their literature analysis, as it can remove many of the knowledge barriers that can limit analysis, as defined in [156]. Through a co-design process engaged with domain experts, we have expanded upon prior research, by creating a Comprehensive Automatic Biomedical Literature Review Framework, with features such as functionality for the accounting for aliases and synonyms during literature search, the grouping of topics within the extracted literature, and the presentation of meta-information extracted from publications for identified topics. The resulting framework permits the analysis of literature based on user-defined queries, which, whilst applied to a given domain, may be extended to others, thus defining an end-to-end solution for literature analysis. The user responses to the questions of technology acceptance and general satisfaction after using the proposed framework demonstrate that the proposed methodology, through the use of the Top2Vec algorithm for literature analysis, and subsequent extension through visualisation, has had a positive effect on the discovery of new knowledge for medical experts. Especially, when considering the positive aspects of the system provided by participants in their qualitative evaluation, 3 participants highlighted that the visualisation of the relationship map (Figure 6.5) was very useful for their study, which is consistent with the result presented in Table 6.2.

To answer **RQ2.6** regarding the perception of users from a medical background when interacting with the proposed framework, first, the result of the cognitive load analysis shown in Table 6.1 suggests that the literature analysis activities during the experiment brought a moderate mental load to the users.

Secondly, the result of the technology acceptance analysis, as shown in Table 6.2, indicates that most participants could easily grasp the operation of the framework and acknowledge the usefulness of SMARTEN. Finally, the analysis of the general satisfaction, as shown in Table 6.3, reveals that most of the participants agreed that the framework helped them have a deeper understanding of the research topics, find new information in the literature, and would like to use the framework in their future work and recommend the SMARTEN system to their colleagues.

Additionally, some participants commented that an interface that presents their first language would be preferred when using SMARTEN. Furthermore, some participants highlighted areas of improvement concerning the interface of the SMARTEN system. Other suggestions included the automatic recommendation of labels, to alleviate the workload of the user during the labelling stage and a notification functionality for finding out when new research was published related to their search.

To answer **RQ2.7** regarding the accuracy of applying Top2Vec as a method for discovery of literature topics in the SMARTEN framework, the analysis of the labels provided by the users indicates that an average of 52.3% of the suggested publications were relevant to the user's analysis goal, as shown in Table 6.4. This varies considerably between users, some users finding 100% of papers suggested to be relevant, compared to as low as 15%.

It is worthy of note that as a method of identifying relevant papers, an average of 52.3% accuracy indicates that on average, 5 out of 10 suggested papers were deemed as relevant by the participants in the topics they chose to analyse. This indicates that the SMARTEN system is beneficial in helping to discern valuable information from the wealth of literature which was identified during the literature searches. When taking into account the lowest labelling accuracy of 15%, this still indicates at the lowest as 1 out of the 10 papers as being relevant to the participant's search. Thus, by allowing users to filter topics within the literature based on the topic-words identified by the topic modelling algorithm, and then presenting users with the 10 most relevant publications to that topic, we argue that the SMARTEN framework and software implementation allow the efficient consolidation and filtering of literature in order to assist in knowledge discovery from a large volume of publications that could not have been manually analysed.

Still, instances of low labelling accuracy indicate room for improvement, which likely stems from several potential reasons. Initially, since participants initiated their analyses using queries of their own selection, it is important to consider the extent to which the literature obtained from PubMed influences labelling accuracy. For example, participants may have formulated search queries that were either excessively vague or, conversely, overly precise, consequently restricting the corpus of literature retrieved via the API query. Such limitations could adversely impact subsequent topic modelling processes. As the determination of query relevance is performed by PubMed, our ability to ascertain the criteria for relevance selection remains restricted; thus, this aspect could benefit from improved post-processing of retrieval results to either filter unrelated literature or identify additional pertinent literature. Moreover, optimising the configuration of Top2Vec with respect to hyper-parameters for UMAP in dimensionality reduction and HDBSCAN in clustering could be investigated to evaluate whether such adjustments enhance the quality of clustering and, by extension, the discernment of topics. Finally, it should be acknowledged that each labelling instance is distinctly associated with the individual participants. Consequently, the variation observed in the reported accuracy metrics of labelling may be attributed to the specific knowledge-base of each participant, which is further shaped by their diverse medical backgrounds as documented in Section 6.7.7. Furthermore, the amount of time participants devote to examining the recommended results, as well as the level of attentiveness maintained during the analysis, may influence their capacity to assess relevance.

Generally, we argue that an average of 52.3% of relevant articles suggested on relevant topics is beneficial for the discovery of new knowledge and contributes to assisting medical experts in finding new information. This is supported by the results of the technology acceptance questionnaire (Table 6.2), which suggests most participants believed the main system functions were useful for improving their literature analysis and the result of the general satisfaction questionnaire (Table 6.3) which suggests most participants believed the system enhances the effectiveness of literature analysis and supports learning in medical subjects.

6.10 Epilogue

The research presented in this chapter has presented a continued expansion of the research provided in Chapter 5, facilitating the design of a fully automated framework capable of assisting in the literature review process, based on a topic modelling approach. The main contribution of this avenue of research can be summarised as facilitating the design of a framework which has been demonstrated through evaluation by medical experts to assist in their literature analysis. Outside of this main contribution, it was possible to explore how the triangulation of external data sources might assist in literature analysis for niche domains such as human protein kinases. The focus of this chapter upon literature analysis presents value to the publishing industry due to the potential adaptation of this approach, which may theoretically be applied to any form of textual media, such as news and magazine publications or social media streams in order to effectively identify topics within that specific domain. In the following chapter, a thorough investigation is carried out on the particular task of dimensionality reduction, which is crucial for cluster-based topic modeling. The aim is to suggest modifications to the algorithms employed in the literature analysis framework to enhance the quality of the topics generated.

Table 6.4: Search terms applied by the users in SMARTEN, and the percentage of papers they labelled as relevant out of the papers Suggested by the system after topic modelling.

	Keyword Group 1	**Keyword Group 2**	**% of Relevant Papers**
1.	Maternal pain, Pain during pregnancy	Postpartum depression, Postnatal depression	16.6%
2.	POCD, Postoperative cognitive disorder, Postoperative cognitive decline	/	40%
3.	Implanted cardio defibrillators	Deactivation	65%
4.	Hypervolemia	Cardiac Surgery	15%
5.	Urinary catheter discomfort	Postoperative	65%
6.	Sevoflurane	Ondansetron	35%
7.	Acute lymphoblastic leukemia	Pediatric	33.3%
8.	Microbiome	Brain-Gut	43.3%
9.	Ammonia	Urinary tract infection	25%
10.	Transcranial magnetic stimulation	Pain	56.6%
11.	Dissection	Bovine	80%
12.	Propofol	Remimazolam	40%
13.	Liver Failure	Acidosis	56.6%
14.	PROM, Preterm premature rupture of the membranes	Antibiotics	50%
15.	Ondansetron	Metoclopramide	100%
16.	Vital Sign	Bleeding	90%
17.	Pancreatoduadenectomy	Fluid	100%
18.	Heart, Coronary	Coronavirus, COVID	30%
	Average		52.3%

Table 6.5: Qualitative responses by participants outlining positive aspects of the SMARTEN system.

	Responses to Positive Aspects of SMARTEN
1.	I came across unexpected related words and expanded my knowledge. I could broad my understanding in related key words.
2.	I was able to learn more about the strength of relationships between surrounding topics and relationships that are of interest to other researchers.
3.	Knowing the publication year of the paper
4.	Map is very good
5.	It was nice that the map was easy to see.
6.	You can narrow down the field you want to search using labeling and graphs.
7.	Where it can be visualized.

Table 6.6: Qualitative responses by participants outlining negative aspects of the SMARTEN system.

	Responses to Negative Aspects of SMARTEN
1.	English.
2.	When moving a relationship graph three-dimensionally, it was inconvenient that you could not move it unless you held a circle. I would like to be able to set the center of the entire graph and move the ball based on that. I want it to be able to rotate no matter where I hold it.
4.	If I type a keyword in Japanese, I want it to be automatically converted to English.
5.	I can't see the label, TOPIC is busy
6.	I want the label to be optional.
7.	labelling is necessary?
8.	Difficult to input labels
9.	Labelling is a hassle, and I want about 3 keyword searches.

Table 6.7: Qualitative responses by participants outlining general suggestions to improve the SMARTEN system.

	Responses to General Suggestions after using SMARTEN
1.	Labelling is a hassle.
2.	I would like to see a Japanese version implemented.
3.	I would like to receive a reminder email when a new paper is published in an area that I have searched for.
4.	I think it has great potential as a search tool. I hope it will be updated.

CHAPTER 7

Investigating the Effects of Dimensionality Reduction upon Clustering and Topic Models

This Chapter focuses on addressing GRQ3 through investigating methods to enhance dimensionality reduction, a key process for cluster-based topic modelling, by proposing architectural designs for parametric dimensionality reduction. The contents of this chapter are accepted into the IEEE Access Journal, under the title of "Partially-Supervised Metric Learning via Dimensionality Reduction of Text Embeddings using Transformer Encoders and Attention Mechanisms[1]", and under review for the 10th International Conference on Machine Learning, Optimization, and Data Science (LOD), under the title of "Enhancing Cluster-Based Topic Models through Parametric Dimensionality Reduction with Transformer Encoders".

7.1 Prologue

In Chapters 5 and 6, when investigating the area of topic modelling of academic literature, the Top2Vec algorithm [6] was essential for identifying topics within the

[1] https://ieeexplore.ieee.org/abstract/document/10536728

literature. This algorithm approaches topic modelling from a clustering perspective, with the details of this process being presented in Chapter 3.2.3. An essential stage in the algorithmic process of Top2Vec is the integration of dimensionality reduction, which permits the reduction of highly dimensional embeddings, which are vector representations of textual information, to a low-dimensionality representation. In Top2Vec, the UMAP algorithm [14] is used to transform highly dimensional document embeddings into a low-dimensional representation which can then be clustered by the HDBSCAN algorithm [79] to identify topics.

For this chapter, an investigation of the UMAP algorithm is performed with the aim of exploring a novel area of research made possible through the more recently introduced parametric UMAP [4]algorithm. The objective of this chapter is initially to establish a robust evaluation of the effects of existing methods for dimensionality reduction on the downstream clustering of embeddings. Subsequently, the chapter proposes a novel approach to dimensionality reduction through the integration of a transformer-encoder-based parametric dimensionality reduction pipeline. Following an examination of the implications of this proposed pipeline in the context of metric learning, it is incorporated into a modified version Top2Vec and subjected to both quantitative and qualitative evaluations to assess the impact on the resultant topics. The notable findings produced in addressing these objectives are three-fold. Firstly, an empirical analysis of UMAP and parametric UMAP across a range of output dimensionalities demonstrates that the algorithm is effective in improving downstream clustering of the k-Means algorithm, achieving notable results even at a minimal output dimensionality of 3. Furthermore, this comparison indicates that elevating the output dimensionality past this value provides minimal benefit to the clustering quality. Secondly, the integration of the transformer-encoder-based parametric dimensionality reduction pipeline is empirically shown to significantly enhance the clustering accuracy of the k-Means algorithm when applied within a metric-learning framework. Concerning these contributions, it is crucial to emphasise that *curse of dimensionality* is prevalent in both k-Means and HDBSCAN, which serves as the core clustering algorithm in Top2Vec. Both algorithms commonly employ Euclidean distance for calculating the distance between points during clustering [72–76, 79] and are thus vulnerable to the notion of distances becoming meaningless in high-dimensional spaces, as discussed in Section 2.1.4. Therefore, by ascertaining the implications of a comprehensive dimensionality reduction study through comparative analysis utilising accuracy measures on benchmark datasets, it is anticipated that the findings will be directly applicable to cluster-based topic modelling. Finally, the influence of the transformer-encoder on enhancing topic modelling, when employed in conjunction with Top2Vec, is also observed, with these enhancements becoming more pronounced when novel additional residual connections are incorporated into the network architecture.

To begin with, an introduction to the parametric UMAP algorithm is first provided, as well as the key issues for why dimensionality reduction is necessary for clustering. Following this, an investigation is performed into the novel paradigm of architectural network design for parametric UMAP, where the transformer-encoder architecture, which has been demonstrated to be revolutionary in a wide number of NLP domains, is proposed as an architecture which can be applied to the parametric UMAP DR pipeline. This is explored with an initial focus on a metric-learning methodology, using partially labelled data, work which generated a paper accepted for publication in *IEEE Access journal*[2], further detailed in Section 7.4. After identifying the notable influence that architectural design in parametric dimensionality reduction appears to incur upon downstream clustering, a continued extension of the findings investigates the introduction of additional residual connections into the transformer-encoder architecture and evaluates the influence that this has upon the quality of the topics produced by a modified version of Top2Vec, where the proposed architecture is introduced in a parametric dimensionality reduction pipeline. This is essential to present the avenue of research in a domain that is the core focus of the thesis, and indicates the considerable influence that the introduction of residual connections incurs upon the quality of topics produced. This is presented in Section 7.6 and is currently being reviewed for publication at the *10th Annual Conference on Machine Learning, Optimization and Data science (LOD)*[3].

[2] `https://ieeexplore.ieee.org/abstract/document/10536728`

[3] `https://link.springer.com/conference/mod`

7.2 Research Questions

As introduced in Section 1.5, the third general research question is investigated in this chapter:

RQ3: How can dimensionality reduction be used to improve accuracy of text clustering, and, subsequently, topic modelling? This can be broken down into the following sub-questions:

- **RQ3.1** *How do current dimensionality reduction algorithms affect accuracy in text clustering tasks?*
- **RQ3.2** *How does output dimensionality of dimensionality reduction models influence performance in text clustering tasks?*
- **RQ3.3** *Can small portions of labelled data used in the metric learning of dimensionality reduction contribute to improvements in downstream clustering accuracy?*
- **RQ3.4** *Can the introduction of attention mechanisms within neural networks (further) improve the metric learning of dimensionality reduction algorithms, in terms of clustering accuracy?*
- **RQ3.5** *Can the transformer-encoder neural network positively influence downstream clustering when applied to both supervised, and unsupervised learning methodologies in dimensionality reduction?*
- **RQ3.6** *What are the implications of introducing additional residual connections into the transformer-encoder pipeline for dimensionality reduction, when dimensionality reduction is used for cluster-based topic modelling?*

7.2.1 Research Objectives

Based on the above motivation and research questions, the research objectives for this Chapter are:

- **RO3.1:** The preliminary evaluation of existing dimensionality reduction algorithms through analysing their effects upon downstream clustering of text embeddings. (Addressing RQ3.1).
- **RO3.2:** The evaluation of attention mechanisms as an architectural component in neural networks within a parametric dimensionality pipeline, through analysing their effects upon downstream clustering of text embeddings. (Addressing RQ3.4).
- **RO3.3:** The provision of a transformer-encoder based parametric dimensionality reduction pipeline, and evaluation of the downstream performance of cluster-based topic modelling algorithms. (Addressing RQ3.5).
- **RO3.4:** The adaptation of the transformer-encoder, by introducing additional residual connections and subsequent evaluation of this with comparison to existing dimensionality reduction algorithms, and evaluation of the effects of this in enhancing cluster-based topic modelling. (Addressing RQ3.6).

7.3 Chapter Contributions

The main contributions of this chapter are as follows. (1) To demonstrate, for the first time, to the best of our knowledge, the effectiveness of the transformer-encoder as an architecture in the metric learning of lower dimensionality embeddings with parametric UMAP for text clustering. We demonstrate it achieves the highest accuracy across three of the four datasets investigated, with no loss in accuracy when our proposed transformer-encoder is compared with the current SoA, UMAP, on the fourth dataset. We showcase this through both a visual analysis of the clustering solution on two datasets, and an evaluation of clustering accuracy based on four datasets. (2) We present the outcomes of the first empirical study into the outcomes of DR, and metric learning. That is, we contribute with an empirical evaluation of all variants of UMAP, as well as of the traditional techniques Principle Components Analysis, Linear Discriminant Analysis, and *t*-SNE, by comparing performance when applied to text-clustering tasks across a range of dimensionalities. (3) We demonstrate, for the first time, the effectiveness of applying attention mechanisms within architectures in the parametric UMAP pipeline, when combined with metric-learning. (4) The proposal of an extension to the transformer-encoder pipeline for parametric UMAP, by introducing additional residual connections into the network in order to mitigate local minima and vanishing gradient problems when modelling long sequences, within a topic modelling framework, and demonstration of the effects this has upon the topic modelling solution, from both a metric, and quality perspective.

7.4 Related Work

7.4.1 Clustering Algorithms and the Curse of Dimensionality

Embeddings produced by techniques such as transformer networks [5] ensure the provision of large amounts of information encoded in highly dimensional vectors, which can be measured using the GLUE benchmark [40, 283]. This has contributed to significant improvements across a range of natural language processing (NLP) tasks, including question answering [40], sentiment classification [34], and text clustering [284]. However, the high degree of dimensionality in embeddings produced by these methods can present difficulties in downstream analysis tasks, due to the prevalence of the "curse of dimensionality" (discussed in Section 2.1.4) [68], which introduces a multitude of problems, as dimensionality increases. Firstly, the volume of computational memory required for storage and processing of high-dimensional vectors is large, and grows exponentially. Second, a greater computational complexity is observed in algorithms, as the number of dimensions increases [68]. Third, the distance measurements necessary for determining distances between embeddings tend to become meaningless in high-dimensional spaces, where the ratio between nearest and farthest points approaches one, such that the points essentially become equidistant from each other [69–71].

In clustering tasks, of which cluster-based topic modelling is one, this is best represented by distance measures used in clustering algorithms, where the distance measure becomes meaningless as the dimensionality increases [69, 70]. Empirical investigations have shown that this phenomenon appears for dimensionalities greater than 10 [69]. The reason being that, as dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point. This presents difficulties in any downstream tasks that apply nearest-neighbour searches, such as k-Means clustering, or systems that use distance measures, such as cosine similarity in nearest-neighbour searches. In the Top2Vec algorithm [16], HDBSCAN clustering [79] is performed for the identification of dense clusters of documents that represent topics.

7.4.2 Dimensionality Reduction Algorithms

To address these limitations, the technique of dimensionality reduction may be applied, which seeks to represent the global and local structure of highly dimensional data in a smaller feature space [70]. Dimensionality reduction may be defined as the transformation of high-dimensional data into a meaningful representation with reduced dimensionality [83]. As noted, this is necessary in many domains, where highly dimensional data can negatively impact computing efficiency and accuracy. In clustering tasks, this problem is best represented through the effect on distance measures, such as k-Means, where it becomes clear that the distance measure becomes meaningless, as the dimensionality increases [69, 70].

Among the techniques proposed for dimensionality reduction, the traditional ones include linear techniques, such as Principal Component Analysis (PCA) [84]. Another technique, Linear Discriminant Analysis (LiDA hereafter, to avoid confusion with Latent Dirichlet Allocation), provided a supervised approach to dimensionality reduction, through a generalisation of Fischer's linear discriminant [85], seeking to identify linear combinations of features, as a means to characterise or separate objects, or documents. However, this technique fails to perform suitably when applied to complex, non-linear data.

More recently, t-distributed stochastic neighbour embedding [86] was proposed as a nonlinear means of dimensionality reduction for the purpose of visualisation.

t-SNE was based upon Stochastic Neighbour Embedding [87], wherein a Gaussian is centred over high-dimensional objects, ensuring that a probability distribution may be defined over potential neighbours of the object. t-SNE expanded upon this, through the implementation of a *Student-t* distribution in place of a Gaussian, when computing the similarity between points in low-dimensional space. In this method, the reduction was typically performed to a dimensionality of 2 or 3, with the resulting vectors being applied as coordinate points in visualisation. t-SNE observed a significant decrease in performance as dimensionality increased [88].

From the perspective of parametric DR, an extension to t-SNE [92] works upon the assumption that a neural network possessing sufficient hidden layers is capable of achieving an approximation of the non-linear functions employed by t-SNE, when mapping a high-dimensional representation to a lower-dimensional representation. In this work, the authors discuss that directly training a neural network through backpropogation is not feasible, due to the tendency for backpropogation to encounter a local minimum, given the complex interactions between layers in the network, which entail a large number of parameters. To address this, the authors applied a training strategy involving the training of autoencoders based upon Restricted Boltzmann Machines (RBMS). In this process, a stack of RBMs is trained, and then used to generate a pre-trained feedforward network, which can subsequently be fine-tuned using backpropogation. The resulting network represents an approximation of the functions of t-SNE. The work demonstrates through experimentation that the parametric model can outperform PCA and an autoencoder in the dimensionality reduction of the MNIST [93], and 20 Newsgroups datasets [94]. In this thesis, the problem of local minima during parametric DR training is considered, with the aim of addressing this through the introduction of additional residual connections, as detailed in Section 7.6.

Another experiment in dimensionality reduction [95] focused on the application of the UMAP algorithm to improve clustering performance in image classification tasks; as representing the current state of the art, albeit in a different field, this approach is introduced in Sections 2.1.6 and 2.1.6, with the technical workings of the algorithms discussed in Sections 3.2.4 and 3.2.5. The authors applied their experiment on four clustering algorithms; k-Means [285], HDBSCAN [79, 166], Gaussian Mixture Models [286] and Agglomerative Clustering [287]. Results indicated a significant improvement in accuracy across multiple datasets, achieving an improvement of 60% when applied to HDBSCAN on the United States Postal Service [288] digit classification dataset. However, there was no reporting of parameter configuration for both UMAP and the clustering algorithms applied. Most importantly, the dimensionality selected for the experiments was not disclosed. This presents a necessity for the disclosure of information for future researchers and forms the basis of our initial experiments in dimensionality reduction, which are detailed in Section 7.5.3.

As discussed, dimensionality reduction algorithms have been shown to be an effective preprocessing tool, which can contribute to downstream clustering [16,95–97]. Given their promise, we seek to investigate whether a novel pipeline based on the cross-domain implementation of neural network architectures, facilitated by the parametric UMAP framework, would lead to improvements. Within the literature, we have not identified any evidence of the transformer encoder architecture or experimentation with any specific neural network architectures within the parametric UMAP framework.

Most notably, the introduction of a neural network in learning low-dimensionality representations in parametric UMAP presents the opportunity for the specification of tailored neural network architectures. These may be tailored to specific domains and will form the main basis of this chapter. The robust mathematical foundation of UMAP allows the extension of the algorithm for use in supervised learning, which is briefly discussed in UMAP [14] and extended by Parametric UMAP [4]. In the case of Parametric UMAP, the introduction of labelled, or partially labelled, data allows training of the network using both classifier loss for labelled data, or UMAP loss for unlabelled data. Through the use of partially labelled data, semi-supervised learning allows the joint learning of data structure with unlabelled data, with labelled data being used for the optimisation of a supervised objective function. For the work presentedin Section 7.4, a partially supervised methodology is applied in conjunction with two proposed neural network architectures, to evaluate how this network architecture influences outcomes in terms of clustering accuracy.

7.4.3 Attention and Recurrence in Neural Networks

The term *Recurrent Neural Network* (RNN) refers to an artificial neural network, wherein artificial neurones send feedback signals to each other [289]. This subsequently allows the output of some nodes to influence the input of the same node. *Long Short-Term Memory* (LSTM) networks [290] have provided significant contributions to several domains, including speech recognition [291], handwriting recognition [292] and machine translation [292]. In speech recognition tasks, bidirectional RNNs [293, 294] have been applied, where both a forward and a backward RNN is present, with each respective RNN reading the input sequence in opposite directions.

As such, LSTM networks have demonstrated success in sequence-to-sequence learning tasks, and it is this which we would seek to evaluate as part of this study, as we can consider the task of a neural network in parametric UMAP to be able to be described as the modelling of a shorter sequence based upon a longer input.

Subsequent improvements to RNN architectures in sequence-to-sequence modelling tasks have involved the inclusion of *attention mechanisms* [295,296] in encoder-decoder networks. Attention mechanisms perform the computation of a context vector, representing the relationship between the layer output and inputs, where the context vector is a weighted sum of the hidden states of the network at each timestep. This guides a model to focus on specific components of the input sequence, rather than the whole vector sequence. In language modelling tasks, this allows a model to focus upon specific words within a sentence or speech, which may provide the most contextual information.

Recently, the use of attention mechanisms was further developed [5], by proposing the concept of *transformer network.* This architecture is based solely on attention mechanisms, with no convolutional or recurrent layers, instead leveraging the proposed "Scaled Dot-Product Attention", and Multi-Head Attention, to achieve improved performance in machine translation tasks [5]. A detailed explanation of the transformer is provided in Section 3.2.6. The transformer architecture has subsequently contributed to improvements in benchmark performance in a number of NLP tasks, including question answering, sentence continuation, named entity recognition, and language understanding [40] [31]. This has been advanced through the introduction of the pre-trained transformer in works such as BERT [40], where a transformer model is trained upon a large corpus, through the omission of certain words and prediction of the correct word.

Of particular value to our investigation is how the introduction of the transformer architecture led to an improvement in downstream NLP tasks, without the need for recurrent or convolutional layers, which in turn allows for a reduction in training time [5].

7.5 Proposing Attention Mechanisms for Metric-Learning in Parametric DR

This section approaches the task of leveraging neural network architectures from other domains to the specific task of metric-learning during parametric dimensionality reduction. The main focus of this is in addressing RQ3.1, RQ3.2, RQ3.3, RQ3.4 and RQ3.5, and serves as an effective method of first evaluating the proposed architectures upon tasks where evaluation metrics are clearly defined, and thus can clearly indicate the influence that the proposed architectures have upon downstream clustering performance. This is conducted through a comparison of existing algorithms for DR, and two proposed architectures implemented into the parametric UMAP algorithm, which have not been previously discussed in the literature for this task. After establishing the positive influence that these architectures have on downstream clustering accuracy, the next section, Section 7.6, investigates expanding the transformer-encoder architecture through the introduction of additional residual connections in order to directly contribute to topic modelling.

7.5.1 Datasets Selected for the Analysis

To demonstrate that the outcomes of dimensionality reduction, when applied to text embeddings, are generalisable, we performed experiments on three datasets: 20 Newsgroups [94], Text REtrieval Conference (TREC) [113] and AG's News[4]. They

[4] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

were selected because each of these provide a textual representation of the data from a range of domains. In the case of the 20 Newsgroups, data is arranged into 20 "newsgroup" categories, with each document being assigned to a single category, representing the topic of the document. The 20 Newsgroups dataset is widely cited [94] and well established as a benchmark dataset for classification [297], clustering [298, 299], and topic modelling tasks [300–303], and was selected to allow reproducibility of results of this work and straightforward comparison with any future investigations. The dataset is accessible easily through the scikit-learn framework[5]. The TREC Question Classification dataset provides 5,500 training documents and 500 test documents consisting of labelled textual questions, resulting in 6000 labelled documents overall. Labels for this dataset are provided as coarse-grained and fine-grained, wherein the coarse-grained set has 6 class labels, with the fine-grained having 47 class labels. We evaluate both labelling formats, to estimate the influence of the number of classes on accuracy. This dataset was originally intended as a classification task [113], however, we investigate this dataset from a clustering perspective for two reasons. Firstly, the dataset is imbalanced, particularly for the finer-grained labels, which presents opportunities to analyse how the DR algorithms in our investigation are affected by such an imbalance. Second, the data set is relatively small by modern standards, which presents the opportunity to again investigate how this affects the DR algorithms. The AG's News corpus provides 127,600 news articles, consisting of titles and description fields of articles collected from more than 200 news sources over a year through the ComeToMyHead search engine, with each document being assigned one of four class labels from either "*World*", "*Sports*", "*Business*" or "*Sci/Tech*" news categories. This dataset features no class imbalance; however, it is significantly larger than the others used in our study, again presenting an opportunity to compare how dataset size influences downstream clustering following DR. These datasets were selected to provide a range in corpus size and class numbers, as we seek to investigate how these may impact upon the proposed methodology. Furthermore, this ensures that additional

[5] `https://scikit-learn.org/stable/index.html`

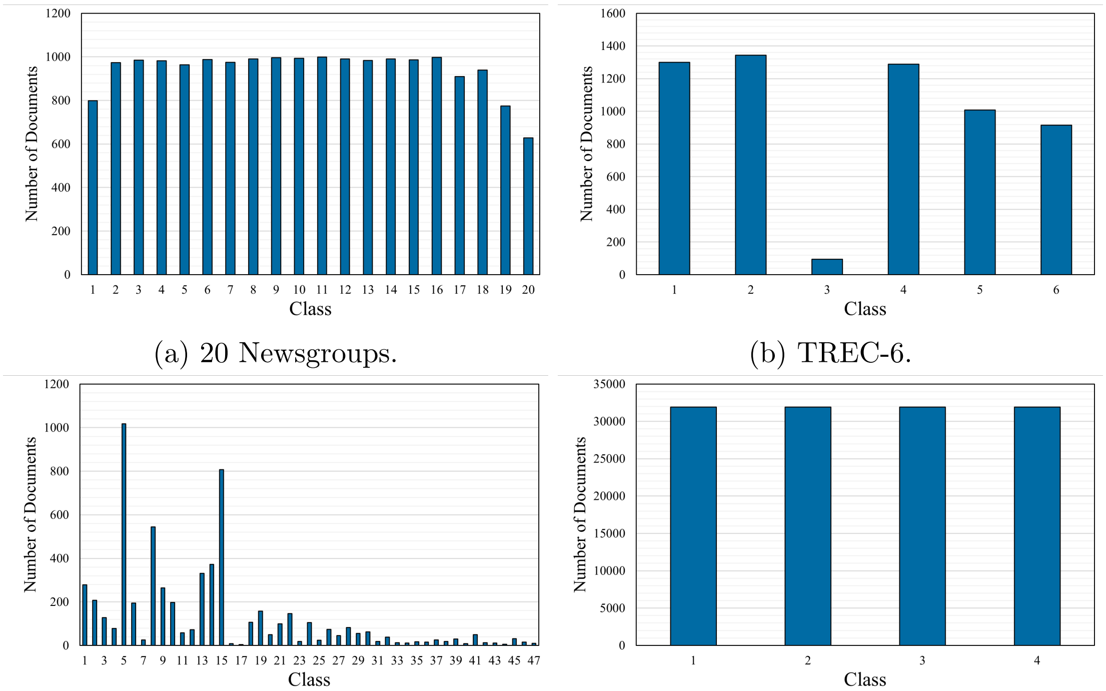
comparisons of DR algorithms are provided on different datasets. An overview of these is provided in Table 7.1.

For each dataset, we use the RoBERTa pre-trained transformer network [31], for the computation of embeddings of documents within the dataset. These may then be passed to the dimensionality reduction algorithms used in the experiments. The embeddings produced by RoBERTa are contextual and therefore it is not necessary to perform "traditional" preprocessing, such as stopword removal or lemmatisation. This is due to the fact that RoBERTa, which is built upon the work of BERT [40], adopts a masked language-model (MLM) strategy for pre-training, where tokens are randomly masked, with the objective of the training being the prediction of the masked term based only on its context [31,40]. Thus, no preprocessing is performed, as this would affect the contextual information entailed within the text. However, one exception to this is the 20 Newsgroups dataset, where the original dataset, consisting of data extracted from online forums, also contains header and footer information, with personal information, such as email addresses and names of the users. We remove these, due to both ethical considerations, as well as due to them not contributing any useful information to the clustering task.

Dataset	**20 Newsgroups**	**TREC-6**	**TREC-50**	**AG News**
Documents	18846	6000	6000	127,600
Classes	20	6	50	4

Table 7.1: Total number of records and classes present in the 20 Newsgroups, TREC and AG's News datasets

Figure 7.1 demonstrates the number of documents present in each class of the selected datasets. In the 20 Newsgroups dataset (Figure. 7.1a), a significant portion of the classes feature a small degree of class imbalance, with the smallest class being class 20 with 628 documents, and the largest, class 16, having 997 documents. In comparison, in Figure 7.1b, it is clear that for the TREC-6 dataset, there is a significant imbalance in class sizes, with the smallest (class 3) containing only 95 documents. This imbalance is even more prevalent in the TREC-50 dataset (Figure 7.1c), which uses the same corpus as TREC-6, however, with a finer-grained labelling scheme. In this dataset, the smallest class contains only 4 documents. Finally, the



(a) 20 Newsgroups.

(b) TREC-6.

(c) TREC-50.

(d) AG's News.

Figure 7.1: Number of documents assigned to each class.

AG's News dataset (Figure 7.1d) features no class imbalances, with each class having 31,900 documents.

7.5.2 Consideration of Overfitting

For the investigation of a supervised learning methodology with UMAP, it is important to consider the impact of class imbalance on the dimensionality reduction process and, subsequently, the downstream clustering task. Anticipated effects include reduced class-specific accuracy in under-represented classes, such as class 20 in the 20 Newsgroups dataset or class 3 in the TREC-6 dataset (as shown later in Results Figure 7.5). We present an analysis of these implications in Section 7.5.8. In downstream clustering, class imbalance has been demonstrated to have a considerable impact on the clustering result for the k-Means algorithm [304]. k-Means was found to tend to produce clusters of uniform size [305], even in diverse datasets with imbalanced data, leading to suboptimal clustering results. To address this, a variety of strategies are available to mitigate the effects of class imbalance, broadly classified into three primary groups: resampling techniques, algorithm-level adjustments, and hybrid methods [306]. Resampling techniques seek to address the imbalance at a data level, typically using over-sampling or under-sampling methods, where the data is adjusted in order to decrease prevalence of the skewed class distribution within the dataset [307]. Of these, one of the most prevalent is the SMOTE algorithm, wherein synthetic minority class examples can be introduced to the dataset to address the imbalance [308]. Algorithm-level adjustments incorporate adaptations to algorithms, such that they take into account the skew in data. For k-Means, examples of this include the introduction of artificial neural networks for the determining of the initial cluster centroids [309], or the introduction of 'multicenter' clustering variant, where multicenters are used to determine each cluster, rather than one centroid per cluster [304]. Additionally, cost-sensitive methods [310] combine algorithm and data-level techniques, to assign a misclassification cost for each class based on evaluation methods [311].

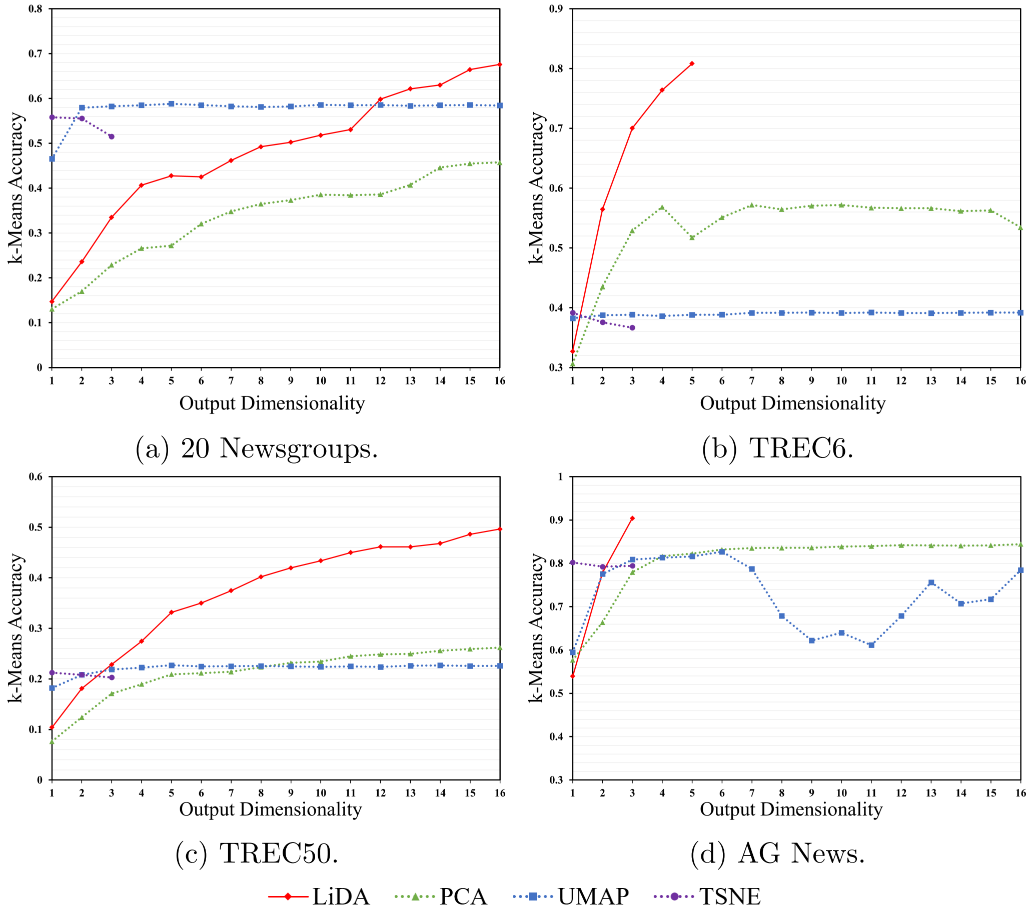
To mitigate overfitting in the transformer-encoder architecture, instead of using sampling techniques that may not accurately represent the data, randomised dropout layers are incorporated within the network architecture. This is applied through the random omission of units within the neural network, which has been shown to be effective in addressing overfitting [312], [5].

For our task, which differs from the language modelling task for which the architecture was originally intended, we apply dropout following the multi-head attention layer of the transformer block, similar to the original transformer encoder [5]. However, a second layer of dropout is then applied to the final feedforward layer of the architecture, prior to the output layer. The amount of dropout in each model is determined through an optimisation strategy. The hyper-parameter optimisation strategy is discussed in Section 7.5.5.

7.5.3 Preliminary Evaluation of Existing Techniques for Dimensionality Reduction

Uniform Manifold Approximation and Projection [14], t-SNE [86], Principal Components Analysis [84], and Linear Discriminant Analysis [85] represent state-of-the-art benchmark methods in dimensionality reduction. We assess these methods in terms of when applied as a DR technique prior to clustering, thereby addressing our initial two research questions (**RQ3.1**, **RQ3.2**). LiDA requires the provision of labelled data. Hence, we provide this as a randomly shuffled subset of 20% of the whole dataset, from which we compute the low-dimensionality representations of the full dataset.

The computational complexity for the algorithms analysed in the preliminary investigation differ considerably, and therefore their use is best suited for different situations and datasets. For LiDA, the time complexity is $O(Ndt+t^3)$, and memory requirement is $O(Nd+Nt+nt)$, where N is the number of samples, d is the number of features, or the dimensionality of the data, and $t = min(N, d)$. In instances where N and d are large, the algorithm becomes infeasible, as discussed by [313], who also evaluated the algorithm on the 20 Newsgroups dataset, where they identified a considerable increase in time complexity for large samples of the dataset. For PCA, a time complexity of $O(min(N^3, d^3))$ is outlined by [314]. Regarding t-SNE, there is a significant limitation due to the computational complexity of the algorithm, which scales at a degree of $O(N^2)$, where N is the number of data points [89]. The application of the Barnes-Hut algorithm as an approximation method for the gradient calculation algorithm can enhance efficiency to $O(\log N)$ time complexity, as demonstrated in [90, 91]. However, it is important to note that this approach is applicable only when the output dimensionality is less than or equal to 3 dimensions. Finally, in the case of UMAP, empirical results indicate an approximate complexity of $O(N^{1.14})$, which is bounded by the complexity of the approximate nearest neighbour algorithm, and has, at this time, no theoretical proof [14, 109]. On the basis of these, it would appear that UMAP presents the best scalability in terms of time complexity. As discussed in Section 7.4.1, the "curse of dimensionality" arises in clustering algorithms, due to the impact that high dimensionalities have upon distance measurements, which are necessary in determining distances between points, when assigning them to clusters. k-Means has a time complexity of $O(N^{dk+1})$ [315], where d is the dimensionality, N is the number of points to cluster, and k is the number of clusters. Therefore, opting for the lowest-dimensional representation prior to clustering is beneficial when working with large datasets, if it can be proven that a low-dimensional representation will perform adequately.



(a) 20 Newsgroups.

(b) TREC6.

(c) TREC50.

(d) AG News.

Figure 7.2: Comparison of LiDA, PCA, t-SNE and UMAP DR techniques upon accuracy in downstream k-Means clustering.

Figures 7.2a-7.2d demonstrate the clustering accuracy attained by k-Means clustering across dimensionality ranges $d = \{1, ..., 16\}$ for UMAP, PCA, LiDA and t-SNE. Notably, LiDA indicates a positive correlation with respect to an increase in accuracy relative to dimensionality across all datasets, outperforming both UMAP and PCA. However, there are limitations to this method, as LiDA relies upon the presence of labelled data. Furthermore, the algorithm is not capable of generating embeddings with a dimensionality greater than $(u - 1)$, where u is the total number of unique labels present in the data. Most notably, as demonstrated in Figures 7.2a and 7.2c, in some cases, the dimensionality must be large, in order to obtain optimal clustering accuracy. This is not ideal based on the increase in computational complexity that is observed by clustering algorithms when analysing highly dimensional data. PCA outperforms, or performs comparably to UMAP, the established state-of-art, on the TREC6, TREC50 and AG News datasets. As with LiDA, accuracy for downstream clustering by k-Means appears to improve as the dimensionality increases. t-SNE performs comparably with UMAP at low-dimensionalities; however, this is evaluated only up to an output dimensionality of 3, due to the Barnes-Hut approximation algorithm used in the algorithm restricting output dimensionality to below 4 dimensions.

7.5.4 Parametric UMAP with Bidirectional Recurrent Networks with Attention

Following the preliminary evaluation of UMAP, t-SNE, PCA and LiDA, a *novel pipeline of a bidirectional RNN with attention mechanism for DR* is proposed, which is facilitated through parametric UMAP [4], using a *metric learning methodology*, wherein the parametric dimensionality reduction model is trained on a small subset of labelled data. It is hypothesised that the introduction of supervised learning to the computation of lower-dimensionality embeddings could improve downstream clustering performance. Furthermore, *this work aims to demonstrate how the definition of a recurrent network with attention could potentially improve clustering accuracy*, given a small amount of training data. Therefore, only a sample of 20% from the dataset is used in the training of the supervised metric learning model. This performance is compared with other configurations of UMAP, including UMAP itself, supervised UMAP, parametric UMAP and parametric supervised UMAP. Parametric UMAP in this case is configured with a default network architecture of 3 fully connected layers consisting of 100 units each.

We propose a recurrent neural network with self-attention mechanism, to *investigate the impact of recurrent networks and attention upon the parametric learning of low-dimensionality embeddings*. Figure 7.3 demonstrates this overall architecture, which consists of 2 blocks of stacked RNN-attention layers, followed by a fully connected layer. This results in a total of $35,119,359$ trainable parameters. The configuration of this architecture and the assigned number of nodes for each layer is identified through the Tree-structured Parzen Estimator [316–318] hyper-parameter optimisation, facilitated by the Optuna framework [319]. For this optimisation strat-

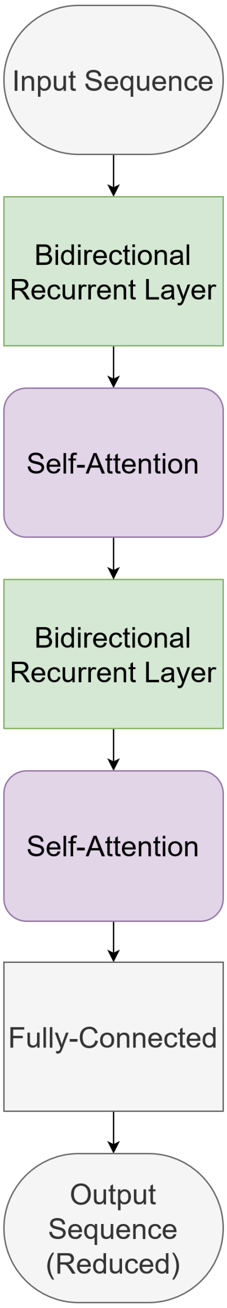

Input Sequence
Bidirectional Recurrent Layer
Self-Attention
Bidirectional Recurrent Layer
Self-Attention
Fully-Connected
Output Sequence (Reduced)

Figure 7.3: Bidirectional recurrent architecture with attention mechanism using gated recurrent units

egy, we set the optimisation objective as the maximising of accuracy score for the k-Means clustering of the whole dataset, with the model being provided with only a 20% subset of the data.

7.5.5 Parametric UMAP with Transformer-Encoder

We define a transformer-encoder, the main crux of our investigation, based on the original architecture [5], consisting of a stack of N transformer blocks, where each transformer block comprises two sub-layers. The first of these sub-layers consists of a multi-head self-attention mechanism, while the second is a fully connected feedforward network with ReLU activation [320]. After both the multi-head attention, and feedforward layer, layer addition is performed. This entails the concatenation of the outputs of the previous layer with the original input sequence. This layer addition process, which is also known as residual connection, is intended to mitigate the vanishing gradient [17] problem, wherein during backpropagation, the multiplication of the gradients of each layer, if they are smaller than 1, leads to exponentially decreasing gradients. It is reported that as the sequence length of a model increases, the gradient magnitude typically decreases, which can slow down or even stop the training process [18].

Using the same hyper-parameter optimisation strategy outlined when designing the RNN with attention, we identify an optimal network configuration consisting of four sequential transformer-encoder blocks, followed by a fully-connected layer. The amount of dropout present within the transformer block, and before the output layer of the model, is also defined through this same hyper-parameter optimisation strategy. Contrary to the implementation of dropout in [5], the strategy for hyper-parameter optimisation pinpoints an ideal setup where the transformer-encoder block does not undergo any dropout. Instead, dropout is exclusively implemented on the model's final feedforward layer, at a rate of 5%. This configuration results in a total of $10,784,272$ trainable parameters, which is more than three times fewer than the total parameters of the RNN with attention architecture. For both the transformer-encoder, and RNN with attention, we investigate our fourth research question (**RQ3.4**).

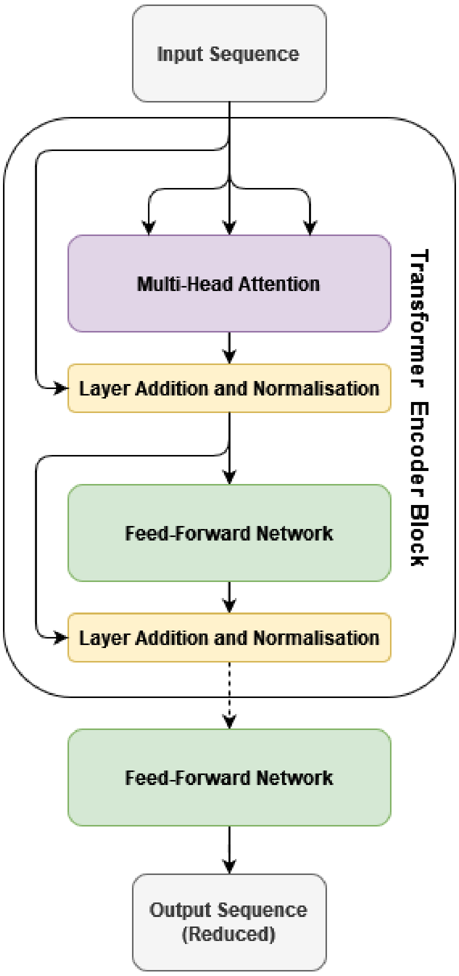

Input Sequence
Multi-Head Attention
Layer Addition and Normalisation
Feed-Forward Network
Layer Addition and Normalisation
Transformer Encoder Block
Feed-Forward Network
Output Sequence (Reduced)

Figure 7.4: Transformer-encoder architecture

7.5.6 Experimental Setup

To evaluate the outcomes of our proposed methods, namely the implementation of the *transformer-encoder*, and *RNN with attention* upon metric learning for DR, we propose the following experiment, comparing clustering accuracy across a range of UMAP variations, where we seek to evaluate both **RQ3.3** and **RQ3.4**. These are performed upon the four benchmark datasets applied for the preliminary evaluation in Section 7.5.3. For each dataset, we evaluate UMAP, Parametric UMAP (P-UMAP), UMAP Supervised, Parametric UMAP Supervised, Parametric UMAP with RNN Supervised, and Parametric UMAP with Transformer-Encoder Supervised, with the latter being our two proposed novel architecture pipelines demonstrated in Figure 7.3 and Figure 7.4. In the supervised cases, dimensionality reduction models were trained upon the same subset of 20% of the overall dataset, to perform metric learning of the reduced embeddings.

Similarly to the preliminary evaluation (7.5.3), we compare the accuracy across dimensionalities ranging from $d = \{1, ..., 16\}$ for each algorithm against a baseline score, where k-Means clustering is performed upon the original RoBERTa [31] embeddings, without dimensionality reduction being applied. Results are based on an average taken over 25 separate experiments, which allows for the calculation of statistical significance using a Wilcoxon-Mann-Whitney U Test [321]. We present the accuracy for each algorithm, across each dataset, for each dimensionality, in all 25 iterations of each experiment[1].

Clustering Evaluation

Evaluation of the performance relative to dimensionality involves calculating the accuracy of the downstream clustering task. The integer label assigned by the k-Means algorithm to a cluster may not directly reflect the integer label assigned as the true label, even if the clustering solution is correct. For example, a clustering solution may assign cluster A a label of 0, and cluster B a label of 1, however, the true label present in the dataset is 1 for cluster A, and 0 for cluster B. When we examine the results, we may find that if the error is low and the clustering solution is correct, then it is necessary to map the predicted clusters to their associated true labels for evaluation. Therefore, the Kuhn-Munkres Algorithm [322] is applied to map the assigned label, by clustering, to the true label. This allows the framing of the evaluation as a supervised learning task to obtain an accuracy score. We calculate the accuracy as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.1}$$

Where TP represents the true positives, TN represents true negatives, FP represents false positives and FN represents false negatives.

7.5.7 Testing for Statistical Significance in Results

We select the Mann-Whitney-Wilcoxon (WMW) U test [321] when testing the significance of experimental results. WMW has been shown to function adequately when applied to a smaller sample size, such as a population of 25 [323]. The WMW test is a nonparametric test that makes no assumptions about the distribution of the data and, as such, is suitable when data is not normally distributed. For our experiments in Section 7.5.8, we use a population size of 25. While it would be preferable to conduct our experiment across a larger population size, this would require extensive compute requirements and therefore take significantly longer to evaluate across all dimensionalities and datasets. To test the significance in the results, we adopt a significance level $\alpha = 0.01$. The significance between two arguments is computed using the complete set of 25 iterations for each experiment, where an iteration represents the training of the dimensionality reduction algorithm, and subsequent clustering by k-Means. This guarantees that the given scores encapsulate a comprehensive depiction of the performance exhibited by the corresponding algorithms. For example, if testing for significant difference between the accuracy of the UMAP algorithm, when compared to the Supervised Parametric UMAP algorithm, with both having an output dimensionality of 3, for the 20 Newsgroups dataset; Sample 1 would entail the 25 accuracy scores, which were attained by k-Means algorithm when clustering the low-dimensional vectors produced by UMAP with an output dimensionality of 3. Sample 2 would be the 25 accuracy scores for the clustering of the vectors produced by the Supervised Parametric UMAP algorithm, using the same criteria. We form a null hypothesis H_0 that there is no significant difference between two comparisons, and an alternative hypothesis H_a that there is significant difference between results. For each statement we make with respect to statistical significance, we provide the U-statistic U, z-score, and probability p, as well as providing the standard deviation SD, and mean accuracy score M for each individual population.

7.5.8 Results

Cluster Analysis

Figures 7.5 and 7.6 represent a two-dimensional plot of the distribution of vectors produced by the dimensionality reduction techniques of UMAP, supervised UMAP, supervised UMAP with RNN and Attention, and supervised UMAP with transformer network. This is provided as a visual demonstration of the clustering solution, which can aid in evaluating how the different DR algorithms can partition the data. Points are colour-coded, to correspond to the true class label associated with each document. To enhance visual comprehension, we did not visualize TREC-50 and 20 Newsgroups in this work due to the large number of classes, which makes it challenging to distinguish colors. Colour maps must be carefully selected to ensure interpretation by readers with colour vision deficiency [324], however, these typically provide a maximum of 10 clearly discernible colours for categorical data, such as the *tableau-colorblind10* colour map of the *matplotlib*[6] visualisation library. Still, visualising a larger number of classes could be achieved by using different point styles, such as dashes and stars.

We observe a notable distinction in the arrangement of points when comparing *unsupervised* (UMAP and UMAP parametric, Figures 7.5a and 7.6c) with *supervised* approaches (UMAP supervised, UMAP parametric supervised, UMAP supervised, RNN with attention, and UMAP supervised transformer, Figures 7.5b, 7.6d, 7.5e and 7.5f). When trained on the subset of data, all supervised configurations of UMAP indicate greater effectiveness in the partitioning of documents based upon their class-assigned label, in comparison with their unsupervised variants. However, there appear to be only five distinct clusters extracted for all supervised variants, reflecting the class imbalance in the TREC-6 dataset identified in 7.1b, where class 3 is highly under-represented, with only 95 documents. When taking into account the 20% sample taken for training, this provides only 19 labelled documents representing class 3. Overall, both the RNN with attention, and transformer architectures, appear

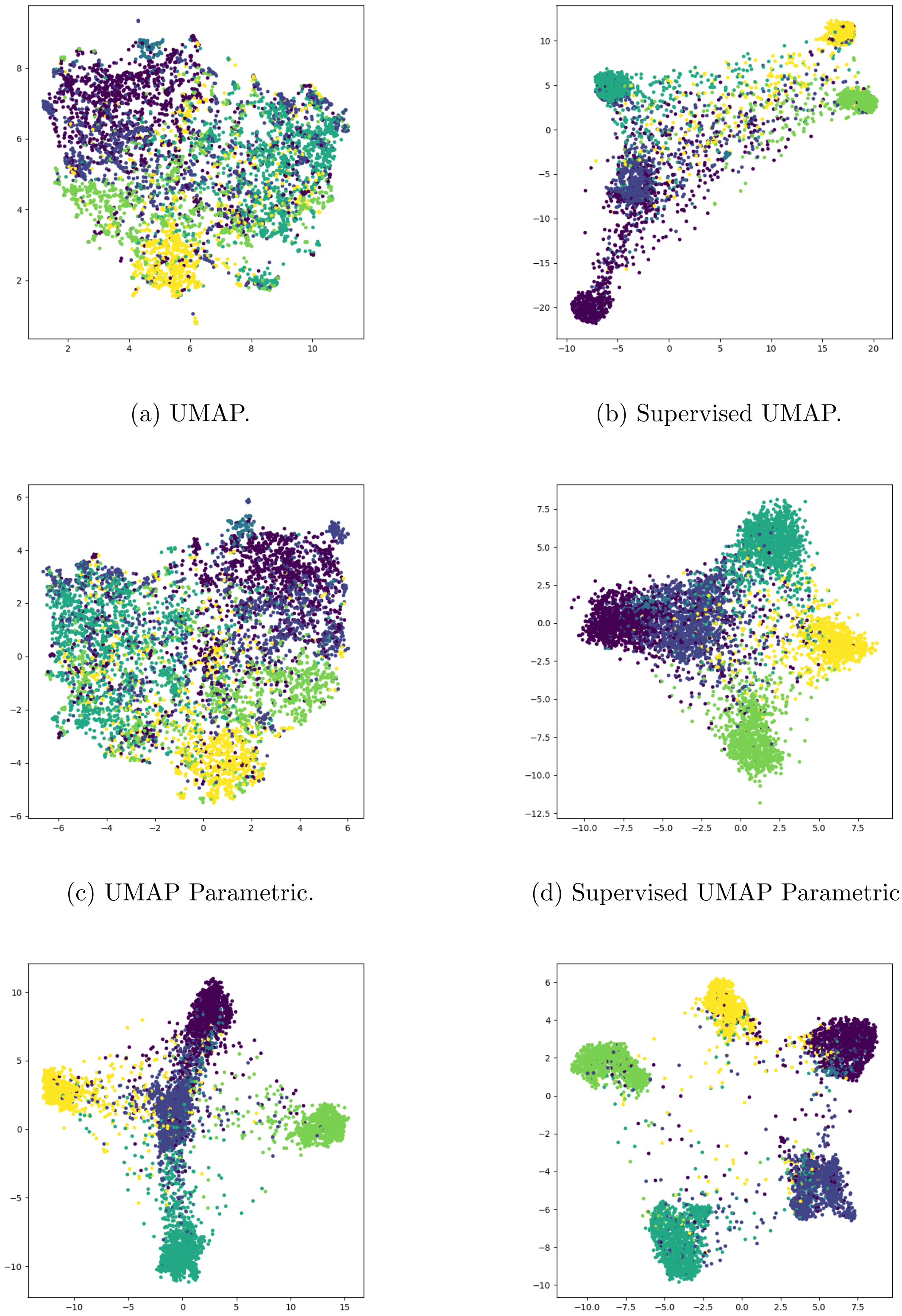
[6] `https://matplotlib.org/stable/gallery/style_sheets/style_sheets_reference.html`

to provide improvements in the separation of clusters, with the transformer providing the clearest separation. However, these appear to form only 5 clusters, which is best demonstrated in Figure 7.5f. It appears that the class imbalance and under-representation of class 3 in the TREC-6 dataset leads to this cluster failing to be represented by UMAP in all its variants.

Accuracy of Clustering

Figure 7.9a demonstrates the clustering accuracy of the models for the 20 Newsgroups dataset. When testing for significance in subsequent analyses in this section, we apply a WMW test as detailed in Section 7.5.7. For a two-tailed WMW test with a population size of 25, the critical value of U is 180, such that any U value greater than this is rejected as being statistically significant. It is worth noting that in our analyses, a U value of 0 is frequently observed. This is due to the prevalence of all observations in one population having a score lower than all observations in the other population, and implies a perfect separation between two groups. Individual results used in our experiments can be accessed at the downloadable repository[1].

At a glance, it is evident that all configurations of UMAP provide an improvement in accuracy when compared to the baseline, with a clear "knee" that can be observed between an output dimensionality of 2 and 3. At a dimensionality of 3, which we adopt for all subsequent calculations, comparing UMAP [14] ($SD = 0.007$, $M = 0.582$) with the k-Means baseline score ($SD = 0.012$, $M = 0.517$) indicates a *significant* improvement of 6.5% ($U = 0$, z-score $= 6.054$, $p < 0.0001$) and parametric UMAP (P-UMAP) ($SD = 0.013$, $M = 0.579$) [4] improving *significantly* by 6.1% ($U = 0$, z-score $= 6.054$, $p < 0.001$), compared to the baseline score. The introduction of a metric learning approach, trained upon a subset of data, is demonstrated by the UMAP Supervised ($SD = 0.013$, $M = 0.668$), where we accept the alternative hypothesis, indicating a *significant* improvement compared baseline score of 15% ($U = 0$, z-score $= 6.054$, $p < 0.001$). For the supervised UMAP parametric (P-UMAP Supervised) ($SD = 0.018$, $M = 0.562$) algorithm, we again accept the alternative hypothesis, indicating a *significant* improvement of 4.4% ($U = 0$, z-score $= 6.054$, $p < 0.001$). This supports our investigation into **RQ3.3**, into *evaluating*



(a) UMAP.

(b) Supervised UMAP.

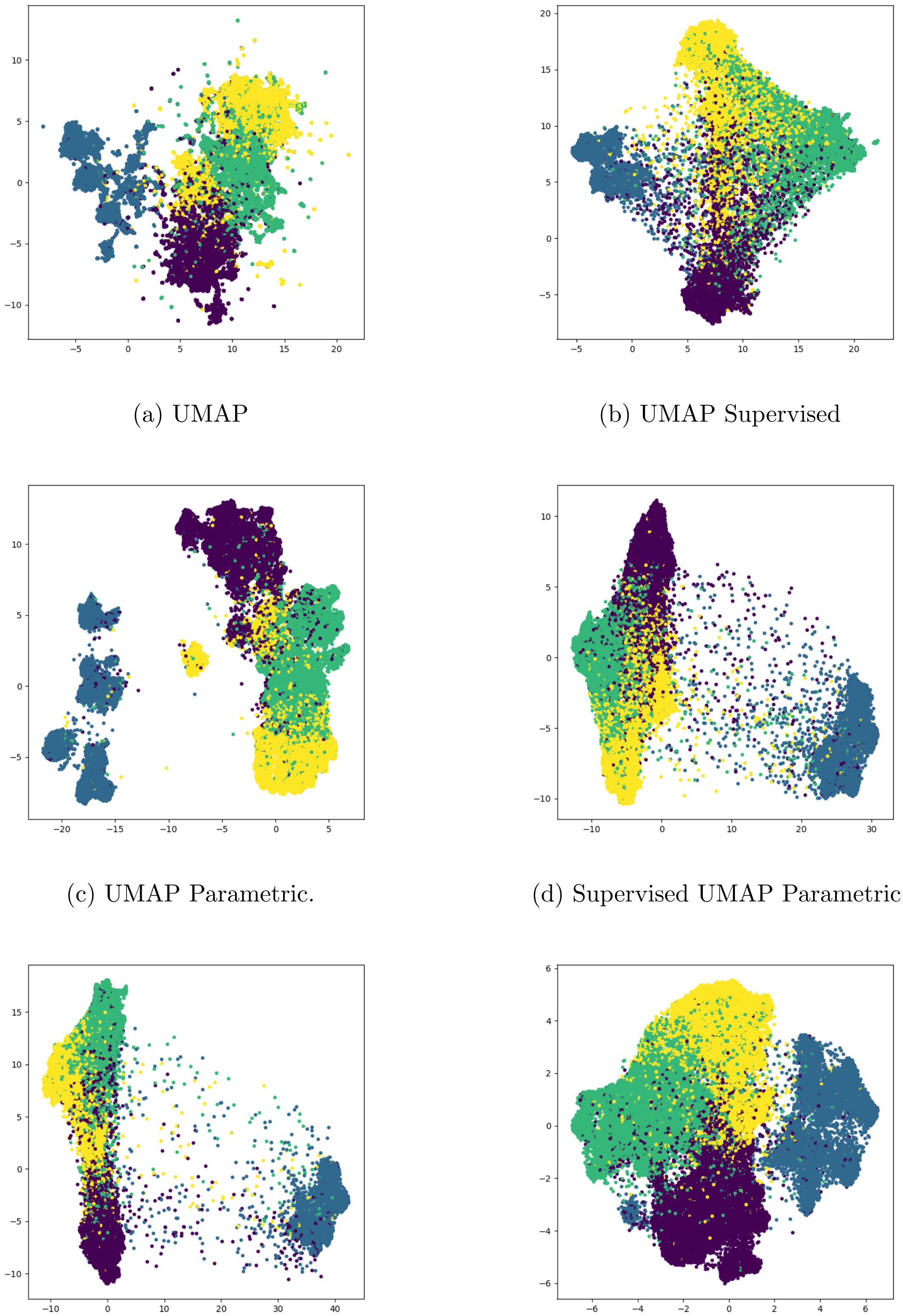
(c) UMAP Parametric.

(d) Supervised UMAP Parametric.

(e) Supervised UMAP with RNN and Attention.

(f) Supervised UMAP Transformer-Encoder.

Figure 7.5: Visualisation of reduced vectors at a dimensionality of 2 for TREC6, for our proposed methods (underlined), compared with existing UMAP configurations, with colours representing the true label for each point.



(a) UMAP

(b) UMAP Supervised

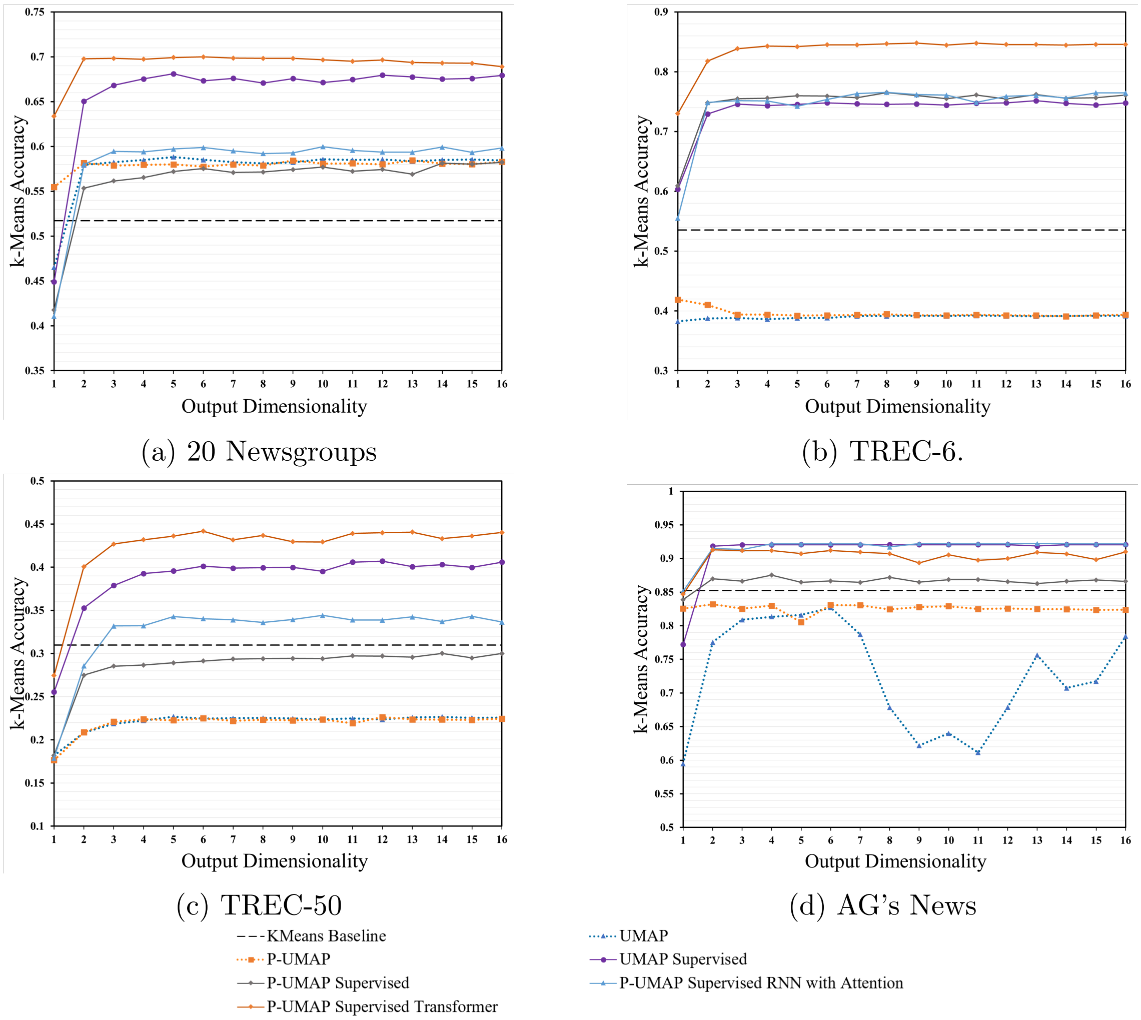
(c) UMAP Parametric.

(d) Supervised UMAP Parametric.

(e) UMAP Supervised RNN With Attention.

(f) UMAP Supervised Transformer-Encoder.

Figure 7.6: Visualisation of reduced vectors at a dimensionality of 2 for AG's News, for proposed methods (underlined), compared with existing UMAP configurations, with colours representing the true label for each point.



(a) 20 Newsgroups

(b) TREC-6.

(c) TREC-50

(d) AG's News

Figure 7.7: Accuracy of k-Means clustering with dimensionality ranging from 1 to 16 for the proposed methods (our top proposed ones underlined), compared with existing UMAP configurations.

metric learning as a suitable method to be used prior to clustering. There is a significant difference between supervised UMAP, and supervised parametric UMAP algorithm, wherein at a dimensionality of 3, the nonparametric algorithm attains an accuracy 10.6% higher than the parametric variant ($U = 0$, z-score $= 6.054$, $p < 0.001$). As the default configuration of the parametric UMAP neural network architecture consists of only fully-connected layers, it is worth evaluating whether the implementation of more complex architectures can contribute to improving the performance of the parametric UMAP model. It is observable in both unsupervised and supervised cases that the nonparametric UMAP algorithm attains a higher accuracy score compared to the parametric version consisting of the default fully connected layers. However, the transformer-encoder (P-UMAP Transformer) (SD = 0.0083, M = 0.698) model attains the greatest *significant* improvement in accuracy relative to the baseline, of 18.1% ($U = 0$, z-score = 6.054, $p < 0.001$). The supervised RNN with attention (SD = 0.0103, M = 0.594), in comparison, provides a *significant* accuracy improvement of 7.7% ($U = 0$, z-score = 6.054, $p < 0.001$) compared to the baseline score.

When comparing the transformer-encoder with the next highest scoring algorithm, UMAP supervised, it is evident that the transformer-encoder attains a *overall significant* higher-scoring accuracy in downstream clustering, even at a lower dimensionality. At a dimensionality of 2, the transformer-encoder (SD = 0.007, M = 0.698) exhibits a *significant* difference 4.7% greater ($U = 0$, z-score = 6.054, $p < 0.001$) than UMAP supervised (SD = 0.013, M = 0.65). At a dimensionality of 3, this *significant* difference is 3% greater ($U = 0$, z-score = 6.054, $p < 0.001$) for the transformer-encoder (SD = 0.008, M = 0.698), compared with UMAP supervised (SD = 0.013, M = 0.668). This indicates that an *advantage of our proposed pipeline using the transformer-encoder*, is that a higher accuracy can be achieved in downstream clustering at lower dimensionalities, which has benefits with regards to computational complexity, and memory efficiency corresponding to the storage of the smaller vectors. These results are depicted in Figure 7.9a.

In Figure 7.9b, when applied to the TREC-6 dataset, we observe a decrease in the average accuracy relative to the k-Means baseline for both UMAP, and parametric UMAP. At an output dimensionality of 3, the UMAP algorithm (SD = 0.002, M = 0.039) is on average 14.7% ($U = 0$, z-score = 6.054, $p < 0.001$) poorer than the k-Means baseline score, which is *significant.* Similarly, parametric UMAP ($SD = 0.008, M = 0.392$) performs 14.3% worse than the baseline ($U = 0$, z-score = 6.054, $p < 0.001$). This decrease is similar for the TREC-50 dataset, where at a dimensionality of 3, UMAP (SD = 0.005, M = 0.22) is on average 9.1% less accurate than the baseline score ($U = 0$, z-score = 6.054, $p < 0.001$), and parametric UMAP (SD = 0.004, M = 0.22) is 7.1% ($U = 0$, z-score = 6.054, $p < 0.001$) less accurate. This is of interest, as it indicates that there is no guarantee that UMAP can contribute to improvements in downstream clustering accuracy. Considering that both TREC-6, and TREC-50 have the lowest number of documents, consisting of only 6000 rows (see Table 7.1), this may have an influence upon the generalisation of the unsupervised model for UMAP. In comparison, all supervised variants of UMAP demonstrate an improvement in clustering accuracy for both TREC-6, and TREC-50 datasets. Most notably, the transformer-encoder model attains the greatest improvement in accuracy relative to the baseline score. At a dimensionality of 3, this is a significant improvement of 30.3% for TREC-6 ($U = 0$, z-score $= 6.054$, $p < 0.001$), attaining an average accuracy that is greater than all other DR algorithms investigated in our study.

The P-UMAP transformer-encoder architecture achieves the greatest accuracy for three of our datasets, 20 Newsgroups, TREC-6, and TREC-50, which reflects the outcomes observed in our visualisation analysis of TREC-6 in Figure 7.5. However, on the fourth, the AG's News dataset, the P-UMAP transformer-encoder does not outperform UMAP Supervised, the current SoA (Figure 7.7d). Nevertheless, both of our proposed architectures, the transformer-encoder, and RNN with attention, perform comparably with UMAP supervised for the AG's News dataset.

An interesting observation, which is evident only for the AG News dataset, is the high degree of variance in downstream clustering accuracy for the UMAP algorithm. This is prevalent between the dimensionalities of 6 and 16, where a considerable decrease in the accuracy of downstream clustering appears. This presents a decrease in the average accuracy of downstream clustering from 82.6% ($SD = 0.054$) at a dimensionality of 6, to a minima of 61.1% ($SD = 0.08$) at a dimensionality of 11. Furthermore, within the scope of our experiments, it was observed that the population of results for UMAP exhibited the highest standard deviation of accuracy, at a dimensionality of 11 ($SD = 0.08$). This was the most significant fluctuation in accuracy across all conducted tests.

A general observation across all experiments (Figure 7.7) indicates the *presence of a "knee" (or "elbow") in accuracy* for all derivatives of UMAP , which becomes apparent between the dimensionalities of $\{2, 3\}$, when the reduced dimensionality embeddings are used in k-Means clustering. Increasing the output dimensionality of embeddings beyond this degree tends to have little or diminishing influence on the quality of the clustering solution for k-Means. This is of interest, and merits a further discussion, as we have, at this time, found *no evidence of this phenomenon within the literature.* Notably, it is empirically apparent that the k-Means algorithm experiences a decrease in the rate of improvement beyond an output dimensionality of 3 for low-dimensional representations produced by UMAP and UMAP derivatives, for both a supervised and unsupervised training manner. When considering computational complexity, there are numerous benefits to choosing a lower output dimensionality. As discussed in Section 7.5.3, clustering algorithms, such as k-Means, can be affected by data dimensionality. Notably in the case of k-Means clustering, the time complexity scales quadratically with relation to the dimensionality and the number of clusters. Therefore, it is often advantageous to perform clustering using a representation of the data with reduced dimensionality. Considering the marginal accuracy improvements beyond the knee curve point, it could be more efficient to select a small output dimensionality, such as 2 or 3, to optimise the runtime of the clustering solution.

Overall, the proposed architecture based on the transformer-encoder is demonstrated to contribute to significant improvements in clustering accuracy across three of the four experiments conducted. The RNN with attention also contributes to improvements in accuracy relative to existing methods, and outperforms the transformer-encoder by a small margin on the AG's News dataset. This addresses both **RQ3.3** and **RQ3.4** outlined in the introduction (Section 7.2).

The maximum average accuracy based upon 25 iterations of each experiment at a dimensionality of 3, for each algorithm across our experiments is summarised in Table 7.2.

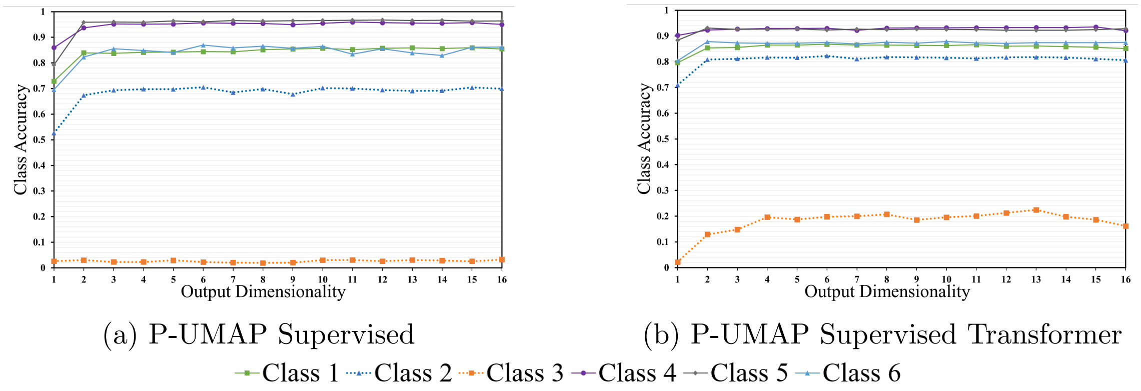
Overall, the results of these experiments are of value to any applications of the parametric UMAP transformer-encoder to dimensionality reduction in downstream tasks, as the model requires fewer trainable parameters, compared to the RNN with attention.

Table 7.2: Average accuracy at an output dimensionality of 3 based on 25 iterations of each experiment

	Dataset			
Algorithm	**20-NG**	**TREC-6**	**TREC-50**	**AG News**
KMeans (No DR)	0.517	0.535	0.31	0.85
LiDA	0.335	0.70	0.228	0.904
PCA	0.229	0.529	0.211	0.779
t-SNE	0.515	0.367	0.203	0.794
UMAP	0.582	0.388	0.219	0.809
P-UMAP	0.579	0.392	0.22	0.84
UMAP Supervised	0.668	0.746	0.379	**0.92**
P-UMAP Supervised	0.562	0.755	0.285	0.866
P-UMAP RNN+Attention	0.595	0.752	0.334	0.913
P-UMAP Transformer	**0.698**	**0.839**	**0.427**	0.911

Analysis of Per-Class Accuracy

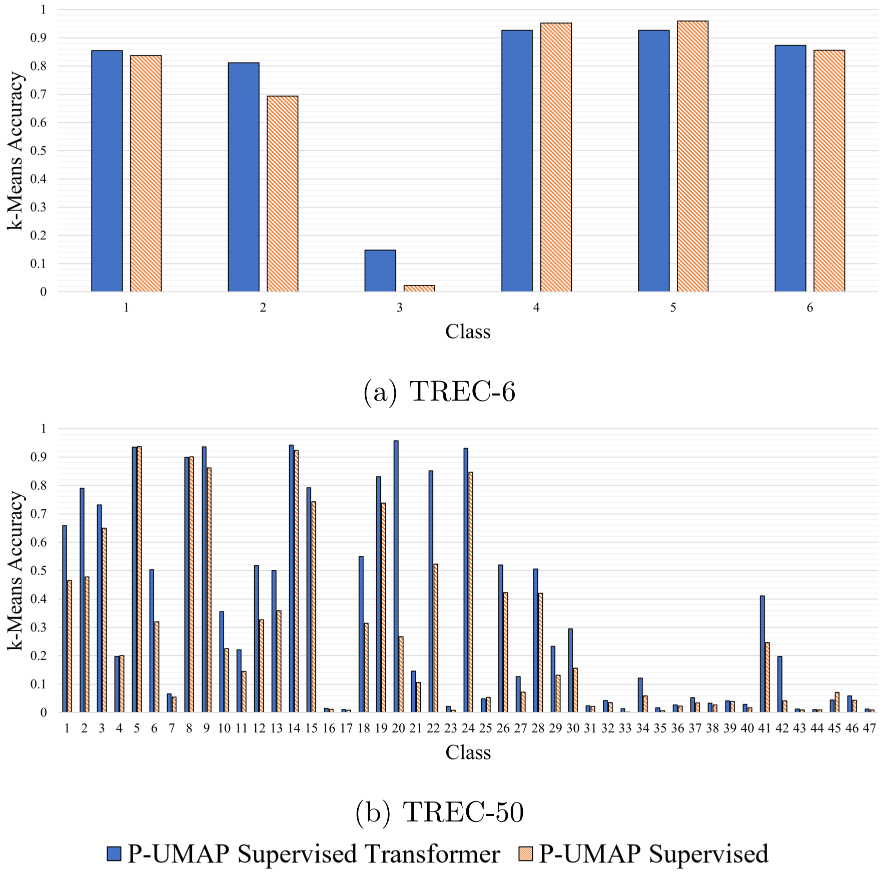
In section 7.5.1, we discussed the presence of a considerable imbalance in the TREC-6, and TREC-50 datasets. Given the supervised learning nature of our methodology, we consider it essential to explore the performance of the proposed transformer-encoder, given this imbalance. We focus upon the proposed transformer-encoder pipeline, as it is apparent, based upon the outcomes of our experiments, that this architecture attains a higher accuracy when compared to the contending pipeline proposed, based on RNN with attention. In any form of supervised learning, there exists the possibility of overfitting, where a model fails to generalise upon unseen data, based on the training data provided [325, 326]. In this case, any supervised derivative of the UMAP algorithm would be susceptible to this phenomena. This is of particular importance when working with highly imbalanced data, such as the TREC-6 and TREC-50 datasets, where the sampling of a training set from the data, which entails only 6000 rows, would lead to an underrepresentation of many of the classes. To facilitate this, we evaluate the accuracy of the supervised transformer-encoder, and compare this to parametric UMAP supervised. While an analysis of the effects of all configurations of UMAP, and the other algorithms used in our preliminary analysis, would be beneficial, it does not fall within the main focus of our study, which is the investigation of the transformer-encoder. Therefore, we focus upon our comparison with the supervised parametric UMAP algorithm. This algorithm, in its default configuration, consists solely of fully-connected layers, thereby offering the most comparable algorithmic structure. The only point of divergence of our transformer-encoder proposal lies in the architecture of the neural network. In Figure 7.8, the average per-class accuracy is presented for the transformer-encoder, and the default configuration of parametric UMAP, for the TREC-6 dataset, showing the accuracy for each individual class across a varying output dimensionality. As with all other experiments in this study, this average is calculated based on 25 individual iterations of the dimensionality reduction and clustering pipeline.



(a) P-UMAP Supervised (b) P-UMAP Supervised Transformer

Figure 7.8: Per-class accuracy of k-Means clustering performed upon across all dimensionalities, for the TREC-6, dataset.

Based on an analysis of the accuracy of each class within the TREC-6 dataset, it is apparent that the transformer-encoder confers an improvement to the accuracy of some classes. At an output dimensionality of 3, when comparing the accuracy of class 1 between the transformer-encoder (SD = 0.024, M = 0.85), and default supervised parametric UMAP architecture (SD = 0.034, M = 0.84), we identify a p value of 0.02, indicating no significant difference between results (U = 187, z-score = 2.42, p = 0.02). For class 2, there is an improvement of 11.8% for the transformer-encoder approach (SD = 0.012, M = 0.81) compared to supervised parametric UMAP (SD = 0.045, M = 0.696), which demonstrates a significant increase ($U = 0$, z-score = 6.054, $p < 0.001$). Of particular interest regarding this comparison, is class 3, where the introduction of the transformer-encoder model demonstrates an increase in the average accuracy of this class, after k-Means clustering. For example, at an output dimensionality of 3, this improves from an average accuracy of 2.3% for parametric UMAP ($SD = 0.016$, $M = 0.023$), to 14.8% for the transformer-encoder ($SD = 0.075$, $M = 0.141$) derived model, a significant increase of 12.5% ($U = 102$, z-score $= 4.07$, $p < 0.001$). At an output dimensionality of 4, the transformer-encoder ($SD = 0.043$, $M = 0.19$) demonstrates a significantimprovement relative to parametric UMAP supervised ($SD = 0.016$, $M = 0.023$) of 16.6% ($U = 5$, z-score $= 5.95$, $p < 0.001$). This is a considerable improvement, given that class 3 of the TREC-6 dataset contains only 95 documents and represents the highest degree of imbalance. It is also worth noting that this underrepresented class also observes the greatest degree of change in accuracy as the dimensionality increases for the transformer-encoder, as denoted by the orange line in Figure 7.8. Returning our focus to an output dimensionality of 3, class 4 observes a decrease in the accuracy of transformer-encoder ($SD = 0.013$, $M = 0.928$) compared to supervised parametric UMAP ($SD = 0.026$, $M = 0.952$) of 2.3% ($U = 110$, z-score $= 3.92$, $p < 0.001$), which is significant based on the analysis using the WMW test. For class 5, the transformer-encoder confers a significant decrease in accuracy compared to supervised parametric UMAP of 3.3% ($U = 25$, z-score $= 5.56$, $p < 0.001$). Finally, for class 6, there is no significant difference observed based on the 25 individual experiments between the transformer-encoder, and supervised parametric UMAP ($U = 291$, z-score $= 0.4$, $p = 0.03$). Taking the assumption that an output dimensionality of 3 is the typical point at which the "knee" is observed in the UMAP algorithm in Figure 7.7, beyond which there are diminishing returns in the clustering result when UMAP derivatives are used for DR, we present the per-class accuracy for the transformer-encoder, and default configuration of supervised parametric UMAP at an output dimensionality of 3 in Figure 7.9. A full reference of the results at other dimensionalities is provided within the repository[1]. As there are too many classes in TREC-50 to analyse within this work succinctly, we sample some which we argue merit discussion. For instance, class 20, which is assigned to only 50 documents within the dataset, observes a significant improvement for the transformer encoder of 69% ($U = 0$, z-score $= 6.054$, $p < 0.001$), based on our average of 25 experiments. Similarly, class 41 also represents 50 documents, and it can be observed that the transformer-encoder leads to a significant increase of 16.5% ($U = 24.5$, z-score $= 5.578$, $p < 0.001$). In comparison, for class 5, the largest within the dataset, there



(a) TREC-6

(b) TREC-50

Figure 7.9: Per-class accuracy of k-Means clustering performed across all dimensionalities, for the TREC-6 and TREC-50 datasets.

is no significant difference between the accuracy of both models ($U = 311.5$, z-score $= 0.009$, $p < 0.95$).Based on the analyses of individual class accuracy, a hypothesis may be formed that *the introduction of the transformer-encoder architecture within parametric UMAP confers a degree of robustness to the sensitivity of imbalanced data when used in a supervised metric-learning methodology.*

7.6 Enhancing Cluster-Based Topic Models through Transformer-Encoder Facilitated Dimensionality Reduction through Additional Residual Connections

While the research conducted into the transformer-encoder, from the perspective of metric-learning, was beneficial in demonstrating the potential value of such an architecture to the field of DR and clustering, the outcomes of the research do not directly influence topic modelling, which has been a significant focus of previous chapters. Thus, in this section, a continued investigation is performed with a focus on applying the parametric DR pipeline with the transformer-encoder in directly contributing to enhancing cluster-based topic modelling with Top2Vec [6]. This is formed of two contributions. First, by considering transformer-encoder in a parametric DR pipeline, implemented into a modified version of Top2Vec. Secondly, by considering the adaptation of the architecture, by introducing additional residual connections. The results of these experiments indicate that the adoption of both architectures in a parametric dimensionality reduction pipeline enhances the topic modelling solution of a modified Top2Vec algorithm from both a metric and quality perspective. However, the introduction of additional residual connections leads to a significant improvement in the topic modelling solution when applied to a smaller dataset, which indicates the value that this novel paradigm of research can have upon the research field of topic modelling.

7.6.1 Defining the Transformer-Encoder with Additional Residual Connection

We propose the architecture of a *transformer-encoder with additional residual connection within parametric UMAP*, which is used in-place of the UMAP algorithm within Top2Vec. This is based on the original transformer-encoder [5], with a multi-head self-attention mechanism preceding a fully-connected feed-forward network.

After both the multi-head attention, and feed-forward layer, layer addition is performed, by combining the output of the previous layer with the initial input sequence. The process of layer addition, known as residual connection, tackles the vanishing gradient problem [17], where, during backpropagation, the multiplication of layer gradients can cause an exponential decrease if they are less than 1. It has been observed that with an increase in the sequence length of a model, there is typically a decrease in the magnitude of the gradient. This can lead to a slowdown or even a halt in the training process [18]. Given that the objective of the study is to enhance the DR process, and input sequences will therefore be large, it is essential to evaluate whether the introduction of residual connections will contribute to enhancing the process. Thus, we propose an additional instance of residual connection, wherein the original input sequence and transformer-encoder outputs are concatenated prior to a final feed-forward layer. The overall structure of the transformer-encoder with additional residual connection proposed in this work is represented in Figure 7.10.

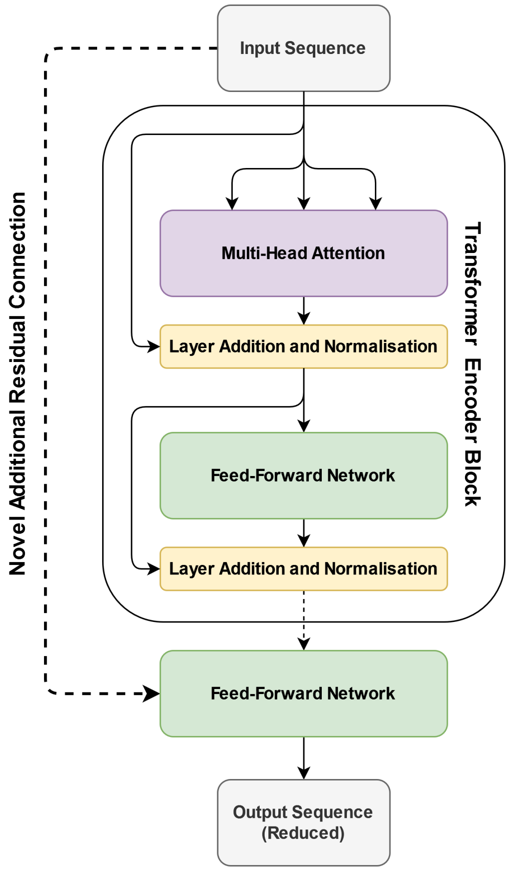

Input Sequence
Novel Additional Residual Connection
Multi-Head Attention
Layer Addition and Normalisation
Feed-Forward Network
Layer Addition and Normalisation
Transformer Encoder Block
Feed-Forward Network
Output Sequence (Reduced)

Figure 7.10: Transformer-encoder architecture with proposed additional residual connections.

To evaluate the proposed architectures for DR, we implement them in a modified version of Top2Vec [16] where we adjust the topic modelling pipeline by introducing parametric UMAP. The general flow of information for the topic modelling process can be summarised as follows: 1. The computation of document embeddings. 2. Application of the UMAP, or in our case, **parametric UMAP**, DR technique for the computation of a low-dimensional representation of the original document embeddings. 3. The clustering of low-dimensional embeddings via the HDBSCAN algorithm, identifying dense clusters of low-dimensional embeddings as documents with a common topic.

To ensure fair comparison between our proposed transformer-encoder approach to topic modelling, and classical UMAP within Top2Vec, we maintain identical hyperparameter configurations across experiments with regards to HDBSCAN configuration, ensuring that the only difference is the presence of the parametric UMAP algorithm with transformer-encoder, in-place of UMAP. In the experimental evaluation discussed in Section 7.6.4, we assess the effectiveness of the transformer-encoder with an additional residual connection, a transformer-encoder without the additional residual connection, the traditional Top2Vec algorithm, and a modified version of the Top2Vec algorithm employing parametric UMAP. In this modified version, a fully-connected network is used, which is the standard encoder network presented by parametric UMAP [4]. This ensures confirmation that the proposed modified transformer-encoder significantly influences the performance of topic modelling, rather than attributing the effects solely to the use of parametric UMAP.

7.6.2 Datasets

We select two datasets for evaluating the outcomes of our proposed transformer-encoder driven DR technique, for topic modelling, namely the 20 NewsGroups and BBC News datasets. These were selected due to their historic use in existing topic modelling research, which permits easy comparison with existing techniques such as Top2Vec. Moreover, these are easily accessible datasets that allow for reproduction of our work. The 20 NewsGroups dataset contains 16309 news articles across 20 categories [327], while the BBC News dataset contains contains significantly fewer, 2225 documents, taken from the BBC News website between 2004 and 2005 [114].

7.6.3 Evaluation Metrics

For evaluating the proposed models, we apply three widely-used metrics, which have been demonstrated in similar works relating to cluster-based topic modelling, and general topic modelling. These consist of normalised pointwise mutual information (C_{NPMI}) [328, 329], C_v coherence [157] and topic diversity (TD). C_{NPMI} is a coherence measure based on C_{UCI}; where C_{UCI} is based on a sliding window and the point-wise mutual information (PMI) of all word-pairs of the given top words (highest frequency words in a set). The word co-occurrence counts are derived using a sliding window with the size 10. For every word-pair, the PMI is calculated. The arithmetic mean of the PMI values is the result of this coherence [158]. In the case of C_{NPMI}, normalised pointwise mutual information (NPMI) is used in place of PMI during the calculation [329]. The resulting coherence measure has been demonstrated to emulate human judgement with reasonable performance [328], with the measure ranging from -1 to 1, where a score of 1 indicates a perfect association.

The C_v coherence measure is based on the co-occurrence counts for words within a topic, using a sliding window. These counts are subsequently used to calculate the NPMI of every topic word to every other topic word, thus, resulting in a set of vectors, one for every topic word. The one-set segmentation of the topic words leads to the calculation of the similarity between every topic-word vector and the sum of all topic-word vectors, with the cosine similarity of the two vectors applied for calculating similarity. Given this set of similarities, C_v can be calculated as the arithmetic mean of all similarities [157]. The C_v coherence metric ranges from 0 to 1, with a value of 1 representing perfectly coherent topics.

Finally, the metric of topic diversity (TD) can be defined as the percentage of unique words within a topic. A TD score close to 0 would indicate redundant topics, while a diversity close to 1 indicates highly varied topics [330].

It is worth considering that quantifying topic coherence measures and comparing them with human judgement of topic coherence is intrinsically difficult [177]. In the case of the C_v and C_{NPMI} measures, the measures will only take into account the top words per topic, and not the full topic-word distributions [177]. Furthermore, coherence measures do not consider the semantic similarity of words within topics. Thus, it is necessary to additionally evaluate the coherence of the resulting topics from the point of view of human judgement [331]. We perform this through an evaluation of the topic words, with respect to the subjects they entail and subsequent assignment of a label we feel best fits the subject of the topic-words.

7.6.4 Results

We evaluate the proposed *transformer-encoder with additional residual connection* with regards to C_v, C_{NPMI}, and TD over 10 experiments, presenting the average for each evaluation metric. Further to this, to assist in the discussion of the outcomes of the experiments, we record the average number of documents that are labelled as noise by the HDBSCAN algorithm [79] within Top2Vec, and the average number of topics generated by each topic model. The results of these are summarised in Table 7.3 for the 20 Newsgroups dataset, and Table 7.4 for the BBC news dataset. For each dataset, we evaluate (1) our proposed models, the adapted Top2Vec with transformer-encoder model (**T2V-P**$_{TE}$), and *transformer-encoder with additional residual connection* (**T2V-P**$_{TE\_RES}$), and compare with (2) a similarly configured model using only a 3-layer 100-neuron fully-connected neural network (**T2V-P**$_{Dense}$), which is the configuration of the encoder network presented in parametric UMAP [4, 332], as well as with (3) the standard Top2Vec algorithm (**T2V**), which uses the default nonparametric UMAP algorithm.

Table 7.3: Performance of the *proposed Top2Vec transformer-encoder pipelines* against state-of-the-art, based on topic evaluation metrics for the *20 Newsgroups* dataset; best values are in **bold**.

Model	Universal Sentence Encoder					MiniLM-L6				
	C_v	TD	C_{NPMI}	Noise	Topics	C_v	TD	C_{NPMI}	Noise	Topics
T2V	0.394	0.776	-0.194	5969.7	76.7	0.427	0.659	-0.149	7066.8	121.3
T2V-P$_{Dense}$	**0.404**	**0.869**	-0.191	4466.8	35.7	0.429	0.802	-0.146	5086	51.6
T2V-P$_{TE}$	0.399	0.723	**-0.171**	7086.1	82.3	0.434	**0.82**	**-0.137**	5041.2	54.9
T2V-P$_{TE\_RES}$	0.399	0.583	-0.243	2669.4	53.8	**0.443**	0.296	-0.159	2829.2	124.1

Table 7.4: Performance of the *proposed Top2Vec transformer-encoder pipelines* against state-of-the-art, based on topic evaluation metrics for the *BBC News* dataset; best values are in **bold**.

Model	Universal Sentence Encoder					MiniLM-L6				
	C_v	TD	C_{NPMI}	Noise	Topics	C_v	TD	C_{NPMI}	Noise	Topics
2V	0.438	**0.914**	-0.259	0	7	0.474	**0.848**	-0.195	111.6	12.2
2V-P$_{Dense}$	0.477	0.659	-0.207	328.2	23.1	0.519	0.765	-0.163	324.8	23.8
2V-P$_{TE}$	0.483	0.681	-0.198	271.1	23	0.521	0.783	-0.16	332.9	23.7
2V-P$_{TE\_RES}$	**0.485**	0.558	**-0.195**	312.6	32.9	**0.57**	0.758	**-0.123**	171.7	20.2

7.6.5 Discussion

Analysis of Evaluation Metrics

Based on the results presented in Tables 7.3 and 7.4, it is apparent that our proposed transformer-encoder architecture confers an effect upon the coherence of topics. When comparing differences between experiments hereafter, we adopt the Mann-Whitney U test to test for significance between results, based upon the sample of ten iterations of each experiment which were conducted. Given the sample size, and significance level of 0.05, the critical U value is 23, such that any U value greater than this indicates no significance between results. For each comparison, we provide the U value and p value.

With regard to C_v measure for the 20 Newsgroups dataset, the introduction of the transformer-encoder architecture (T2V-P$_{TE}$), indicates an improvement compared to T2V, of 0.005 (U = 29, p = 0.121) for USE embeddings, which we reject as significant. T2V-P$_{TE\_RES}$ indicates an improvement 0.016 (U=18, p=0.017) for MiniLM-L6 embeddings, when compared with classical Top2Vec. Similarly, for the BBC News dataset, an improvement of 0.047 (U=0, $p < 0.001$)is observed for T2V-P$_{TE\_RES}$ when using USE embeddings, and an improvement of 0.57 (U=0, $p < 0.001$) when using embeddings derived from MiniLM-L6 Consequently, incorporating additional residual connections into the transformer-encoder results in the highest C_v score in three of the four experiments.

From the perspective of C_{NPMI} measure, T2V-P$_{TE\_RES}$ confers the greatest improvement relative to classical Top2Vec, only for BBC News, where T2V-P$_{TE\_RES}$

Table 7.5: Top 10 words for the 10 largest topics for Top2Vec, trained using the Universal Sentence Encoder embedding model on the 20 Newsgroups dataset

ID	Topic Words	Human Label
1	recchi, potvin, nyi, nyr, lindros, lemieux, ahl, mets, defenseman, phillies	Sport - Ice Hockey
2	firearms, batf, guns, armed, firearm, atf, massacres, gun, homicides, militia	Firearms
3	diagnosed, symptoms, severe, diagnosis, vax, treatments, immune, candida, cure, chronic	Medicine
4	idiot, intrinsics, xlib, uucp, crap, rayshade, oops, fool, alleged, xfree	Computers - Programming/Utterances
5	email, mailing, sending, linked, addresses, regards, addressing, pm, im, mx	Computers - Email
6	spacecraft, jpl, orbiter, satellites, orbiting, satellite, astronaut, propulsion, nasa, orbital	Space Flight
7	xterm, implementations, args, printf, bitmap, compile, argv, fprintf, ansi, gnu	Computers - Programming
8	cryptography, encrypted, encryption, encrypt, cryptographic, decrypt, ciphers, pgp, plaintext, cipher	Computers - Cryptography
9	floppies, obo, dma, selling, maxtor, sgi, scsi, buyer, trade, atari	Computers - Storage
10	targa, mustang, honda, vw, bmw, ford, toyota, car, cars, engines	Automotive

achieves an improvement of 0.195 (U=0, $p < 0.001$) for USE embeddings, and 0.071 when for MiniLM-L6 embeddings (U=0, $p < 0.001$). For 20 Newsgroups, T2V-P$_{TE}$ indicates the greatest improvement of 0.021 compared to Top2Vec when using the USE embedding model (U=0, $p < 0.001$), and an improvement of 0.01 when using MiniLM-L6 embeddings (U=6, p=0.001).

Regarding topic diversity measure (TD), there is little consistency across experiments to indicate whether either T2V-P$_{TE\_RES}$, or T2V-P$_{TE}$ have a consistent effect upon the measure. In some cases, the standard Top2Vec algorithm expresses optimal topic diversity, for instance when applied to the BBC News dataset. Furthermore, the feed-forward parametric Top2Vec model (T2V-P$_{Dense}$) expresses the optimal topic diversity measure for the 20 Newsgroups dataset, when using USE embeddings. It is only when using MiniLM-L6 embeddings on the 20 Newsgroups datasets where T2V-P$_{TE}$ attains the greatest improvement in topic diversity. Notably, the transformer-encoder with additional residual connections (T2V-P$_{TE\_RES}$)

Table 7.6: Top 10 words for the 10 largest topics for Top2Vec using our proposed T2V-P$_{TE\_RES}$, trained using the Universal Sentence Encoder embedding model on the *20 Newsgroups* dataset.

ID	Topic Words	Human Label
1	theology, scriptures, scripture, verses, christians, believer, bible, beliefs, doctrines, biblical	Religion - Christianity
2	idiot, huh, fool, seriously, crap, sigh, xlib, oops, silly, trivial	Utterances
3	irq, processor, mhz, motherboard, risc, dma, scsi, processors, cmos, baud	Computers - Hardware
4	spacecraft, jpl, orbiter, satellites, orbiting, satellite, astronaut, propulsion, nasa, aerospace	Space Flight
5	xterm, implementations, args, printf, argv, fprintf, compile, bitmap, ansi, gnu	Computers - Programming
6	cryptography, encrypted, encryption, encrypt, cryptographic, decrypt, ciphers, pgp, plaintext, cipher	Computers - Cryptography
7	diagnosed, symptoms, severe, diagnosis, vax, treatments, chronic, lyme, candida, immune	Medicine
8	floppies, obo, dma, selling, sgi, maxtor, buyer, trade, scsi, need	Computers - Storage
9	waco, koresh, davidians, fbi, feds, atf, militia, firing, executed, suspect	Events - Waco Siege
10	pitchers, pitching, mets, alomar, clemens, inning, innings, phillies, recchi, batting	Sport - Baseball

presents the greatest decrease in topic diversity, which indicates that this has led to a greater number of topics which contain similar words. Hence, it is not possible to assert that the suggested method offers any advantage in terms of the diversity metric.

Evaluation of Topics

In comparing the ten largest topics produced by both Top2Vec, and the proposed T2V-P$_{TE\_RES}$, it is apparent that the introduction of the proposed pipeline has an effect on the general quality of the resulting topics. For brevity, we do not present topics for every experiment, however, we provide access to all experimental results and their associated topics in the provided repository[7].

Table 7.5 presents the topic words identified by the Top2Vec algorithm, when

[7] `https://github.com/ryanon4/Attention-Mechanisms-In-UMAP`

Table 7.7: Top 10 words for the topics produced by Top2Vec, trained using the Universal Sentence Encoder embedding model on the *BBC News* dataset

ID	Topic Words	Human Label
1	tories, ukip, lse, labour, economist, tory, reuters, imf, economic, gazprom	Politics
2	broadband, telecoms, consumers, technologies, consumer, aimed, electronic, spyware, launched, handsets	Technology
3	oscars, actress, nominated, oscar, awards, actor, famous, producers, films, critics	Oscar Awards
4	hodgson, premiership, gerrard, mourinho, celtic, downing, wenger, cardiff, glazer, newcastle	Sport - Football
5	roddick, federer, nadal, tennis, hewitt, tournament, henman, finals, final, murray	Sport - Tennis
6	olympics, olympic, athletics, marathon, competing, medal, doping, iaaf, athletes, compete	Sport - Olympics
7	doping, olympics, allegations, olympic, athletics, iaaf, athletes, suspended, accused, alleged	Sport - Olympics

applied to the 20 Newsgroups dataset using USE embeddings. Generally, the topics are of high quality, with the model even providing a separation of different computing topics (E.g. *Email*, *Terminals*, and *Cryptography*). However, topic 4, does not appear coherent and appears to consist of computing and programming terminology ("*xlib*" is a programming library[8], "*uucp*" is a unix command[9]), with general utterances (E.g. "*idiot*", "*fool*", "*oops*"). Comparing this with T2V-P$_{TE\_RES}$ (7.6), similar topics are identified, such as *cryptography*, *programming*, *computer storage*, and *medicine*.

When comparing Top2Vec with T2V-P$_{TE\_RES}$ for the BBC News dataset, it is apparent that the proposed pipeline provides a finer-grained breakdown of topics. For example, when comparing T2V with the proposed T2V-P$_{TE\_RES}$ pipeline applied to USE embeddings, it can be observed that more topics are produced (for brevity we only display the 10 largest topics, however for this example, T2V-P$_{TE\_RES}$ produced 21 topics in total). Furthermore, the topics are generally more fine-grained, with T2V-P$_{TE\_RES}$ providing additional topics on the subjects of *Finance - Investment*,

[8] https://en.wikipedia.org/wiki/Xlib

[9] https://en.wikipedia.org/wiki/UUCP

Table 7.8: Top 10 words for the 10 Largest Topics for Top2Vec using our proposed T2V-P_{TE_RES} trained using the Universal Sentence Encoder embedding model on the BBC News Dataset.

ID	Topic Words	Human Label
1	tories, ukip, labour, tory, parliamentary, mps, ministers, conservative, politicians, blair	UK Politics
2	oscars, actress, nominated, oscar, awards, actor, famous, producers, films, critics	Entertainment - Oscars Awards
3	mourinho, premiership, wenger, glazer, arsenal, gerrard, liverpool, chelsea, hodgson, downing	Sport - Football
4	hodgson, rugby, wales, cardiff, premiership, scotland, welsh, gerrard, clarke, england	Sport - Rugby
5	economist, economy, economic, analysts, deficit, exports, imf, inflation, lse, unemployment	Economics
6	lse, gazprom, reuters, shareholders, earnings, profits, stock, shares, profit, demand	Finance - Investment
7	technologies, technology, gadget, electronic, devices, gadgets, handheld, electronics, wireless, billion	Technology
8	roddick, federer, nadal, tennis, hewitt, tournament, henman, finals, final, murray	Sport - Tennis
9	fraud, allegations, lawsuit, investigation, bankruptcy, alleged, analysts, attorney, executives, accused	Finance - Fraud
10	spyware, viruses, virus, malicious, secure, spam, security, attacks, broadband, fraud	Cybersecurity

Finance - Fraud, *Economics* and *Cybersecurity*, which would otherwise not have been identified within the corpus.

Similarly, when applied to MiniLM-L6 embeddings, topics are identified by T2V-P_{TE_RES} which were not produced by standard T2V. For example, topics related to *Entertainment - Music* and a technology topic specific to *communications and telecommunications*. Again, this fine-grained topic result would not have been obtained without the introduction of the transformer-encoder pipeline with additional residual connections.

Table 7.9: Top 10 words for the topics produced by Top2Vec, trained using the MiniLM-L6 embedding model on the BBC News dataset

ID	Topic Words	Human Label
1	ukip, tory, banking, labour, reuters, finance, firms, mps, economy, financial	UK Politics
2	viewers, media, microsoft, multimedia, piracy, digital, entertainment, devices, tv, television	Entertainment
3	mourinho, premiership, hodgson, liverpool, gerrard, owen, rugby, tackle, arsenal, wenger	Sport - Football
4	iaaf, competing, olympic, olympics, marathon, athletics, compete, athletes, doping, running	Sport - Olympics
5	tennis, federer, nadal, competing, tournament, hewitt, roddick, matches, doping, match	Sport - Tennis
6	iaaf, doping, olympic, athletes, olympics, athletics, accused, competing, tennis, suspended	Sport - Olympics

7.7 Discussion on the Implications of Parametric Dimensionality Reduction for Clustering and Topic Modelling

This work has investigated how two neural network architectures can affect the accuracy of downstream clustering tasks, when used in a parametric UMAP dimensionality reduction pipeline, namely the P-UMAP Supervised RNN with Attention, and the P-UMAP Supervised Transformer. Our analysis highlights several interesting findings. Firstly, we provide an empirical investigation into the effects of "traditional" dimensionality reduction algorithms PCA, LiDA, t-SNE, and UMAP upon downstream k-Means clustering upon four benchmark datasets. Through our evaluation of these traditional algorithms across a range of dimensionalities, we demonstrate the effectiveness of each with respect to the output dimensionality, as well as discussing the benefits and disadvantages of each technique in relation to computational complexity.

In our subsequent analysis of UMAP, Parametric UMAP, and their supervised-learning alternatives, we empirically identify a consistently observable "knee" curve in relation to the accuracy of k-Means clustering upon the low-dimensional repre-

Table 7.10: Top 10 words for the 10 largest topics for Top2Vec using our proposed T2V-P$_{TE\_RES}$, trained using the MiniLM-L6 embedding model on the BBC News dataset

ID	Topic Words	Human Label
1	tory, ukip, blair, mps, minister, ministers, parliamentary, tories, politicians, labour	UK Politics
2	economy, inflation, economic, currency, economist, deficit, exports, income, treasury, markets	Economics
3	mourinho, premiership, liverpool, chelsea, arsenal, gerrard, wenger, striker, hodgson, madrid	Sport - Football
4	oscar, oscars, cast, cinema, films, actor, film, movies, movie, awards	Entertainment - Oscars Awards
5	rugby, hodgson, tackle, owen, welsh, referee, premiership, wales, competing, cardiff	Sport - Rugby
6	takeover, firms, stock, shareholders, shares, investors, market, profits, profit, earnings	Finance - Investment
7	concert, singer, songs, band, album, music, song, rock, robbie, tour	Entertainment - Music
8	tennis, federer, nadal, competing, tournament, hewitt, roddick, matches, doping, match	Sport - Tennis
9	mobiles, mobile, phones, telecoms, phone, devices, handsets, broadband, telephone, multimedia	Technology - Communications
10	fraud, bankruptcy, lawsuit, firms, shareholders, claims, banking, takeover, securities, executives	Finance - Fraud

sentations produced by variations of the UMAP algorithm. This finding merits a deeper discussion, particularly as it may be beneficial for researchers or industry who seek to perform any clustering of embeddings, as we have demonstrated that there is no benefit from opting for a large output dimensionality when using the UMAP algorithm as a preprocessing step prior to downstream clustering, where it is evident that an output dimensionality of 2 or 3 is suitable. There are several considerations of this to take into account. First, it is evident that dimensionality reduction of text embeddings has a benefit upon k-Means clustering, where many UMAP derivatives outperform the k-Means baseline, as denoted in Figure 7.5.8. As discussed in Section 2.1.4, it becomes difficult to perform distance calculations on highly dimensional vectors, and in k-Means, the distance measure used in determining clusters is commonly euclidean distance. Thus, it is understandable why the introduction of dimensionality reduction prior to clustering enhances the overall accuracy of k-Means by enabling a greater diversity in distance calculations. However, observation of the "knee" in UMAP variants may indicate several potential important considerations.

Firstly, from an algorithmic perspective, UMAP, and its closest algorithmic comparison t-SNE, approach dimensionality reduction in a different manner. UMAP has been shown to more effectively preserve the global structure of data in comparison to t-SNE [96], however this has been reported to be due to the random initialisation performed in t-SNE, whereas by default UMAP applies Laplacian Eigenmaps [333] during initialisation of the embedding [98], and by swapping these initialisation methods, t-SNE was found to better capture global structure. In our experiments using UMAP, the default initialisation using Laplacian Eigenmaps was performed, and it may be for this reason that the "knee" is observed in UMAP derivatives but not in t-SNE if this is an artefact of the initialisation process.

Secondly, while not peer-reviewed, and thus open to interpretation, the UMAP documentation[10] highlights several caveats of using UMAP for clustering that are also claimed to be apparent in t-SNE, where both algorithms are susceptible to not completely preserving the density of the original data during dimensionality reduction, and also may create false tears in clusters, which can result in a finer clustering solution than what is present in the original data. Both of these claims would understandably influence the clustering solution with k-Means, and therefore influence the accuracy reported, particularly if these phenomena are more apparent in lower output dimensionalities.

Thirdly, it is crucial to acknowledge that at a dimensionality of 1, both UMAP and k-Means may struggle to effectively partition the dataset. It is likely that a considerable amount of information is lost during the reduction to a single dimension, wherein the compression of global and local features presents a significant challenge. Furthermore, a single dimension is inherently limited in its capacity to represent complex relationships among data points, which could result in an overlap of points and a consequential loss of meaningful cluster separation. This phenomenon po-

[10] `https://umap-learn.readthedocs.io/en/latest/clustering.html`

tentially accounts for the pronounced increase in clustering accuracy observed from a dimensionality of 1 to 2. However, since this trend is not evident in the other dimensionality reduction algorithms analysed, we argue that this characteristic is exclusive to UMAP.

Finally, it is imperative to consider the method employed in generating embeddings for the experiments. High-dimensional embeddings were generated using RoBERTa [31], and it is plausible that this embedding technique may have exerted a direct influence on dimensionality reduction and, consequently, clustering outcomes. In this context, the creation of embeddings can be conceptualised as an information bottleneck, where any subsequent data transformation is constrained by the quality of information encoded in the original embeddings. Consequently, it is worth considering whether the manifestation of the "knee" in UMAP clustering accuracy is affected by the use of various embedding techniques. Additionally, the potential influence of UMAP's hyperparameters on this outcome should also be contemplated. One such hyperparameter to consider is the impact of varying the number of neighbors considered during the approximation of the local neighborhood graph, or the distance measure used during high-dimensional graph construction, which are, by default, 15 neighbours, and Euclidean distance in our experiments.

Through an investigation across a range of different dimensionalities, we have identified that attention mechanisms can have a significant effect upon clustering accuracy, when used in a metric learning framework within parametric UMAP. Moreover, our second proposed architecture, entailing a transformer-encoder, achieved the best overall improvement across three of the four datasets used in our investigation. Through a visual analysis of the lower-dimensionality embeddings produced by the transformer-encoder, we are able to demonstrate the effectiveness of the transformer-encoder pipeline for downstream clustering for a parametric UMAP dimensionality reduction pipeline when used for metric-learning, when compared to several other architectures. Additionally, we demonstrate that the transformer-encoder ensures an improvement in accuracy, while also maintaining significantly fewer trainable model parameters compared to an RNN, which extends our findings from investigating **(RQ3.4)**.

Through an analysis of the accuracy of individual classes within a dataset, it is apparent that the transformer-encoder architecture confers a benefit when faced with imbalanced data. More specifically, underrepresented classes have been found to benefit from the introduction of the transformer-encoder. These findings open novel avenues of research, particularly in relation to the individual components of the architecture, which confer the robustness of the model in handling underrepresented data.

Our research has demonstrated the feasibility of *our proposed pipeline, employing a transformer-encoder within a Parametric UMAP metric learning framework*. However, there are still unresolved issues and unanswered questions which may affect researchers and real-world applications. Firstly, the adoption of the transformer-encoder into a parametric dimensionality reduction pipeline will incur an effect upon the training time and computing requirements of the DR process. While the training time of the transformer-encoders for parametric DR was not recorded or disclosed in this research, and therefore presents a potential avenue of future investigation, it is worth considering comparative studies of similar architectures. With regard to transformers during pre-training via masked-language-modelling, it has been reported that model depth has a significant effect upon the time required for training [334]. A similar study reinforces this, highlighting that increasing the depth of transformers significantly increases computational and memory requirements as the number of parameters increases [335]. In light of these implications, it is worth considering whether the computational demands associated with training a transformer-encoder-based pipeline for dimensionality reduction outweigh the potential advantages in terms of accuracy. Although the low-dimensional representations produced by the proposed UMAP architecture can contribute to a reduction in the time required for clustering, it is worth considering that a larger number of trainable parameters of the architecture confers a longer time required for preprocessing. However, once the model has been trained, the cost of inference is actually cheaper and faster for parametric UMAP than it would be for the conventional UMAP, as reported by [4]. This initial computational cost could, however, be mitigated through optimisation strategies such as adopting GPU training of the neural network architecture. Furthermore, our study has focused upon a metric-learning methodology, wherein we have used a small portion of labelled data to contribute to the dimensionality reduction process. For many real-world applications of our methodology, labelled data may not be available. Therefore, it would be beneficial for future works to investigate how the architecture proposed would perform when applied in an unsupervised manner. Additionally, we have conducted our study with a focus upon the downstream clustering of low-dimensionality vectors by the k-Means algorithm. While this clustering technique is widely known, the algorithm requires the prior specification of a known number of clusters, and as such is not suitable for applications, where the number of clusters is not known. Consequently, it remains to be determined if the transformer-encoder architecture can be utilized for calculating low-dimensional representations and subsequently subjected to alternative clustering algorithms. Among these, the potential enhancement of results through the application of density-based clustering methods such as HDBSCAN [79, 166] or DBSCAN [336], which have been demonstrated in tandem with UMAP for topic modelling, is of particular interest [16].

Aside from our contribution of showing how to apply a transformer-encoder to parametric UMAP, we have provided further experimentation into the outcomes of using various existing dimensionality reduction algorithms to contribute to improvements to clustering accuracy. Most notably, we demonstrate how the combination of the transformer-encoder with metric learning, when using parametric UMAP, can provide significant improvements to the clustering solution.

After outlining the potential value of architectural design in parametric dimensionality reduction, we have presented a modified approach to cluster-based topic modelling through the implementation of a transformer-encoder based parametric DR pipeline via the parametric UMAP algorithm, where additional residual connections were introduced to further address the vanishing gradient problem, therein addressing RQ3.5 from an unsupervised learning perspective, and RQ3.6 by the introduction of further residual connections. This novel paradigm of DR research is relatively unexplored within the literature, and we argue that further improvements in downstream tasks such as the topic modelling focus of this work could be obtained through continued investigation of the effects of network architecture on parametric DR. Through a comparative analysis of the performance of the proposed pipeline with respect to topic evaluation metrics of coherence and topic diversity, it is possible to demonstrate the significance that the proposed architectures have upon the topic modelling solution of the Top2Vec algorithm, where the proposed parametric DR with modified transformer-encoder pipeline achieves an improvement with respect to coherence measures in 3 of the 4 experiments conducted. However, it is also evident that this has had a negative influence on the topic diversity measure. Although evaluating with respect to coherence measures is valid, we further investigated the topics produced from the perspective of human judgement, where we identified that on the smaller BBC News dataset, the proposed pipeline achieves a finer-grained breakdown of topics, which greatly improves its value in the topic modelling of smaller datasets.

An interesting observation of the results of the experiments in both Section 7.5 and Section 7.6, is the apparent considerable improvement that transformer-encoders provide in parametric DR on small datasets. In Section 7.5, from a metric-learning perspective, the transformer-encoder provides the greatest improvement on the small datasets of TREC-6 and TREC-50, significantly outperforming an RNN configured with attention mechanisms. In contrast, on the significantly larger AG News dataset, there is a significantly smaller difference in downstream accuracy, with the nonparametric supervised UMAP algorithm attaining the greatest downstream clusering accuracy. This is reflected in Section 7.6, where it is clear that the introduction of the transformer-encoder has a significant beneficial effect upon both the metric, and quality of the topics produced for the BBC News dataset, which is a small dataset. Based on these evaluations, it can be considered that transformer-encoders for parametric DR appear to generalise well to small datasets, where other models would require a significantly larger dataset to achieve the same quality results.

7.8 Epilogue

This chapter has focused upon contributing to the novel paradigm of parametric dimensionality reduction in order to enhance downstream clustering performance. Generally, based on the results of the experiments, it is apparent that the introduction of the transformer-encoder, and subsequent modified transformer-encoder through additional residual connections has had a considerable positive influence in the quality of clustering solutions from both a metric-learning and fully unsupervised perspective. This satisfies the research questions set forth, and opens up the potential to improve the academic literature analysis framework set forth in Chapter 6 by improving the quality of the topics found within literature. However, open questions, particularly in regards to alternative architectural designs, remain. Furthermore, it is apparent that the proposed parametric DR pipeline performs better on small datasets, however it is also apparent that on larger datasets, nonparametric UMAP remains comparable. In industry, where datasets are typically large, the benefits of parametric DR may not be demonstrable, and thus continued research is necessary in order to assure that direct contributions can be made to industrial processes. In the next chapter, an in-depth discussion on the overall findings of the research enclosed within this thesis is provided, as well as the limitations, and potential future extensions.

CHAPTER 8

Discussion

This Chapter addresses the general outcomes of the works detailed in Chapters 4, 5, 6, and 7 from the perspective of the research questions that were addressed in the respective chapters. This can be generally summarised into the concepts of social media meta-embeddings for information retrieval (Chapter 4), topic modelling of academic literature - as an additional domain to highlight the approach of automatic processing of articles, outside journalism (Chapters 5 and 6), and parametric dimensionality reduction for enhancing clustering performance - a technical contribution of greater generality, applicable also directly to the final product for the industry sponsor (Chapter 7). After summarising the main points of contribution for each chapter, the limitations and potential areas of expansion and future works are set forth, with regard to each main aspect of the thesis.

8.1 Construction of Meta-Embeddings to Improve Journalist Retrieval Based on Full-Text Queries

In Chapter 4, the first general research question GRQ1 was addressed, in which it was proposed to investigate the retrieval and recommendation of relevant social media influencers based on full-text article queries, in order to directly address the time-consuming nature of such tasks for key players in the media and publishing industry. This was a unique industry problem, which was expressed by the industrial sponsor of the Ph.D. project, to address the key needs of the publishing industry. This required the formulation of a suitable method to address the unique needs of publishers for the identification of relevant social media accounts based on the context of a given article of journalistic content. A notable outcome of this aspect of research is the subsequent **commercial implementation**, where a revised system based on the research is frequently used by customers within the publishing industry for the originally intended purpose. For this aspect of research, the following umbrella research question was addressed:

GRQ1: How can journalistic influencers be retrieved with accuracy for a publishing enterprise?

When approaching the task of social media account recommendation based on full-text querying of journalistic content, the notable obstacle of a lack of prior existing research, and subsequently, lack of existing evaluation data in the niche domain incurred limitations to the approaches that could be adopted. Specifically, the '*cold-start*' problem of collaborative filtering recommender systems, wherein they often experience poor recommendation accuracy when first instantiated, and require appropriate time to 'warm up' before recommendations become accurate [15], is dealt with here. By adopting an information retrieval approach to recommendation, based on the semantic similarity comparison of query and candidate embeddings, it was possible to overcome this challenge, which ensured accurate retrieval of results, by eliminating the need for training a recommender system. Still, a brute-force comparison of semantic similarity in an ever-growing industry dataset of social media accounts is impracticable and certainly not scalable. Hence, the adoption of approximate-nearest-neighbour searches through HNSW algorithm was considered to mitigate this limitation and ensure that an industrial implementation could account for growth in the dataset, with this being the contribution made by addressing RQ1.1. The contributions thus far with regard to this RQ are of applied nature, using existing techniques in a novel domain.

Importantly, the use of HNSW for conducting a semantic search to find relevant social media profiles using journalistic queries was found to offer a satisfactory level of recommendation accuracy in terms of Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) metrics, as **assessed by industry expert evaluators for a viable commercial product**. However, the emphasis of RQ1.2 was on investigating potential improvements in retrieval accuracy. For this, it was clearly demonstrable that the adoption of meta-embeddings constructed from multiple data sources, in this case, social media bios and posted content, enhanced the top-5 retrieval accuracy, as evaluated by these domain experts. The findings from this research enabled the identification of an optimally configured meta-embedding, through a weighting of 0.5, which resulted in achieving an MRR of 0.81 and MAP of 0.72, respectively, for the top-5 suggested accounts that were highly relevant, based on expert labels. When considering recommendations that were both strongly and weakly relevant, these values improved to 0.93 for MAP and 0.86 for MRR. This demonstrated an improvement over both the state of the art (at time of publishing) embedding model SGPT in terms of top-5 recommendation accuracy, and the accuracy of recommendations, when embeddings were derived from either tweet, or user biography sources, independently.

The findings of this research have clear implications to the media and publishing industry, particularly in regard to the industry implementation of the recommender system, which has already seen commercial use since December 2022. Moreover, for the wider research domain, the concept of meta-embedding for information retrieval tasks shows promise in contributing to enhancing retrieval accuracy for a range of tasks, where multiple data sources are available. With the advent of transfer-learning, the size of pre-trained language models is continually growing, along with the requirements of greater compute resources that comes with continual improvements in the quality of embeddings produced by such pre-trained models. Similarly, recently proposed models in research are typically released 'as-is', with optimised or 'lite' versions of such models often being released after a long period of time (e.g. ALBERT [337] was released 1 year after BERT [40]). This presents a limitation to industry, due to the costs incurred during inference of embeddings when using large models, particularly if these are bleeding-edge models with little optimisation, which is often not the first consideration during NLP research. Thus, it is worth considering that the construction of meta-embeddings in a manner similar to that proposed in Chapter 4 can mitigate the costly nature of using large emebdding models, through leveraging smaller, cheaper models, while also enabling an improvement in the retrieval accuracy. This objective could be investigated via two primary avenues which deserve exploration in future research. Firstly, as explained in Chapter 4, the utilisation of meta-embeddings derived from dual textual data sources has demonstrated advantages for text retrieval tasks. Consequently, it is worth considering whether this binary meta-embedding approach could be further refined by incorporating additional data sources. This would require investigation as to whether the integration of supplementary sources in meta-embedding construction could introduce noise, thereby diminishing retrieval accuracy. Secondly, there is a potential to reorient the focus towards investigating the feasibility of integrating diverse embedding models within an ensemble framework to produce a unified meta-embedding. This approach could be particularly beneficial in harnessing the optimal characteristics of disparate models, such as the amalgamation of a domain-specific model with one trained on a broader corpus. In undertaking this line of inquiry, consideration must be made concerning disparities in the dimensionalities of embeddings produced by different techniques, which would complicate the straightforward application of the weighted averaging strategy as discussed in Chapter 4.

Another contribution to the research presented in Chapter 4 was through the evaluation of forward time-decay functions in the ranking of retrieval results, which was investigated through RQ1.3. When evaluating the literature, it was apparent that **temporal characteristics played a significant role in information retrieval and recommender systems** [137], where temporally-aware methods had shown competitive performance compared to collaborative recommenders. These systems often utilised a time-decay function to assign weights during result ranking, typically employing a '*backwards*' function that considers the age of an item relative to the current time. In the dynamic data streaming environments of the commercial implementation of the research presented in Chapter 4, this would incur the need to recalculate the weights of the elements, as their age changes over time. Alternatively, a forward time decay model was identified in [117], which measured decay from a specific time point and was demonstrated to model existing decay functions in a manner which was scaleable in a data streaming environment. Thus, a monomial forward decay function introduced by [117] was selected as a weighting strategy to rank reommendations based on the recency of their posted content.

After evaluating across a range of weightings with regard to the importance of temporal relevance in recommendations, it was identified that the introduction of the time-decay ranking function presented an improvement in the recency of the recommended results; however, conferred a small decrease in the accuracy of the recommendations. A slight reduction in accuracy is often acceptable when prioritising up-to-date results, particularly given that in the publishing sector, it is crucial to target social media users who have recently engaged in a topic, to ensure that the content is perceived as valuable to them. At the same time, high accuracy relating to dormant users is irrelevant.

8.2 Topic Modelling of Academic Literature to Enhance the Discovery of New Knowledge

In Chapters 5 and 6, an in-depth investigation was conducted into another aspect of unsupervised learning methodologies, with the aim of **contributing to the publishing industry through assisting in the comprehension of large volumes of literature**. As discussed in Section 1.5, there is a continuous production of textual media in the publishing industry, requiring industry leaders to stay up-to-date on trends and connections within various subjects. This led to addressing the second main research question, which considered the use of topic modelling as a tool to aid in understanding large amounts of unstructured text. Here, a dual approach was taken, starting with an initial exploratory phase, as described in Chapter 5, followed by a more detailed experimental investigation, presented in the final solution for handling large volumes of academic literature, in Chapter 6. Due to limitations on sharing Intellectual Property held by the industry sponsor, a case-study methodology was necessary to protect sensitive industry information. Thus, a focus on the topic modelling of academic literature, was devised. This investigation aimed to enhance the comprehension of literature through the topic modelling of academic content, and subsquent functionality to assist in making sense of topics within the literature. The outcomes of this research can be applied to the publishing sector, by either concentrating on journalism based on academic literature, or by analysing the abundant textual data sources available, such as news articles, magazines, and social media sources. The general research question for this aspect of the project was set forth as:

GRQ2: What AI techniques can be defined to assist readers in the comprehension of literature?

GRQ2 presented an overarching objective of investigating how AI techniques can be applied to the task of assisting in literature analysis, and was initiated through an exploration of existing topic modelling algorithms and their value to identifying topics in academic literature. In addressing RQ2.1, we initially explored the classical Bayesian topic modelling algorithm LDA as a method of making sense of academic literature. This way it was possible to identify topics within a larger cross-section of literature than would be possible via a manual literature survey; the solution was based on the framework presented by [156]. Notably, using LDA incurs the need for an exhaustive search of hyperparameters to identify an optimal topic modelling solution, with manual intervention necessary in the curation of the 201 topics, before the definition of 7 topics that were considered relevant to the original objective of the literature analysis. This small number of topics identified out of the total presented by LDA topic modelling indicates an innate issue with the algorithmic approach, where manual intervention by a human reviewer was necessary at multiple stages, such as defining the number of topics to identify, and the subsequent curation of the large number of topics identified into a meaningful selection of useful topics.

In order to address these limitations of LDA, Section 5.3 proposed applying an alternative algorithm, based on a separate case-study of literature regarding intelligent tutoring systems research. This answers RQ2.1 more effectively than the initial LDA exploration, through the Top2Vec [6] algorithm, which adopts a cluster-based approach to the identification of topics in literature. This eliminates the need for a thorough assessment of model coherence across a range of topic numbers, a requirement when utilising LDA, which makes the approach significantly more suited to the overarching objective of the provision of an automated literature analysis framework. Furthermore, it was possible to take advantage of contextual information of text through neural embedding models, detailed in Section 2.1 when using Top2Vec, which would have been ignored in the bag-of-words style of topic analysis which was facilitated by LDA.

Both aspects of the research so far demonstrate the effectiveness of topic modelling algorithms as an AI tool to assist in the comprehension of literature through the identification of meaningful topics relevant to the literature analysis goal, thus addressing RQ2.1. Specifically, both algorithms were successful in identifying useful topics for the literature analysis goal; however, the Top2Vec-based approach presents the best solution, addressing several of the limitations encountered when using LDA which were observed in the existing literature on using topic modelling for literature analysis [156]. From the perspective of the media and publishing industry, the techniques adopted can be easily interchanged with different data sources, such as news media and social media data. For example, the identification of distinct topics of discussion on social media platforms can be an effective method for journalists and publishers to keep up-to-date with recent trends in public opinion, aiding with the consideration of prompts on particular topics, thus improving the volume of reads on news media produced on these subjects.

RQ2.2 is addressed in several stages throughout this thesis, beginning with Chapter 5, Section 5.4.2, where a method for the analysis of semantic relationships between topics is first established, based on a comparison of the semantic similarity of vectors derived from topics. This method of visualising was found to present relationships between literature topics in intelligent tutoring systems research, which were evaluated in terms of their relevance to what could be identified in the literature. In Chapter 6, this is further developed through an interactive visual tool based on topics selected by experts, from, ambitiously, a completely new domain - the biomedical domain, specifically relevant to their case-study research into the human kinome.

Based on the outcomes of this case-study, it is apparent that the mapping of semantic relationships between the topics identifies meaningful relationships via our **automated visualisation tool**, which contributes to literature understanding by permitting researchers to identify potential subject areas to consider in their analysis without the need for exhaustive prior manual literature analysis. Building upon this, in Chapter 6, Section 6.5 an expansion of the semantic relationship visualisation tool is presented, which enables the further functionality for literature analysis through the addition of comparisons of semantic relationships with the keywords adopted during the automated literature search. Based on an experimental **evaluation by** 18 **medical experts and medical students**, it is identified that the proposed semantic mapping functionality presents value to the literature analysis with respect to the general usefulness of the functionality from both quantitative and qualitative evaluation perspectives. However, it was also identified that the general usability of the semantic mapping functionality introduced a degree of cognitive load, which negatively influences the effectiveness of the functionality for literature analysis. The construction of semantic relationships between topics is thus an effective method for *surface-level analysis* of topics, which is valuable to any individual who is exploring a subject in which they may have limited prior knowledge. A worthy consideration of expanding the concept of semantic mapping of topics is touched upon through the inclusion of individual search keywords in Section 6.5. One way in which this could be enhanced would be through the identification of distinct named entities within the literature of a topic, and fine-grained breakdown of the semantic relationships of this entity with relation to the publications that discuss it. This would, in turn, enable the construction of a significantly finer-grained relationship map, which would enable researchers to consider specific aspects of a topic, in relation to other semantic entities. In constructing such a semantic map, it would be possible to establish relationships hierarchically, similar to a hierarchical ontology [338], with the most abstract relationships presenting the overall semantic links between topics, and detailed relationships would be provided deeper in this hierarchy.

From the perspective of the RQ2.3, Chapter 5, Section 5.2.2 presented a temporal analysis from the perspective of neural network architectures, social networks, and dialectal techniques identified during topic modelling. This temporal analysis consisted of tracking the frequencies of topic words and assessing the frequencies of these based on the year of publication of the papers within the corpus. The analysis identified several trends, such as the consistent volume of research on Bayesian analysis present in ASA research, something supported by the manual survey accessible at [154]. Generally, the temporal analysis discussed in this section of the thesis is shown to support the findings of the manual study, and demonstrates the effectiveness that visualisation of temporal trends in data can have in assisting literature analysis. Similarly, in Chapter 5, Section 5.4.2, an alternative approach to temporal analysis was performed. This differs from that of Section 5.2.2 by the focus on topics, which are dense clusters of documents rather than the frequencies of individual words. By visualising the changes in these topics over time, it was possible to present an **overview of the temporal changes of entire semantic topics, rather than individual words**. Based on the analysis of the topics identified in a case-study of intelligent tutoring system research, which was presented in Section 5.4.2, it was possible to identify general themes and trends, such as the growth of research into MOOCs in recent years, or a decline in research related to adaptive hypermedia and intelligent dialogue systems. Overall, the contributions identified when analysing temporal trends in academic topics present the **opportunity for automatic support in the discovery of new knowledge when approaching a new research task**, through identifying potentially under explored areas of research, outside of the trending topics.

As a next step, the integration of information from various external data sources in Section 6.4 was aimed at addressing RQ2.4. This process facilitates the expert annotation and understanding of topics discussed in the literature. Triangulation was achieved through an internal relational database that established links to well-known protein databases, such as the UniProt Knowledgebase (UniProtKB) [277] and InterPro [278]. Although this case study specifically focused on the literature on protein kinases, the robust methodology employed in collaboration with bioscientists has the potential for broader applications and can be easily adapted for analysing the literature in different domains, as we demonstrate by application to several areas.

The main contributions obtained when addressing RQ2.5 are set forth in Chapter 6, where an early iteration of a framework for general literature analysis was evaluated, based on a case-study of the human kinome in Section 6.4. The introduction of a new framework for examining literature topics demonstrated the consideration of semantic connections and temporal patterns of the topics, while allowing for the involvement of domain experts in labelling and refining the topics found in the literature through crowdsourcing [276]. Experts like educational instructors or experienced researchers can contribute their knowledge to label the identified topics. As a result, learners in the field can utilise the visualisation tools offered to enhance their understanding of the subject. To improve knowledge discovery, external data sources were triangulated to extract and link named kinases mentioned in the literature. This process aids in expert labelling and understanding of topics within the literature. In evaluating the usefulness of the proposed framework, the case-study analysed the topic modelling results, and the subsequent semantic mapping visualisation with respect to the domain knowledge of the supporting experts. As detailed in Section 6.4.2, the topics identified within the literature by the proposed framework are coherent with regards to the domain expert's knowledge. Furthermore, the semantic relationships were explained from the knowledge of the domain expert, indicating their effectiveness in the identification of relationships between literature topics.

While the findings of the framework presented in Section 6.4 demonstrated the value of a literature analysis framework, based on cluster-based topic modelling, the scale of the study was small and focused only on a case study of literature related to the human kinome. Thus, it was necessary to expand the scope of the study through a revised framework proposed with the aim of contributing to a generally applicable framework for assisting in literature analysis. This was presented in the work titled *A co-designed Framework for the Analysis of Large Volumes of Biomedical Literature*, presented in Section 6.5. Through a co-design process, the findings of the previous chapters were revised based on the requirements set forth by a medical expert, resulting in in a framework which permits the analysis of medical literature form the PubMed platform based on user-devised queries. However, it was necessary to demonstrate that the framework would be beneficial in the discovery of new knowledge. The previous works presented in Chapters 5 and 6 had been evaluated through case-studies in distinct domains, which demonstrate the effectiveness of the framework. Further work was necessary to analyse how it facilitates knowledge discovery for a wider range of literature tasks.

Thus, RQ2.6 was proposed, which focused upon an evaluation by medical experts. The framework, which was shown to be applicable to any type of literature, search based on feedback from 18 medical experts, was evaluated with respect to the concepts of cognitive load, technology acceptance, and general satisfaction, as well as based on the general relevance of the topic modelling solution and qualitative reports of the participants, with the full results of these being presented in Section 6.8. Based on participant feedback, it was identified that the proposed framework had a positive effect upon the discovery of new knowledge for medical experts. With regard to cognitive load, results suggested that the literature analysis activities during the experiment brought a moderate mental load to the users. By examining participants' perspectives with respect to technology acceptance, it was found that the majority quickly understood how the framework operates and recognised the benefits of SMARTEN. Finally, the assessment of overall satisfaction indicated that a large number of participants agreed that the framework enhanced their comprehension of research subjects, facilitated the discovery of new information in academic sources, and expressed interest in using the framework for their upcoming projects. Additionally, they were inclined to suggest the SMARTEN system to their peers. From the perspective of qualitative feedback from participants, it was identified that the SMARTEN system presented opportunities for the discovery of new knowledge, through encountering unexpected related terms during the literature search; however, participants additionally specified that the general usability of the system, particularly the labelling of topics, was unsuitable.

Generally, by conducting an experiment with medical professionals in the proposed literature analysis framework, it is possible to deduce that the framework presents a benefit to the discovery of new knowledge during literature analysis, satisfying RQ2.5 through the contribution of an **end-to-end framework capable**

of handling user-provided literature queries and subsequently analysing them through cluster-based topic modelling with Top2Vec. However, open questions remain, especially when considering the implications of this research to the media and publishing industry, and the wider research domain.

Firstly, the demonstration of the framework for literature analysis was focused upon medical literature extracted from the PubMed archives. PubMed offers several advantages that make it an ideal repository for literature, due to the fact that publications are provided through API resources in a structured XML format. This presents many opportunities that were unexplored in the research of this thesis, such as considering the analysis of distinct aspects of publications, which can be clearly separated into sections - such as methodological approaches, results, and related work sections, to name a few. It is worth taking into consideration that researchers may wish to analyse literature from the perspective of methodological approaches and disregard other aspects of publications, which would be easily performed in the manner presented in the SMARTEN framework. This is something which can be easily performed by extracting the relevant entries from the XML structure of PubMed resources, however, XML structuring of academic literature is not universally performed. For example, the standardised Journal Article Tag Suite[1] is an XML format which enables structured, and therefore, machine-readable, formatting of scientific literature, however is adopted by a small portion of repositories, with these including the National Institute of Health and their PubMed database[2], Taylor & Francis[3], SciELO[4], and Redalyc[5]. While this list is not exhaustive, it is worth highlighting that JATS standardisation is not universally and therefore a significant portion of knowledge in academic repositories cannot be extracted by XML processing. Thus, in any future expansions of this research, caveats in the extraction of information from publications will need to be addressed, particularly as many repositories do not easily provide full-text access through API resources, or provide

[1] `https://jats.niso.org/`

[2] `https://jats.nlm.nih.gov/`

[3] `https://jats.taylorandfrancis.com/`

[4] `https://www.scielo.org/en/`

[5] `https://www.redalyc.org/`

results as PDFs. In this case, PDF extraction techniques such as Optical Character Recognition (OCR) or similar technologies would need to be performed.

In the media and publishing sector, the methodology for analysing literature can be adjusted to accommodate various forms of textual media like news articles or conversations on social media. Ongoing research beyond the completion of this thesis includes a year-long post-doctoral project focused on exploring topic modelling of social media content to produce prompts for journalistic writing, aiming to **amplify the direct impact on the media and publishing field**.

8.3 Transformer Encoders for Parametric Dimensionality Reduction in Clustering

Following the largely *applied* aspect of the research presented in Chapters 4, 5 and 6, the main focus of Chapter 7 embarked upon the specific avenue of dimensionality reduction research. This was conducted based on the identification of UMAP [14] as a necessity for cluster-based topic modelling within the Top2Vec [6] algorithm, due to the limitations encountered when applying clustering algorithms to highly dimensional vectors produced by neural-embedding techniques, as detailed in Section 2.1.4. By investigating the effects of dimensionality upon clustering, it was possible to identify that the introduction of attention-mechanisms through the implementation of an architecture based on the transformer-encoder could contribute to improving dimensionality reduction of text embeddings, with the resulting contribution being a method for improving the topic modelling solution provided by Top2Vec through a transformer-encoder based dimensionality reduction pipeline. Thus, the third general research question was defined as:

RQ3: How can dimensionality reduction be used to improve the accuracy of text clustering and, subsequently, topic modelling?

In addressing this research question, it was necessary to first establish an understanding of the current state of research in DR with respect to clustering, through a preliminary evaluation of existing DR algorithms, as detailed in Section 7.5.3. By comparing these algorithms in their effectiveness as dimensionality reduction pipelines for text clustering with k-Means, it was possible to identify, for the first time, trends in the effects of output dimensionality, which were not identified in the literature prior. Notably, for UMAP, a “knee” curve was identified, generally around output dimensionalities of 2 or 3, where diminishing returns are observed in downstream clustering accuracy, as dimensionality increases beyond this point. This phenomenon has not been identified in the wider literature and presents an opportunity for future industry applications of UMAP and its derivatives, where a output dimensionality of 3 appears to be suitable to maintain the overall quality of original text embeddings in a significantly compressed format. This is of particular value to industries such as publishing, through enabling a significant reduction in the computing resources required for the storage of low-dimensional embeddings, while also ensuring that downstream applications of such embeddings can circumvent the curse of dimensionality. It remains, however, necessary to analyse this phenomenon further from the perspective of other types of information, such as images, to evaluate whether UMAP can also compress the dimensionality of these in a similar manner.

In further expanding the analysis of DR with respect to downstream clustering performance, an alternative approach was explored, based on metric learning through the UMAP, and parametric UMAP algorithms. The main contributions when investigating this RQ were presented through identifying that the **introduction of metric-learning by using a small portion of labelled data during DR with UMAP was beneficial in improving the downstream clustering solution** across all experiments. It is worth noting that for this training strategy, metric learning with UMAP, and the proposed attention-based architectures for UMAP, conferred a significant improvement in the downstream clustering accruacy, even providing a robustness in undersampled classes even when using a significantly smaller dataset. This presents opportunities for future research and industry applications where only a small amount of labelled data can be obtained. For example, given the time and resource costs associated with manual labelling, it is feasible to propose that the introduction of UMAP in metric learning pipelines for DR can assist not only in reducing the dimensionality of data for more efficient clustering, but also provide a benefit to the accuracy of downstream analysis.

In addressing RQ3.4 and RQ3.5, a novel paradigm in dimensionality reduction was explored, facilitated by the recently proposed parametric UMAP algorithm. The proposed study of two architectures which have demonstrated success in other NLP domains, namely the RNN with Attention and the Transformer-Encoder. Based on the outcomes of the previous RQs, it was proposed to evaluate both of these architectures in a metric-learning pipeline, in order to compare the effects that these would have on downstream clustering with respect to existing state-of-the-art algorithms UMAP and parametric UMAP. The findings of these experiments were two-fold. First, the study demonstrated that the transformer-encoder based pipeline enabled a significant improvement in the downstream clustering accuracy of k-Means, attaining the highest accuracy in three of the four datasets used during the experiment, beating the established state-of-the-art UMAP supervised algorithm. Second, after establishing the benefit that the introduction of the transformer-encoder pipeline conferred on downstream clustering, a deeper analysis of the accuracy of individual classes within the data indicated that the proposed implementation enabled a degree of robustness to imbalances in the data when used in a metric-learning pipeline.

The adoption of the transformer-encoder into a parametric DR pipeline for topic modelling has evidently had a considerable influence upon the quality of the topics identified, however it is interesting to note that these are most prevalent in the smaller, BBC News dataset. In Section 7.5, viewed from a metric learning angle, the transformer-encoder shows the most substantial improvement on the smaller datasets like TREC-6 and TREC-50, surpassing an RNN equipped with attention mechanisms by a significant margin. This trend is also evident in Section 7.6, where it becomes apparent that the integration of the transformer-encoder has a notably positive impact on both the metric and the quality of topics generated for the BBC News dataset, which is relatively small. Based on these assessments, it can be inferred that transformer-encoders for parametric dimensionality reduction tend to generalise effectively to small datasets, unlike the other techniques investigated, that would necessitate a significantly larger dataset to achieve comparable results in quality. It is also evident that the introduction of additional residual connections into the network architecture is also beneficial in enhancing the quality of the topic modelling solution, possibly due to addressing the vanishing gradient problem to some degree, however a further analysis of this is necessary, and is one of the considerations of continued research after the completion of the Ph.D.

Similarly with regard to transformer-encoders, the use of attention mechanisms, which have proven revolutionary in NLP, **may not be the best approach for parametric DR**. Attention mechanisms, in essence, permit a model to focus on specific aspects of a sequence rather than the entire sequence. This is ideal for modelling text, as some words in a sentence are more important than others in understanding the overall semantics. In comparison, it should be considered that the entire input sequence is important for DR. If attention mechanisms in transformer-encoders for parametric DR are only focusing on specific areas of the input sequence, then it is a valid argument that the model is failing to account for the entirety of the information encoded within the input sequence. If this is the case, then a better approach would be to consider alternative network architectures that can ensure accounting for the entire sequence. A simple, feed-forward network is one example of this, however it was apparent from the experiments conducted in Chapter 7 that the transformer-encoder outperformed these in DR tasks. Still, more advanced architectures should be considered. Alternatively, if it can be proven that attention mechanisms are, in fact, focusing upon specific aspects of the input sequence, then there are many interesting outcomes of investigating this, which could have considerable impact on the field of NLP. It may be possible to demonstrate, for instance, that parametric DR is capable of producing low-dimensional embeddings based on partial input sequences, or that certain areas of embeddings are more important than others.

8.4 Value of Research to Broader Research Domains

The research presented in this thesis was carried out with the objective of specifically contributing to the media and publishing industry. In Chapter 4, this is performed through an information retrieval approach to the retrieval of social media influencers based on full-text querying, and the evaluation of meta-embedding weightings constructed from multiple data sources. In Chapters 5 and 6, the focus was adapted to the task of assisting in the analysis of literature, wherein topic modelling techniques were first evaluated through case-studies of specific domains, before proposing a framework that enables the analysis of large volumes of literature, based on custom queries defined by users. After the identification of dimensionality reduction as a key component in cluster-based topic modelling, Chapter 7 set out to contribute to improving downstream clustering through the proposal of attention mechanisms as encoder networks for parametric dimensionality reduction, with the resulting proposal of transformer-encoder pipelines for DR being demonstrated to provide a benefit to the granularity of topic modelling results.

The findings discussed in Chapter 4 offer valuable insights into how the creation of meta-embeddings from various data sources enhances semantic search tasks, which presents value to the broader research domain, where the approach could be applied to other downstream analyses involving semantic similarity. Similarly, the contributions of Chapters 5 and 6, through the provision of an assisted framework for literature review, offer a degree of flexibility in the types of data to which the framework could be adapted. Specifically, by building the framework upon unsupervised learning via cluster-based topic modelling and subsequent semantic similarity measures for evaluating similarities between topics, it is feasible to apply the framework to any textual domain in order to assist with the comprehension of text. Examples of where such a framework could present value include the analysis of news publications, in order to identify themes and similarities between news topics, or social media posts to identify online discussion topics.

Finally, in Chapter 7, the proposed transformer-encoder pipeline for parametric dimensionality reduction was investigated with a focus on clustering and topic modelling. However, in the wider domain of research involving high-dimensional vectors, a common approach to the visualisation of embeddings involves the prior dimensionality reduction of embeddings, in order to enable the visualisation of them in a 2-dimensional, or 3-dimensional manner [339]. Thus, given the contributions that the transformer-encoder based pipeline for parametric DR made evident, it is worth considering whether this technique would also be valuable to high-dimensional visualisation. Similarly, text-embeddings were the focus of the research of this thesis, and it remains to be seen whether the results of the investigation into parametric DR can be replicated in different domains such as images, or graph embeddings.

8.5 Limitations and Future Work

A challenge to the retrieval and recommendation of social media influencers based on full-text queries, as presented in Chapter 4, stems from the niche domain in which the study was applied. The **presented framework is the first to address the problem of account recommendation from a journalistic perspective, by taking into account full-text queries of journalistic content**. This presents the problem of benchmarking such an approach. Therefore, there is a need for suitable evaluation datasets. This would ensure easy quantification of new research. Furthermore, the computation of meta-embeddings focused on data obtained from user biographies and posted content produced by a user, taken from Twitter. However, alternative research avenues could investigate the triangulation of additional sources of data.

In terms of research supporting the literature review process, a significant challenge lies in the quality of the data available. Publishers provide access to research through in-house API tools, with the amount of data that can be obtained by these resources differing between publishers. For example, some platforms may provide abstracts only, while others provide access to full-text publications, or provide full-text only for open-access articles. This results in a significant disparity in the data at hand when multiple data sources are used, a practice that is essential to provide a complete picture of the state of the literature during analysis. This problem is further exacerbated by the common practice of providing publications as PDF files, so that extraction techniques need to be employed in order to convert the publication content into a textual medium that permits downstream analysis.

In Chapter 7, metric-learning was performed upon a subset of of data when investigating attention mechanisms as a method for semi-supervised dimensionality reduction. For this approach, learning was conducted only upon this small sample. However, one alternative is to perform masking of the full dataset, such that a sample of labelled could be provided, with the rest of the unlabelled data also being used in the training process. This approach has its merits as a potential avenue of further investigation in how to address overfitting in undersampled classes; however, it does not fall within the main scope of this thesis. Thus, in future works it is worth considering the implications of this alternative training strategy in improving the performance of downstream clustering.

Several potential research directions can be suggested to enhance the evidence supporting the findings that the suggested transformer-encoder pipeline with parametric UMAP offers advantages for minority classes and improves overall clustering accuracy. First, from a supervised metric-learning perspective, an in-depth ablation study of the transformer-encoder architecture, evaluated in relation to per-class accuracy, could contribute to reinforcing findings. This could be applied to domains other than the clustering of text embeddings. Secondly, from an unsupervised learning perspective, it would be beneficial to investigate the performance of the transformer-encoder when applied outside of the metric-learning methodology used in this study. This could be evaluated in a manner similar to that of the study provided in Chapter 7, based on accuracy. Alternatively, there remain open questions as to the effects that the clustering method selected, k-Means, has had upon the outcomes of the study. Subsequent investigations could evaluate the transformer-encoder within parametric UMAP in conjunction with different clustering algorithms.

A novel avenue of further research into DR, which has been considered in a post-doctoral research project, is based on the advancement of parametric DR to account for pre-trained language models. This is based on the assumption that research thus far has focused on training a neural network for the specific tasks of learning an approximation of the DR function of UMAP. Thus, the network does not have any prior knowledge of the underlying semantics of the text, and aims to produce low-dimensional representations based only on optimising UMAP loss. In order to address this limitation, future research will seek to investigate the adoption of pre-trained language models as encoder networks for parametric DR, in order to evaluate whether this can produce low-dimensional embeddings. The general methodology of this approach is already established and works under the assumption that a graph representation of the original high-dimensional data can be produced from embeddings of text. However, these embeddings are not used as an input sequence to the parametric DR process. Instead, after computing the high-dimensional graph, a pre-trained encoder network based on the weights of the original embedding model is applied, to optimise a low-dimensional graph representation based on the original text sequence. This ensures that the encoder network could leverage the contextual understanding of pre-trained language models when computing low-dimensional representations of text. In addressing this future research, significant modifications to the underlying parametric UMAP algorithm are necessary, and this could demonstrate a significant impact to the domain of DR, through the unexplored domain of pre-training parametric DR algorithms. In industry, this is particularly advantageous, since it allows for the computation of low-dimensional embeddings, instead of depending on high-dimensional ones. This could lead to a substantial decrease in expenses related to storing extensive vectors and circumvent the curse of dimensionality during downstream analysis entirely.

CHAPTER 9

Conclusion

This thesis has presented research on the challenges of making sense of unlabelled textual media through the avenue of unsupervised learning. Beginning with a focus on semantic search, research identified the construction of meta-embeddings based on posted content and user biographies, as an effective method in enhancing the quality of semantic searching, based on full-text querying of journalistic content, with the results of this research being deployed in a commercial setting, enabling a direct contribution to impact in the publishing industry. From the perspective of topic modelling, research presented a series of investigations into topic modelling techniques for the analysis and comprehension of academic literature, concluding with the provision of a framework, and software implementation, which enabled the identification of topics within literature, and an analysis of the semantic relationships of topics, in order to assist in the finding of new knowledge for medical experts and medical students. Based on a study of medical experts and students, it was possible to demonstrate the benefit that the framework, and software implementation, could contribute to knowledge discovery, something which could ultimately be adapted to facilitate the analysis of a plethora of potential textual media. Finally, in order to directly contribute to the domain of text clustering, which is essential for the topic modelling techniques adopted in this thesis, it was proposed to investigate the introduction of architectural designs through parametric dimensionality reduction. Through the proposal of attention-based neural networks, it was possible to demonstrate the significant effects of the transformer-encoder in enhancing downstream clustering, when using a metric-learning approach when training the parametric encoder. By further exploring this research direction, it was feasible to determine that integrating transformer-encoders into the parametric dimensionality reduction technique UMAP could enhance the effectiveness of topic modeling in terms of metrics and quality. Furthermore, by the introduction of additional residual connections into the network architecture, further improvements were observed from this perspective. This is something which can have a direct contribution to the wider field of topic modelling, rather than just the media and publishing industry.

The initial scope of this PhD, as defined in the project brief, outlined the need for predictive and prescriptive analytics, which can contribute to the media and publication industry. This was defined by the industrial sponsor Distinctive Publishing[1] and their Reveela[2] platform, which aims to address current barriers to journalistic and publishing processes by taking advantage of advances in data mining and natural language processing. As part of the industrial Ph.D. partnership with Distinctive Publishing, opportunities and obstacles were present throughout the project. Unique opportunities for the real-world application of Ph.D research results ensured that industrial impact has been achieved. This is best presented through the research described in Chapter 4, which has since been finalised into a commercial product, available on the Reveela[3] platform. Furthermore, the findings of Chapters 5, 6, and 7 have since resulted in a successful application for a year-long impact project aimed at continuing research and implementation of topic modelling architectures with the same industry sponsor, Distinctive Publishing. Obstacles throughout the Ph.D. project were present and unique to the industrial nature of the research. Among the most significant considerations were intellectual property requirements for the shar-

[1] https://www.distinctivepublishing.co.uk/

[2] https://www.reveela.com

[3] `www.reveela.com`

ing of industrial data relevant to the publishing industry, such as large databases of collated journalists and associated publications. This information could not be freely shared in published academic literature and presented the need to apply open-access data in place of confidential industry data. This is best represented in Chapters 5 and 6, where research investigated the topic modelling of academic texts. Based on the findings of the research discussed in this thesis, the next phase of the research will aim to make a further direct contribution to the media and publishing sector, by utilising the topic modelling and dimensionality reduction advancements that have been previously described, and evaluating the generated impact.

Bibliography

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013. (document), 2.1, 2.1.1, 2.1.1, 2.1, 3.2.3

[2] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*, pp. 1188–1196, PMLR, 2014. (document), 2.1.1, 2.2, 5.3.1

[3] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 824–836, 2018. (document), 2.1.3, 3.2.2, 3.1, 4.3.4, 4.3.6

[4] T. Sainburg, L. McInnes, and T. Q. Gentner, “Parametric umap embeddings for representation and semisupervised learning,” *Neural Computation*, vol. 33, no. 11, pp. 2881–2907, 2021. (document), 1.3, 2.1.6, 3.1, 3.1.3, 3.2.3, 3.2.4, 3.2.4, 3.2, 3.2.5, 7.1, 7.4.2, 7.5.4, 7.5.8, 7.6.1, 7.6.4, 7.7

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017. (document), 2.1.2, 3.1, 3.2.6, 3.3, 3.4, 3.2.6, 4.2.2, 4.3.3, 7.4.1, 7.4.3, 7.5.2, 7.5.5, 7.5.5, 7.6.1

[6] D. Angelov, “Top2vec: Distributed representations of topics,” *CoRR*, vol. abs/2008.09470, 2020. (document), 1.3, 2.1.5, 3.2.3, 3.2.4, 5.1, 5.4.3, 6.1, 6.1, 6.4.1, 7.1, 7.6, 8.2, 8.3

[7] Department for Culture, Media and Sport, “The cairncross review: a sustainable future for journalism,” 2019. 1.1, 4.1, 4.1.1

[8] N. Newman, R. Fletcher, A. Kalogeropoulos, D. Levy, and R. K. Nielsen, “Reuters institute digital news report 2018, figure q3,” Oct 2018. 1.1

[9] Statista, "Share of individuals reading or downloading online news, newspapers or magazines in great britain from 2007 to 2020," tech. rep., 2020. 1.1, 4.1

[10] Jul 2023. 1.1

[11] P. Harrigan, T. M. Daly, K. Coussement, J. A. Lee, G. N. Soutar, and U. Evers, "Identifying influencers on social media," *International Journal of Information Management*, vol. 56, p. 102246, 2021. 1.2.1

[12] R. Sagayam, S. Srinivasan, and S. Roshni, "A survey of text mining: Retrieval, extraction and indexing techniques," *International Journal of Computational Engineering Research*, vol. 2, no. 5, pp. 1443–1446, 2012. 1.3, 2.1.3

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003. 1.3, 2.1.5, 5.1, 5.1.1, 5.2, 5.3.1, 5.3.4, 5.4.3

[14] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018. 1.3, 2.1.5, 2.1.6, 3.2.3, 3.2.4, 3.2.4, 3.2.4, 3.2.5, 6.3.1, 6.7.3, 7.1, 7.4.2, 7.5.3, 7.5.8, 8.3

[15] S.-T. Park and W. Chu, "Pairwise preference regression for cold-start recommendation," in *Proceedings of the third ACM conference on Recommender systems*, pp. 21–28, 2009. 1.5, 4.2.1, 8.1

[16] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*, 2020. 1.5, 2.1, 2.1.6, 5.1, 5.1.1, 5.3, 5.3.1, 5.3.3, 5.3.4, 6.3.1, 6.3.1, 6.7, 6.7.3, 6.7.4, 6.9, 7.4.1, 7.4.2, 7.6.1, 7.7

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1.5, 3.2.6, 3.2.6, 7.5.5, 7.6.1

[18] S. Basodi, C. Ji, H. Zhang, and Y. Pan, "Gradient amplification: An efficient way to train deep neural networks," *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 196–207, 2020. 1.5, 3.2.6, 7.5.5, 7.6.1

[19] S. Selva Birunda and R. Kanniga Devi, "A review on word embedding techniques for text classification," *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, pp. 267–281, 2021. 2.1

[20] E. Poslavskaya and A. Korolev, "Encoding categorical data: Is there yet anything'hotter'than one-hot encoding?," *arXiv preprint arXiv:2312.16930*, 2023. 2.1

[21] W. Pu, N. Liu, S. Yan, J. Yan, K. Xie, and Z. Chen, "Local word bag model for text categorization," in *Seventh IEEE international conference on data mining (ICDM 2007)*, pp. 625–630, IEEE, 2007. 2.1

[22] L. S. da Costa, I. L. Oliveira, and R. Fileto, "Text classification using embeddings: a survey," *Knowledge and Information Systems*, vol. 65, no. 7, pp. 2761–2803, 2023. 2.1

[23] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial Intelligence Review*, pp. 1–81, 2023. 2.1

[24] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972. 2.1

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013. 2.1, 3.2.3

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. 2.1, 2.1.1, 2.1.2, 2.1.4, 2.1.6, 4.3.3, 5.3.1, 5.3.4

[27] K. Sinha, R. Jia, D. Hupkes, J. Pineau, A. Williams, and D. Kiela, "Masked language modeling and the distributional hypothesis: Order word matters pre-training for little," *arXiv preprint arXiv:2104.06644*, 2021. 2.1

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. 2.1, 2.1.2

[29] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," 2016. 2.1

[30] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," 2018. 2.1

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. 2.1, 2.1.2, 2.1.6, 3.2.6, 7.4.3, 7.5.1, 7.5.6, 7.7

[32] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022. 2.1, 2.1.5

[33] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, Nov. 2019. 2.1, 2.1.6, 4.3.3, 5.3.4

[34] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using bert," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, vol. 1, pp. 1–5, 2019. 2.1, 7.4.1

[35] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, “Large language models in machine translation,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (J. Eisner, ed.), (Prague, Czech Republic), pp. 858–867, Association for Computational Linguistics, June 2007. 2.1.1

[36] J. Coates and D. Bollegala, “Frustratingly easy meta-embedding - computing meta-embeddings by averaging source word embeddings,” *CoRR*, vol. abs/1804.05262, 2018. 2.1.1, 3.1.1

[37] B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (K. Knight, H. T. Ng, and K. Oflazer, eds.), (Ann Arbor, Michigan), pp. 115–124, Association for Computational Linguistics, June 2005. 2.1.1

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, and S. Bethard, eds.), (Seattle, Washington, USA), pp. 1631–1642, Association for Computational Linguistics, Oct. 2013. 2.1.1

[39] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011. 2.1.1

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018. 2.1.1, 2.1.2, 3.2.6, 5.3.4, 5.4.2, 6.3.1, 6.3.1, 7.4.1, 7.4.3, 7.5.1, 8.1

[41] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “Mpnet: Masked and permuted pre-training for language understanding,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020. 2.1.1, 3.2.6, 4.3.3, 4.3.4, 4.3.6

[42] A. Nayak, H. Timmapathini, K. Ponnalagu, and V. G. Venkoparao, “Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words,” in *Proceedings of the first workshop on insights from negative results in NLP*, pp. 1–5, 2020. 2.1.1

[43] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, *An Introduction to Information Retrieval*, pp. 3–11. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. 2.1.3

[44] G. Salton, *Automatic Information Organization and Retrieval.* McGraw-Hill computer science series, McGraw-Hill, 1968. 2.1.3

[45] B. O'Neill, "Chapter 2 - frame fields," in *Elementary Differential Geometry (Second Edition)* (B. O'Neill, ed.), pp. 43–99, Boston: Academic Press, second edition ed., 2006. 2.1.3

[46] J. Han, M. Kamber, and J. Pei, "2 - getting to know your data," in *Data Mining (Third Edition)* (J. Han, M. Kamber, and J. Pei, eds.), The Morgan Kaufmann Series in Data Management Systems, pp. 39–82, Boston: Morgan Kaufmann, third edition ed., 2012. 2.1.3

[47] C. Li, M. Zhang, D. G. Andersen, and Y. He, "Improving approximate nearest neighbor search through learned adaptive early termination," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 2539–2554, 2020. 2.1.3

[48] Q. Huang, J. Feng, Y. Zhang, Q. Fang, and W. Ng, "Query-aware locality-sensitive hashing for approximate nearest neighbor search," *Proceedings of the VLDB Endowment*, vol. 9, no. 1, pp. 1–12, 2015. 2.1.3

[49] L. Gong, H. Wang, M. Ogihara, and J. Xu, "idec: indexable distance estimating codes for approximate nearest neighbor search," *Proceedings of the VLDB Endowment*, vol. 13, no. 9, 2020. 2.1.3

[50] A. Arora, S. Sinha, P. Kumar, and A. Bhattacharya, "Hd-index: Pushing the scalability-accuracy boundary for approximate knn search in high-dimensional spaces," *arXiv preprint arXiv:1804.06829*, 2018. 2.1.3

[51] C. Silpa-Anan and R. Hartley, "Optimised kd-trees for fast image descriptor matching," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008. 2.1.3

[52] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010. 2.1.3

[53] Z. Pan, L. Wang, Y. Wang, and Y. Liu, "Product quantization with dual codebooks for approximate nearest neighbor search," *Neurocomputing*, vol. 401, pp. 59–68, 2020. 2.1.3

[54] C. Fu, C. Xiang, C. Wang, and D. Cai, "Fast approximate nearest neighbor search with the navigating spreading-out graph," *arXiv preprint arXiv:1707.00143*, 2017. 2.1.3

[55] M. Wang, X. Xu, Q. Yue, and Y. Wang, "A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search," *arXiv preprint arXiv:2101.12631*, 2021. 2.1.3

[56] M. Aumüller, E. Bernhardsson, and A. Faithfull, "Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms," *Information Systems*, vol. 87, p. 101374, 2020. 2.1.3

[57] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin, "Approximate nearest neighbor search on high dimensional data—experiments, analyses, and improvement," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1475–1488, 2019. 2.1.3

[58] Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov, "Approximate nearest neighbor algorithm based on navigable small world graphs," *Information Systems*, vol. 45, pp. 61–68, 2014. 2.1.3

[59] J. Wang and S. Li, "Query-driven iterated neighborhood graph search for large scale indexing," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 179–188, 2012. 2.1.3

[60] J. Wang, J. Wang, G. Zeng, Z. Tu, R. Gan, and S. Li, "Scalable k-nn graph construction for visual descriptors," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1106–1113, IEEE, 2012. 2.1.3

[61] C. Fu, C. Xiang, C. Wang, and D. Cai, "Fast approximate nearest neighbor search with the navigating spreading-out graph," *Proc. VLDB Endow.*, vol. 12, p. 461–474, jan 2019. 2.1.3

[62] K. Zhao, P. Pan, Y. Zheng, Y. Zhang, C. Wang, Y. Zhang, Y. Xu, and R. Jin, "Large-scale visual search with binary distributed graph at alibaba," in *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 2567–2575, 2019. 2.1.3

[63] M. Iwasaki, "Pruned bi-directed k-nearest neighbor graph for proximity search," in *International Conference on Similarity Search and Applications*, pp. 20–33, Springer, 2016. 2.1.3

[64] M. Iwasaki and D. Miyazaki, "Optimization of indexing based on k-nearest neighbor graph for proximity search in high-dimensional data," *arXiv preprint arXiv:1810.07355*, 2018. 2.1.3

[65] K. Sugawara, H. Kobayashi, and M. Iwasaki, "On approximately searching for similar word embeddings," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2265–2275, 2016. 2.1.3

[66] B. Naidan, L. Boytsov, and E. Nyberg, "Permutation search methods are efficient, yet faster search is possible," *arXiv preprint arXiv:1506.03163*, 2015. 2.1.3

[67] OpenAI, "Embeddings documentation," 2023. Available at `https://platform.openai.com/docs/guides/embeddings/second-generation-models`, Accessed 05/10/2023. 2.1.4

[68] R. Bellman, R. Corporation, and K. M. R. Collection, *Dynamic Programming*. Rand Corporation research study, Princeton University Press, 1957. 2.1.4, 2.1.4, 7.4.1

[69] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," in *Database Theory — ICDT'99* (C. Beeri and P. Buneman, eds.), (Berlin, Heidelberg), pp. 217–235, Springer Berlin Heidelberg, 1999. 2.1.4, 2.1.4, 7.4.1, 7.4.2

[70] I. Assent, "Clustering high dimensional data," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 340–350, 2012. 2.1.4, 7.4.1, 7.4.2

[71] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*, pp. 420–434, Springer, 2001. 2.1.4, 7.4.1

[72] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition*, pp. 273–309, 2004. 2.1.4, 2.1.4, 7.1

[73] T. S. Madhulatha, "An overview on clustering methods," *CoRR*, vol. abs/1205.1117, 2012. 2.1.4, 7.1

[74] N. Bouhmala, "How good is the euclidean distance metric for the clustering problem," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 312–315, 2016. 2.1.4, 7.1

[75] C. X. Gao, D. Dwyer, Y. Zhu, C. L. Smith, L. Du, K. M. Filia, J. Bayer, J. M. Menssink, T. Wang, C. Bergmeir, S. Wood, and S. M. Cotton, "An overview of clustering methods with guidelines for application in mental health research," *Psychiatry Research*, vol. 327, p. 115265, 2023. 2.1.4, 7.1

[76] J. Clatworthy, D. Buick, M. Hankins, J. Weinman, and R. Horne, "The use and reporting of cluster analysis in health psychology: A review," *British Journal of Health Psychology*, vol. 10, no. 3, pp. 329–358, 2005. 2.1.4, 7.1

[77] S. Xia, Z. Xiong, Y. Luo, WeiXu, and G. Zhang, "Effectiveness of the euclidean distance in high dimensional spaces," *Optik*, vol. 126, no. 24, pp. 5614–5619, 2015. 2.1.4

[78] D. M. Blei and J. D. Lafferty, "Topic models," in *Text mining*, pp. 101–124, Chapman and Hall/CRC, 2009. 2.1.5

[79] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013. 2.1.5, 3.2.3, 3.2.4, 5.3.4, 6.4.1, 6.7.3, 7.1, 7.4.1, 7.4.2, 7.6.4, 7.7

[80] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 09 2019. 2.1.5, 5.3.4, 6.3

[81] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020. 2.1.5

[82] T.-H. Tran, T.-D. Cao, and T.-T.-H. Tran, *HDBSCAN: Evaluating the Performance of Hierarchical Clustering for Big Data*, pp. 273–283. Cham: Springer International Publishing, 2021. 2.1.5

[83] L. van der Maaten, E. O. Postma, and J. van den Herik, "Dimensionality reduction: A comparative review," 2009. 2.1.6, 5.1, 7.4.2

[84] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987. 2.1.6, 7.4.2, 7.5.3

[85] R. A. FISHER, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936. 2.1.6, 5.3.1, 7.4.2, 7.5.3

[86] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008. 2.1.6, 7.4.2, 7.5.3

[87] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *NIPS*, 2002. 2.1.6, 7.4.2

[88] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)," *Computer Science Review*, vol. 40, p. 100378, 2021. 2.1.6, 7.4.2

[89] N. Pezzotti, A. Mordvintsev, T. Hollt, B. Lelieveldt, E. Eisemann, and A. Vilanova, "Linear tsne optimization for the web," *arXiv preprint arXiv:1805.10817*, vol. 2, 2018. 2.1.6, 2.1.6, 7.5.3

[90] J. Barnes and P. Hut, "A hierarchical o (n log n) force-calculation algorithm," *nature*, vol. 324, no. 6096, pp. 446–449, 1986. 2.1.6, 7.5.3

[91] L. van der Maaten, "Barnes-hut-sne," 2013. 2.1.6, 7.5.3

[92] L. Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial intelligence and statistics*, pp. 384–391, PMLR, 2009. 2.1.6, 7.4.2

[93] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 2.1.6, 7.4.2

[94] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization," in *ICML*, 1997. 2.1.6, 3.1, 7.4.2, 7.5.1

[95] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study," in *Image and Signal Processing* (A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud, eds.), (Cham), pp. 317–325, Springer International Publishing, 2020. 2.1.6, 7.4.2

[96] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019. 2.1.6, 7.4.2, 7.7

[97] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel, "A review of umap in population genetics," *Journal of Human Genetics*, vol. 66, no. 1, pp. 85–91, 2021. 2.1.6, 7.4.2

[98] D. Kobak and G. C. Linderman, "Umap does not preserve global structure any better than t-sne when using the same initialization," *bioRxiv*, 2019. 2.1.6, 7.7

[99] F. Trozzi, X. Wang, and P. Tao, "Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: A comparison study," *The Journal of Physical Chemistry B*, vol. 125, no. 19, pp. 5022–5034, 2021. 2.1.6

[100] Z. Ghahramani, "Unsupervised learning," in *Summer school on machine learning*, pp. 72–112, Springer, 2003. 3.1

[101] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 2065–2073, 2014. 3.1.1

[102] W. Yin and H. Schütze, "Learning word meta-embeddings," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (K. Erk and N. A. Smith, eds.), (Berlin, Germany), pp. 1351–1360, Association for Computational Linguistics, Aug. 2016. 3.1.1

[103] Y. A. Malkov and A. Ponomarenko, "Growing homophilic networks are natural navigable small worlds," *PloS one*, vol. 11, no. 6, p. e0158162, 2016. 3.2.2

[104] Y. Lifshits and S. Zhang, "Combinatorial algorithms for nearest neighbors, near-duplicates and small-world design," in *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pp. 318–326, SIAM, 2009. 3.2.2

[105] A. Karbasi, S. Ioannidis, and L. Massoulié, "From small-world networks to comparison-based search," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3056–3074, 2015. 3.2.2

[106] W. Xiao, Y. Zhan, R. Xi, M. Hou, and J. Liao, "Enhancing hnsw index for real-time updates: Addressing unreachable points and performance degradation," *arXiv preprint arXiv:2407.07871*, 2024. 3.2.2

[107] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation.," *Psychological review*, vol. 114, no. 2, p. 211, 2007. 3.2.3

[108] T. Mikolov, W.-t. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013. 3.2.3

[109] W. Dong, C. Moses, and K. Li, “Efficient k-nearest neighbor graph construction for generic similarity measures,” in *Proceedings of the 20th international conference on World wide web*, pp. 577–586, 2011. 3.2.4, 7.5.3

[110] R. Hodgson, J. Wang, A. I. Cristea, and J. Graham, “Hybrid weighted retrieval of twitter users for temporally relevant full-text querying in the media industry,” in *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, pp. 38–44, 2022. 3.1, 4.1

[111] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, “The protein kinase complement of the human genome,” *Science*, vol. 298, pp. 1912 – 1934, 2002. 3.1, 6.4.1

[112] “Protein kinases, kinomes and evolution, at kinase.com.” 3.1, 6.4.1

[113] X. Li and D. Roth, “Learning question classifiers,” in *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING ’02, (USA), p. 1–7, Association for Computational Linguistics, 2002. 3.1, 7.5.1

[114] D. Greene and P. Cunningham, “Practical solutions to the problem of diagonal dominance in kernel document clustering,” in *Proc. 23rd International Conference on Machine learning (ICML’06)*, pp. 377–384, ACM Press, 2006. 3.1, 7.6.2

[115] “Entrez programming utilities help,” Jan 1970. 3.3, 6.4.1, 6.7.2

[116] Statista, “Share of respondents who had their own social network profile in the united kingdom (uk) from 2015 to 2020, by age,” tech. rep., 2022. 4.1

[117] G. Cormode, V. Shkapenyuk, D. Srivastava, and B. Xu, “Forward decay: A practical time decay model for streaming systems,” in *2009 IEEE 25th International Conference on Data Engineering*, pp. 138–149, 2009. 4.1.2, 4.2.2, 4.3.5, 8.1

[118] R. Burke, A. Felfernig, and M. Göker, “Recommender systems: An overview,” *Ai Magazine*, vol. 32, pp. 13–18, 2011. 4.2.1

[119] C. C. Aggarwal, *Attack-Resistant Recommender Systems*, pp. 385–410. Cham: Springer International Publishing, 2016. 4.2.1

[120] M. P. O’Mahony, N. J. Hurley, and G. C. Silvestre, “Promoting recommendations: An attack on collaborative filtering,” Database and Expert Systems Applications, pp. 494–503, Springer, Springer Berlin Heidelberg, 2002. 4.2.1

[121] M. O’Mahony, N. Hurley, N. Kushmerick, and G. Silvestre, “Collaborative recommendation: A robustness analysis,” *ACM Transactions on Internet Technology (TOIT)*, vol. 4, no. 4, pp. 344–377, 2004. 4.2.1

[122] H. Bast, B. Buchhold, and E. Haussmann, “Semantic search on text and knowledge bases,” *Foundations and Trends® in Information Retrieval*, vol. 10, pp. 119–271, 2016. 4.2.2

[123] M. Verleysen and D. François, “The curse of dimensionality in data mining and time series prediction,” Computational Intelligence and Bioinspired Systems, pp. 758–770, Springer, Springer Berlin Heidelberg, 2005. 4.2.2

[124] V. Pestov, “Is the k-nn classifier in high dimensions affected by the curse of dimensionality?,” *Computers & Mathematics with Applications*, vol. 65, no. 10, pp. 1427–1437, 2013. 4.2.2

[125] N. Reimers and I. Gurevych, “The curse of dense low-dimensional information retrieval for large index sizes,” Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 605–611, Association for Computational Linguistics, 2020. 4.2.2

[126] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018. 4.2.2, 5.3.1

[127] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin, “Approximate nearest neighbor search on high dimensional data–experiments, analyses, and improvement,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1475–1488, 2019. 4.2.2

[128] P. Indyk and R. Motwani, “Approximate nearest neighbors: Towards removing the curse of dimensionality,” in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC ’98, (New York, NY, USA), p. 604–613, Association for Computing Machinery, 1998. 4.2.2

[129] A. Farseev, K. Lepikhin, H. Schwartz, E. K. Ang, and K. Powar, “Somin.ai: Social multimedia influencer discovery marketplace,” in *Proceedings of the 26th ACM International Conference on Multimedia*, MM ’18, (New York, NY, USA), p. 1234–1236, Association for Computing Machinery, 2018. 4.2.2

[130] F. Abel, I. Celik, G.-J. Houben, and P. Siehndel, “Leveraging the semantics of tweets for adaptive faceted search on twitter,” in *International Semantic Web Conference*, pp. 1–17, Springer, Springer, 2011. 4.2.2

[131] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, *Dbpedia: A nucleus for a web of open data*, pp. 722–735. Springer, 2007. 4.2.2

[132] N. Craswell, *Mean Reciprocal Rank*, pp. 1703–1703. Boston, MA: Springer US, 2009. 4.2.2

[133] J. Lin, K. Sugiyama, M.-Y. Kan, and T.-S. Chua, “Addressing cold-start in app recommendation: Latent user models constructed from twitter followers,” in *Proceedings of the 36th International ACM SIGIR Conference on Research

and Development in Information Retrieval, SIGIR '13, (New York, NY, USA), p. 283–292, Association for Computing Machinery, 2013. 4.2.2

[134] C. De Maio, M. Gallo, F. Hao, and E. Yang, "Who and where: context-aware advertisement recommendation on twitter," *Soft Computing*, vol. 25, no. 1, pp. 379–387, 2021. 4.2.2

[135] P. Bansal, S. Jain, and V. Varma, "Towards semantic retrieval of hashtags in microblogs," in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, (New York, NY, USA), p. 7–8, Association for Computing Machinery, 2015. 4.2.2

[136] A. Belhadi, Y. Djenouri, J. C.-W. Lin, C. Zhang, and A. Cano, "Exploring pattern mining algorithms for hashtag retrieval problem," *IEEE Access*, vol. 8, pp. 10569–10583, 2020. 4.2.2

[137] V. W. Anelli, T. Di Noia, E. Di Sciascio, A. Ragone, and J. Trotta, "Local popularity and time in top-n recommendation," Advances in Information Retrieval, pp. 861–868, Springer International Publishing, 2019. 4.2.2, 8.1

[138] U. Government, "The cairncross review, a sustainable future for journalism," government report, Department for Digital Culture, Media & Sport, Feb 2019. 4.3

[139] J. Jiang, J. Thomason, F. Barbieri, and E. Ferrara, "Geolocated social media posts are happier: Understanding the characteristics of check-in posts on twitter," in *Proceedings of the 15th ACM Web Science Conference 2023*, pp. 136–146, 2023. 4.3.1

[140] T. Elmas, "The impact of data persistence bias on social media studies," in *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, (New York, NY, USA), p. 196–207, Association for Computing Machinery, 2023. 4.3.1

[141] T. Elmas, R. Overdorf, and K. Aberer, "Misleading repurposing on twitter," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 209–220, Jun. 2023. 4.3.1

[142] S. F. Seyfosadat and R. Ravanmehr, "Systematic literature review on identifying influencers in social networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 567–660, 2023. 4.3.1

[143] C. M. Alis, M. T. Lim, H. S. Moat, D. Barchiesi, T. Preis, and S. R. Bishop, "Quantifying regional differences in the length of twitter messages," *PLOS ONE*, vol. 10, pp. 1–10, 04 2015. 4.3.1

[144] K. Żyłka, "Shorter or longer tweets? One year with the expanded character limit — sotrender.com." `https://www.sotrender.com/blog/2018/10/shorter-longer-tweets-one-year-expanded-character-limit-analysis/`. [Accessed 09-09-2024]. 4.3.1

[145] A. Kantrowitz, "Twitter Users Like Long Tweets More Than Short Ones — buzzfeednews.com." `https://www.buzzfeednews.com/article/alexkantrowitz/early-data-shows-longer-tweets-are-a-hit-with-twitters-users`. [Accessed 09-09-2024]. 4.3.1

[146] X. G. Xu Han and S. Peng, "Analysis of tweet form's effect on users' engagement on twitter," *Cogent Business & Management*, vol. 6, no. 1, p. 1564168, 2019. 4.3.1

[147] Wikipedia contributors, "Scunthorpe problem — Wikipedia, The Free Encyclopedia," 2024. [Online; accessed 6-October-2024]. 4.3.1

[148] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, and M. Zhou, "Mind: A large-scale dataset for news recommendation," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3597–3606, Association for Computational Linguistics, 2020. 4.3.2

[149] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019. 4.3.3

[150] N. Muennighoff, "Sgpt: Gpt sentence embeddings for semantic search," *ArXiv*, vol. abs/2202.08904, 2022. 4.3.4

[151] R. Subhashini and V. J. S. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," in *2010 First International Conference on Integrated Intelligent Computing*, pp. 27–31, IEEE, 2010. 4.3.4

[152] L. Muflikhah and B. Baharudin, "Document clustering using concept space and cosine similarity measurement," in *2009 International Conference on Computer Technology and Development*, vol. 1, pp. 58–62, IEEE, 2009. 4.3.4

[153] D. Irani, S. Webb, C. Pu, and K. Li, "Study of trend-stuffing on twitter through text classification," in *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, Citeseer, 2010. 4.5.1

[154] L. Almurqren, R. Hodgson, and A. Cristea, "Arabic text sentiment analysis: Reinforcing human-performed surveys with wider topic analysis," *arXiv preprint arXiv:2403.01921*, 2024. 5.1, 5.2.2, 5.2.2, 5.2.3, 8.2

[155] R. Hodgson, A. Cristea, L. Shi, and J. Graham, "Wide-scale automatic analysis of 20 years of its research," in *Intelligent Tutoring Systems* (A. I. Cristea and C. Troussas, eds.), (Cham), pp. 8–21, Springer International Publishing, 2021. 5.1, 5.3

[156] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data*,

vol. 6, no. 1, pp. 1–18, 2019. 5.2, 5.2.2, 5.3, 5.3.3, 5.3.4, 5.4.3, 5.5, 6.3.1, 6.3.1, 6.9, 8.2

[157] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, 2015. 5.2, 5.3.1, 7.6.3

[158] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 100–108, 2010. 5.3, 7.6.3

[159] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *Journal of Information Science*, vol. 43, no. 1, pp. 88–102, 2017. 5.3

[160] D. Dermeval, R. Paiva, I. I. Bittencourt, J. Vassileva, and D. Borges, "Authoring tools for designing intelligent tutoring systems: a systematic review of the literature," *International Journal of Artificial Intelligence in Education*, vol. 28, pp. 336–384, 2018. 5.3.1

[161] B. D. Nye, "Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context," *International Journal of Artificial Intelligence in Education*, vol. 25, pp. 177–203, 2015. 5.3.1

[162] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999. 5.3.1

[163] A. Panichella, B. Dit, R. Oliveto, M. Di Penta, D. Poshynanyk, and A. De Lucia, "How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms," in *2013 35th International conference on software engineering (ICSE)*, pp. 522–531, IEEE, 2013. 5.3.1

[164] M. J. B. Yee Whye Teh, Michael I Jordan and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006. 5.3.1

[165] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018. 5.3.1, 5.3.4, 6.3.1

[166] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013. 5.3.1, 7.4.2, 7.7

[167] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine learning proceedings 1995*, pp. 331–339, Elsevier, 1995. 5.3.1, 5.3.2, 6.3.1

[168] X. Wang, H. Edison, D. Khanna, and U. Rafiq, “How many papers should you review? a research synthesis of systematic literature reviews in software engineering,” in *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 1–6, IEEE, 2023. 5.3.2

[169] X. Huang, Z. Li, C. Wang, and H. Ning, “Identifying disaster related social media for rapid response: a visual-textual fused cnn architecture,” *International Journal of Digital Earth*, 2020. 5.3.3

[170] D. Jurafsky and J. H. Martin, “Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.” 5.3.3

[171] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”, 2009. 5.3.3

[172] A. Schofield, M. Magnusson, L. Thompson, and D. Mimno, “Understanding text pre-processing for latent dirichlet allocation,” in *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, vol. 2, pp. 432–436, 2017. 5.3.3

[173] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111–3119. 5.3.4

[174] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015. 5.3.4

[175] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019. 5.3.4, 6.3

[176] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987. 5.3.4

[177] R. M. Taylor and J. A. du Preez, “Simlda: A tool for topic model evaluation,” in *Proceedings of the Future Technologies Conference*, pp. 534–554, Springer, 2022. 5.4.1, 7.6.3

[178] P. Brusilovsky, “Adaptive hypermedia: From intelligent tutoring systems to web-based education,” in *International Conference on Intelligent Tutoring Systems*, pp. 1–7, Springer, 2000. 5.4.2

[179] J. Noguera, F. E. Ayeni, S. Okuboyejo, and S. Adusumilli, “Towards a web based adaptive and intelligent tutoring system,” *International Journal of Computing and Informatics (IJCANDI)*, vol. 1, no. 1, 2017. 5.4.2

[180] B. L. Bayasut, G. Pramudya, and H. B. Basiron, “Ulul-ilm: The design of web-based adaptive educational hypermedia system based on learning style,” in *2013 13th International Conference on Intellient Systems Design and Applications*, pp. 147–152, IEEE, 2013. 5.4.2

[181] B. L. Bayasut, G. Pramudya, and H. Basiron, “The application of multi layer feed forward artificial neural network for learning style identification,” *Advanced Science Letters*, vol. 20, no. 10-11, pp. 2180–2183, 2014. 5.4.2

[182] J. Mota, “Using learning styles and neural networks as an approach to elearning content and layout adaptation,” in *Doctoral Symposium on Informatics Engineering*, 2008. 5.4.2

[183] S. Chimalakonda and K. V. Nori, “An ontology based modeling framework for design of educational technologies,” *Smart learning environments*, vol. 7, no. 1, pp. 1–24, 2020. 5.4.2

[184] E. Gouli, H. Kornilakis, K. Papanikolaou, and M. Grigoriadou, “Adaptive assessment improving interaction in an educational hypermedia system,” in *Proceedings of the PanHellenic Conference with International Participation in Human-Computer Interaction*, pp. 217–222, 2001. 5.4.2

[185] P. Jordan, P. Albacete, M. J. Ford, S. Katz, M. Lipschultz, D. Litman, S. Silliman, and C. Wilson, “Interactive event: The rimac tutor-a simulation of the highly interactive nature of human tutorial dialogue,” in *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*, pp. 928–929, Springer, 2013. 5.4.2

[186] H. C. Lane and K. VanLehn, “A dialogue-based tutoring system for beginning programming.,” in *FLAIRS Conference*, pp. 449–454, Citeseer, 2004. 5.4.2

[187] R. Moreno, R. Mayer, and J. Lester, “Life-like pedagogical agents in constructivist multimedia environments: Cognitive consequences of their interaction,” in *EdMedia+ Innovate Learning*, pp. 776–781, Association for the Advancement of Computing in Education (AACE), 2000. 5.4.2

[188] M. Moundridou and M. Virvou, “Evaluating the impact of interface agents in an intelligent tutoring systems authoring tool,” in *Proceedings of the Panhellenic Conference with International participation in Human-Computer interaction*, 2001. 5.4.2

[189] C.-Y. Chou, T.-W. Chan, and C.-J. Lin, “Redefining the learning companion: the past, present, and future of educational agents,” *Computers & Education*, vol. 40, no. 3, pp. 255–269, 2003. 5.4.2

[190] A. L. Baylor and Y. Kim, “Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role,” in *International conference on intelligent tutoring systems*, pp. 592–603, Springer, 2004. 5.4.2

[191] O. K. Akputu, K. P. Seng, Y. Lee, and L.-M. Ang, "Emotion recognition using multiple kernel learning toward e-learning applications," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, pp. 1–20, 2018. 5.4.2

[192] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. de Jesus Filipe, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," in *International Conference on Technology and Innovation in Learning, Teaching and Education*, pp. 52–68, Springer, 2022. 5.4.2

[193] T. W. Liew, N. A. Mat Zin, and N. Sahari, "Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment," *Human-centric Computing and Information Sciences*, vol. 7, no. 1, pp. 1–21, 2017. 5.4.2

[194] C. Limongelli, F. Sciarrone, and G. Vaste, "Personalized e-learning in moodle: the moodle_ls system," *Journal of e-Learning and Knowledge Society*, vol. 7, no. 1, pp. 49–58, 2011. 5.4.2

[195] A. I. Mørch, S. Jondahl, and J. A. Dolonen, "Supporting conceptual awareness with pedagogical agents," *Information Systems Frontiers*, vol. 7, pp. 39–53, 2005. 5.4.2

[196] R. Monson, D. Bunney, and T. Lawrence, "Moocs, learning analytics and learning advisors," *eCULTURE*, vol. 6, no. 1, p. 3, 2013. 5.4.2

[197] D. F. O. Onah, E. L. Pang, J. E. Sinclair, and J. Uhomoibhi, "Learning analytics for motivating self-regulated learning and fostering the improvement of digital mooc resources," in *Mobile Technologies and Applications for the Internet of Things: Proceedings of the 12th IMCL Conference*, pp. 14–21, Springer, 2019. 5.4.2

[198] G. Alexandron, L. Y. Yoo, J. A. Ruipérez-Valiente, S. Lee, and D. E. Pritchard, "Are mooc learning analytics results trustworthy? with fake learners, they might not be!," *International journal of artificial intelligence in education*, vol. 29, pp. 484–506, 2019. 5.4.2

[199] T. Bystrova, V. Larionova, E. Sinitsyn, and A. Tolmachev, "Learning analytics in massive open online courses as a tool for predicting learner performance," , no. 4 (eng), pp. 139–166, 2018. 5.4.2

[200] A. Alamri, M. Alshehri, A. Cristea, F. D. Pereira, E. Oliveira, L. Shi, and C. Stewart, "Predicting moocs dropout using only two easily obtainable features from the first week's activities," in *Intelligent Tutoring Systems: 15th International Conference, ITS 2019, Kingston, Jamaica, June 3–7, 2019, Proceedings 15*, pp. 163–173, Springer, 2019. 5.4.2

[201] J. Gardner and C. Brooks, "Dropout model evaluation in moocs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018. 5.4.2

[202] M. Grigoriadou, H. Kornilakis, K. A. Papanikolaou, and G. D. Magoulas, “Fuzzy inference for student diagnosis in adaptive educational hypermedia,” in *Methods and Applications of Artificial Intelligence: Second Hellenic Conference on AI, SETN 2002 Thessaloniki, Greece, April 11–12, 2002 Proceedings 2*, pp. 191–202, Springer, 2002. 5.4.2

[203] F. Sgrò, P. Mango, S. Pignato, A. L. Piccolo, S. Nicolosi, R. Schembri, and M. Lipoma, “A neuro-fuzzy approach for student module of physical activity its,” *Procedia-Social and Behavioral Sciences*, vol. 9, pp. 189–193, 2010. 5.4.2

[204] M. Gumiñska and J. Madejski, “Assessment of the didactic measurement results using fcm type networks,” *Archives of Materials Science*, vol. 46, p. 46, 2009. 5.4.2

[205] E. Naganathan, R. Venkatesh, and N. U. Maheswari, “Intelligent tutoring system: Predicting students results using neural networks.,” *J. Convergence Inf. Technol.*, vol. 3, no. 3, pp. 22–26, 2008. 5.4.2

[206] H. P. Maffon, J. S. Melo, T. A. Morais, P. B. Klavdianos, L. M. Brasil, T. L. Amaral, and G. Curilem, “Architecture of an intelligent tutoring system applied to the breast cancer based on ontology, artificial neural networks and expert systems,” in *The Sixth International Conference on Advances in Computer-Human Interactions, ACHI*, Citeseer, 2013. 5.4.2

[207] M. A. Hogo, “Evaluation of e-learning systems based on fuzzy clustering models and statistical tools,” *Expert systems with applications*, vol. 37, no. 10, pp. 6891–6903, 2010. 5.4.2

[208] M. Soliman and C. Guetl, “Simulating interactive learning scenarios with intelligent pedagogical agents in a virtual world through bdi-based agents,” 2013. 5.4.2

[209] C. Buche, R. Querrec, P. De Loor, and P. Chevaillier, “Mascaret: A pedagogical multi-agent system for virtual environments for training,” *International Journal of Distance Education Technologies (IJDET)*, vol. 2, no. 4, pp. 41–61, 2004. 5.4.2

[210] L.-P. Huang and C.-L. Ho, “Building and adaptive learning mechanism to assist elearning students,” 2009. 5.4.2

[211] V. Furtado Filho, “J.: A multi-agent simulator for teaching police allocation,” in *Proceedings of the National Conference on Artificial Intelligence*, pp. 1521–1528, 2005. 5.4.2

[212] C. E. Giuffra and R. A. Silveira, “An agent based model for integrating intelligent tutoring system and virtual learning environments,” in *Advances in Artificial Intelligence–IBERAMIA 2012: 13th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 13-16, 2012. Proceedings 13*, pp. 641–650, Springer, 2012. 5.4.2

[213] G. Papagiannakis, P. Zikas, N. Lydatakis, S. Kateros, M. Kentros, E. Geronikolakis, M. Kamarianakis, I. Kartsonaki, and G. Evangelou, “Mages 3.0: Tying the knot of medical vr,” in *ACM SIGGRAPH 2020 Immersive Pavilion*, pp. 1–2, 2020. 5.4.2

[214] A. Terracina, R. Berta, F. Bordini, R. Damilano, and M. Mecella, “Teaching stem through a role-playing serious game and intelligent pedagogical agents,” in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pp. 148–152, IEEE, 2016. 5.4.2

[215] A. Steinmaurer, J. Pirker, and C. Gütl, “scool-game based learning in stem education: a case study in secondary education,” in *The Challenges of the Digital Transformation in Education: Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018)-Volume 1*, pp. 614–625, Springer, 2020. 5.4.2

[216] D. Hooshyar, R. Binti Ahmad, M. Wang, M. Yousefi, M. Fathi, and H. Lim, “Development and evaluation of a game-based bayesian intelligent tutoring system for teaching programming,” *Journal of Educational Computing Research*, vol. 56, no. 6, pp. 775–801, 2018. 5.4.2

[217] S. Saha, T. I. Dhamecha, S. Marvaniya, P. Foltz, R. Sindhgatta, and B. Sengupta, “Joint multi-domain learning for automatic short answer grading,” *arXiv preprint arXiv:1902.09183*, 2019. 5.4.2

[218] M. Cozma, A. M. Butnaru, and R. T. Ionescu, “Automated essay scoring with string kernels and word embeddings,” *arXiv preprint arXiv:1804.07954*, 2018. 5.4.2

[219] M. Liu, Y. Wang, W. Xu, and L. Liu, “Automated scoring of chinese engineering students’ english essays,” *International Journal of Distance Education Technologies (IJDET)*, vol. 15, no. 1, pp. 52–68, 2017. 5.4.2

[220] A. N. Bhat, *Sketchography-Automatic Grading of Map Sketches for Geography Education*. PhD thesis, 2017. 5.4.2

[221] Y. Hou, P. Zhou, T. Wang, L. Yu, Y. Hu, and D. Wu, “Context-aware online learning for course recommendation of mooc big data,” *arXiv preprint arXiv:1610.03147*, 2016. 5.4.2

[222] V. Demertzi and K. Demertzis, “A hybrid adaptive educational elearning project based on ontologies matching and recommendation system,” *arXiv preprint arXiv:2007.14771*, 2020. 5.4.2

[223] S. Kellogg and A. Edelmann, “Massively open online course for educators (mooc-e d) network dataset,” *British Journal of Educational Technology*, vol. 46, no. 5, pp. 977–983, 2015. 5.4.2

[224] R. Hodgson, J. Wang, A. Cristea, F. Matsuzaki, and H. Kubota, "A topic-centric crowdsourced assisted biomedical literature review framework for academics," in *Proceedings of the 15th International Conference on Educational Data Mining*, p. 652, 2022. 6.1, 6.3.1, 6.5, 6.6, 6.6.1

[225] A. E. Jinha, "Article 50 million: an estimate of the number of scholarly articles in existence," *Learned publishing*, vol. 23, no. 3, pp. 258–263, 2010. 6.3

[226] Y. Hu, Z. Yu, X. Cheng, Y. Luo, and C. Wen, "A bibliometric analysis and visualization of medical data mining research," *Medicine*, vol. 99, no. 22, 2020. 6.3

[227] H. Kilicoglu, F. Ensan, B. McInnes, and L. L. Wang, "Semantics-enabled biomedical literature analytics," *Journal of Biomedical Informatics*, p. 104588, 2024. 6.3

[228] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics*, vol. 18, no. 12, pp. 1553–1561, 2002. 6.3

[229] J. Nicolas, *Artificial Intelligence and Bioinformatics*, pp. 209–264. Cham: Springer International Publishing, 2020. 6.3, 6.4

[230] C. Rodríguez-Penagos, H. Salgado, I. Martínez-Flores, and J. Collado-Vides, "Automatic reconstruction of a bacterial regulatory network using natural language processing," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–11, 2007. 6.3

[231] R. Islamaj Doğan, S. Kim, A. Chatr-aryamontri, C. S. Chang, R. Oughtred, J. Rust, W. J. Wilbur, D. C. Comeau, K. Dolinski, and M. Tyers, "The BioC-BioGRID corpus: full text articles annotated for curation of protein–protein and genetic interactions," *Database*, vol. 2017, 01 2017. baw147. 6.3

[232] K. Raja, M. Patrick, Y. Gao, D. Madu, Y. Yang, and L. C. Tsoi, "A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries," *International Journal of Genomics*, vol. 2017, p. 6213474, Feb 2017. 6.3

[233] Y. Pu, D. Beck, and K. Verspoor, "Graph embedding-based link prediction for literature-based discovery in alzheimer's disease," *Journal of Biomedical Informatics*, vol. 145, p. 104464, 2023. 6.3

[234] L. D. Dang, U. T. Phan, and N. T. Nguyen, "Gena: A knowledge graph for nutrition and mental health," *Journal of Biomedical Informatics*, vol. 145, p. 104460, 2023. 6.3

[235] M. Pérez-Pérez, T. Ferreira, G. Igrejas, and F. Fdez-Riverola, "A novel gluten knowledge base of potential biomedical and health-related interactions extracted from the literature: Using machine learning and graph analysis methodologies to reconstruct the bibliome," *Journal of Biomedical Informatics*, vol. 143, p. 104398, 2023. 6.3

[236] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021. 6.3

[237] A. Nentidis, T. Chatzopoulos, A. Krithara, G. Tsoumakas, and G. Paliouras, "Large-scale investigation of weakly-supervised deep learning for the fine-grained semantic indexing of biomedical literature," *Journal of Biomedical Informatics*, vol. 146, p. 104499, 2023. 6.3

[238] A. Khader and F. Ensan, "Learning to rank query expansion terms for covid-19 scholarly search," *Journal of Biomedical Informatics*, vol. 142, p. 104386, 2023. 6.3

[239] K. Roberts, T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff, E. Voorhees, L. L. Wang, and W. R. Hersh, "TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1431–1436, 07 2020. 6.3

[240] "Trends for open access to publications." 6.3

[241] F. Shu and V. Larivière, "The oligopoly of open access publishing," *Scientometrics*, vol. 129, pp. 519–536, Jan 2024. 6.3

[242] C. Zhang, L. Zhao, M. Zhao, and Y. Zhang, "Enhancing keyphrase extraction from academic articles with their reference information," *Scientometrics*, vol. 127, no. 2, pp. 703–731, 2022. 6.3

[243] Y. Wang, C. Zhang, and K. Li, "A review on method entities in the academic literature: extraction, evaluation, and application," *Scientometrics*, vol. 127, no. 5, pp. 2479–2520, 2022. 6.3

[244] T. S, B. J, and G. T.V., "Semi-supervised bootstrapping approach for named entity recognition," *ArXiv*, vol. abs/1511.06833, 2015. 6.3

[245] A. Zheng, H. Zhao, Z. Luo, C. Feng, X. Liu, and Y. Ye, "Improving on-line scientific resource profiling by exploiting resource citation information in the literature," *Information Processing Management*, vol. 58, no. 5, p. 102638, 2021. 6.3

[246] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen, "Biomedical text mining and its applications in cancer research," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 200–211, 2013. 6.3

[247] L. Qing, W. Linhong, and D. Xuehai, "A novel neural network-based method for medical text classification," *Future Internet*, vol. 11, no. 12, 2019. 6.3

[248] Y. Bao, Z. Deng, Y. Wang, H. Kim, V. D. Armengol, F. Acevedo, N. Ouardaoui, C. Wang, G. Parmigiani, R. Barzilay, D. Braun, and K. S. Hughes, "Using machine learning and natural language processing to review and classify the medical literature on cancer susceptibility genes," *JCO Clinical Cancer Informatics*, no. 3, pp. 1–9, 2019. PMID: 31545655. 6.3

[249] L. Soldaini and N. Goharian, "Quickumls: a fast, unsupervised approach for medical concept extraction," in *MedIR workshop, sigir*, pp. 1–4, 2016. 6.3

[250] H. Ding and X. Luo, "Attention-based unsupervised keyphrase extraction and phrase graph for covid-19 medical literature retrieval," *ACM Trans. Comput. Healthcare*, vol. 3, oct 2021. 6.3

[251] J. Portenoy and J. D. West, "Constructing and evaluating automated literature review systems," *Scientometrics*, vol. 125, pp. 3233–3251, Dec 2020. 6.3

[252] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, p. 993–1022, mar 2003. 6.3.1, 6.3.1, 6.9

[253] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl_1, pp. 5228–5235, 2004. 6.3.1, 6.3.1

[254] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, (New York, NY, USA), p. 1980–1984, Association for Computing Machinery, 2012. 6.3.1

[255] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, "Content-driven detection of cyberbullying on the instagram social network.," in *IJCAI*, vol. 16, pp. 3952–3958, 2016. 6.3.1

[256] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Constrained lda for grouping product features in opinion mining," in *Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part I 15*, pp. 448–459, Springer, 2011. 6.3.1

[257] H. Wang, Y. Ding, J. Tang, X. Dong, B. He, J. Qiu, and D. J. Wild, "Finding complex biological relationships in recent pubmed articles using bio-lda," *PloS one*, vol. 6, no. 3, p. e17243, 2011. 6.3.1

[258] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. II–1188–II–1196, JMLR.org, 2014. 6.3.1

[259] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering.," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017. 6.3.1

[260] R. Hodgson, A. Cristea, L. Shi, and J. Graham, "Wide-scale automatic analysis of 20 years of its research," in *International Conference on Intelligent Tutoring Systems*, pp. 8–21, Springer, 2021. 6.3.1, 6.5, 6.6

[261] M. Saha and D. Logofătu, "An approach to predict optimal configurations for lda-based topic modeling," in *International Conference on Engineering Applications of Neural Networks*, pp. 17–27, Springer, 2024. 6.3.1

[262] K. Du, "Evaluating hyperparameter alpha of lda topic modeling.," in *DHd*, 2022. 6.3.1

[263] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, "The expertise reversal effect," *Educational Psychologist*, vol. 38, no. 1, pp. 23–31, 2003. 6.3.2, 6.5, 6.6, 6.7.7, 6.8.1

[264] A. Baddeley, "Working memory oxford: Oxford university press," *Clarendon Press*, vol. 666, p. 667, 1986. 6.3.2, 6.5, 6.8.1

[265] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information.," *Psychological review*, vol. 63, no. 2, p. 81, 1956. 6.3.2, 6.5, 6.8.1

[266] G. D. White, "Mental load: helping clinical learners," *The Clinical Teacher*, vol. 8, no. 3, pp. 168–171, 2011. 6.3.2

[267] D. Miller-Cotto and R. Gordon, "Revisiting working memory fifty years after baddeley and hitch: A review of field-specific conceptualizations, use and misuse, and paths forward for studying children," 6.3.2

[268] J. Sweller, J. J. Van Merrienboer, and F. G. Paas, "Cognitive architecture and instructional design," *Educational psychology review*, vol. 10, pp. 251–296, 1998. 6.3.2

[269] D. Gopher and R. Braune, "On the psychophysics of workload: Why bother with subjective measures?," *Human factors*, vol. 26, no. 5, pp. 519–532, 1984. 6.3.2, 6.8.1

[270] T. M. Jingyun Wang and T. Hoel, "Strategies for multimedia learning object recommendation in a language learning support system: Verbal learners vs. visual learners," *International Journal of Human–Computer Interaction*, vol. 35, no. 4-5, pp. 345–355, 2019. 6.3.2, 6.8.1

[271] J. Wang, T. Mendori, and J. Xiong, "A language learning support system using course-centered ontology and its evaluation," *Computers Education*, vol. 78, pp. 278–293, 2014. 6.3.2, 6.6, 6.7.7, 6.8.1

[272] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989. 6.3.2, 6.7.7

[273] R. Schultz, D. Slevin, E. Krannert Graduate School of Industrial Administration. Institute for Research in the Behavioral, and M. Sciences, *Implementation and Organizational Validity: An Empirical Investigation.* Reprint / Herman C. Krannert Graduate School of Industrial Administration of Purdue University, Lafayette, Ind, Krannert Graduate School of Industrial Administration, Purdue University, 1975. 6.3.2, 6.7.7

[274] D. Robey, "User attitudes and management information system use," *Academy of management Journal*, vol. 22, no. 3, pp. 527–538, 1979. 6.3.2, 6.7.7

[275] H.-C. Chu, G.-J. Hwang, C.-C. Tsai, and J. C. Tseng, "A two-tier test approach to developing location-aware mobile learning systems for natural science courses," *Computers Education*, vol. 55, no. 4, pp. 1618–1627, 2010. 6.3.2, 6.7.7

[276] L. Hammon and H. Hippner, "Crowdsourcing," *Business & Information systems engineering*, vol. 4, no. 3, pp. 163–166, 2012. 6.4, 8.2

[277] T. U. Consortium, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, pp. D480–D489, 11 2020. 6.4.1, 6.4.1, 6.4.1, 8.2

[278] M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, and R. D. Finn, "The InterPro protein families and domains database: 20 years on," *Nucleic Acids Research*, vol. 49, pp. D344–D354, 11 2020. 6.4.1, 6.4.1, 6.4.1, 8.2

[279] R. Hodgson, A. Cristea, L. Shi, and J. Graham, "Wide-scale automatic analysis of 20 years of its research," in *Intelligent Tutoring Systems* (A. I. Cristea and C. Troussas, eds.), (Cham), pp. 8–21, Springer International Publishing, 2021. 6.4.1

[280] J. Wang, A. Shimada, M. Oi, H. Ogata, and Y. Tabata, "Development and evaluation of a visualization system to support meaningful e-book learning," *Interactive Learning Environments*, vol. 31, no. 2, pp. 836–853, 2020. 6.6, 6.7.7

[281] K. Schwaber and J. Sutherland, *The Definitive Guide to Scrum: The Rules of the Game.* Retrospectiva del Sprint de Nexus, Scrum.org., 2017. 6.6.1

[282] "Entrez help," May 2016. 6.7.2

[283] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018. 7.4.1

[284] A. Subakti, H. Murfi, and N. Hariadi, “The performance of bert as data representation of text clustering,” *Journal of big Data*, vol. 9, no. 1, pp. 1–21, 2022. 7.4.1

[285] J. MacQueen, “Classification and analysis of multivariate observations,” in *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297, 1967. 7.4.2

[286] C. Rasmussen, “The infinite gaussian mixture model,” *Advances in neural information processing systems*, vol. 12, 1999. 7.4.2

[287] A. Subasi, “Chapter 7 - clustering examples,” in *Practical Machine Learning for Data Analysis Using Python* (A. Subasi, ed.), pp. 465–511, Academic Press, 2020. 7.4.2

[288] J. Hull, “A database for handwritten text recognition research,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994. 7.4.2

[289] S. Grossberg, “Recurrent neural networks,” *Scholarpedia*, vol. 8, no. 2, p. 1888, 2013. revision #138057. 7.4.3

[290] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997. 7.4.3

[291] S. Fernández, A. Graves, and J. Schmidhuber, “An application of recurrent neural networks to discriminative keyword spotting,” in *Proceedings of the 17th International Conference on Artificial Neural Networks*, ICANN’07, (Berlin, Heidelberg), p. 220–229, Springer-Verlag, 2007. 7.4.3

[292] A. Graves and J. Schmidhuber, “Offline handwriting recognition with multidimensional recurrent neural networks,” in *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), vol. 21, Curran Associates, Inc., 2008. 7.4.3

[293] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks-signal processing, ieee transactions on,” 1998. 7.4.3

[294] A. Graves, N. Jaitly, and A. rahman Mohamed, “Hybrid speech recognition with deep bidirectional lstm,” *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013. 7.4.3

[295] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014. 7.4.3

[296] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015. 7.4.3

[297] Y. Lin, Y. Meng, X. Sun, Q. Han, K. Kuang, J. Li, and F. Wu, “Bertgcn: Transductive text classification by combining gcn and bert,” *arXiv preprint arXiv:2105.05727*, 2021. 7.5.1

[298] T. Jo, "Clustering news groups using inverted index based ntso," in *2009 First International Conference on Networked Digital Technologies*, pp. 1–7, IEEE, 2009. 7.5.1

[299] N. S. E. K. Jasila and K. A. A. Nazeer, "An efficient document clustering approach for devising semantic clusters," *Cybernetics and Systems*, vol. 0, no. 0, pp. 1–18, 2023. 7.5.1

[300] W. Xu, X. Jiang, S. H. Sengamedu, F. Iannacci, and J. Zhao, "vontss: vmf based semi-supervised neural topic modeling with optimal transport," *arXiv preprint arXiv:2307.01226*, 2023. 7.5.1

[301] J. Xu and G. Durrett, "Spherical latent spaces for stable variational autoencoders," *arXiv preprint arXiv:1808.10805*, 2018. 7.5.1

[302] A. B. Dieng, F. J. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. 7.5.1

[303] R. Wang, X. Hu, D. Zhou, Y. He, Y. Xiong, C. Ye, and H. Xu, "Neural topic modeling with bidirectional adversarial training," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 340–350, Association for Computational Linguistics, July 2020. 7.5.1

[304] J. Liang, L. Bai, C. Dang, and F. Cao, "The k-means-type algorithms versus imbalanced data distributions," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 4, pp. 728–745, 2012. 7.5.2

[305] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: a data distribution perspective," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 779–784, 2006. 7.5.2

[306] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2011. 7.5.2

[307] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004. 7.5.2

[308] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. 7.5.2

[309] O. Sitompul, E. Nababan, *et al.*, "Optimization model of k-means clustering using artificial neural networks to handle class imbalance problem," in *IOP*

conference series: materials science and engineering, vol. 288, p. 012075, IOP Publishing, 2018. 7.5.2

[310] Y. Qian, Y. Liang, M. Li, G. Feng, and X. Shi, "A resampling ensemble algorithm for classification of imbalance problems," *Neurocomputing*, vol. 143, pp. 57–67, 2014. 7.5.2

[311] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining and Knowledge Discovery*, vol. 17, pp. 225–252, 2008. 7.5.2

[312] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014. 7.5.2

[313] D. Cai, X. He, and J. Han, "Training linear discriminant analysis in linear time," in *2008 IEEE 24th international conference on data engineering*, pp. 209–217, IEEE, 2008. 7.5.3

[314] I. M. Johnstone and A. Y. Lu, "Sparse principal components analysis," *arXiv preprint arXiv:0901.4392*, 2009. 7.5.3

[315] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted voronoi diagrams and randomization to variance-based k-clustering," in *Proceedings of the tenth annual symposium on Computational geometry*, pp. 332–339, 1994. 7.5.3

[316] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, (Red Hook, NY, USA), p. 2546–2554, Curran Associates Inc., 2011. 7.5.4

[317] Y. Ozaki, Y. Tanigaki, S. Watanabe, and M. Onishi, "Multiobjective tree-structured parzen estimator for computationally expensive optimization problems," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, GECCO '20, (New York, NY, USA), p. 533–541, Association for Computing Machinery, 2020. 7.5.4

[318] Y. Ozaki, Y. Tanigaki, S. Watanabe, M. Nomura, and M. Onishi, "Multiobjective tree-structured parzen estimator," *Journal of Artificial Intelligence Research*, vol. 73, pp. 1209–1250, 2022. 7.5.4

[319] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019. 7.5.4

[320] K. Fukushima, "Visual feature extraction by a multilayered network of analog threshold elements," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, no. 4, pp. 322–333, 1969. 7.5.5

[321] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947. 7.5.6, 7.5.7

[322] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957. 7.5.6

[323] M. Cohen and J. Arthur, "Randomization analysis of dental data characterized by skew and variance heterogeneity," *Community Dentistry and Oral Epidemiology*, vol. 19, no. 4, pp. 185–189, 1991. 7.5.7

[324] J. R. Nuñez, C. R. Anderton, and R. S. Renslow, "Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data," *PLOS ONE*, vol. 13, pp. 1–14, 08 2018. 7.5.8

[325] X. Ying, "An overview of overfitting and its solutions," in *Journal of physics: Conference series*, vol. 1168, p. 022022, IOP Publishing, 2019. 7.5.8

[326] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004. 7.5.8

[327] T. Joachims *et al.*, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization," in *ICML*, vol. 97, pp. 143–151, Citeseer, 1997. 7.6.2

[328] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, vol. 30, pp. 31–40, 2009. 7.6.3

[329] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*, pp. 13–22, 2013. 7.6.3

[330] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. 7.6.3

[331] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Advances in neural information processing systems*, vol. 22, 2009. 7.6.3

[332] "Parametric (neural network) Embedding &x2014; umap 0.5 documentation — umap-learn.readthedocs.io." `https://umap-learn.readthedocs.io/en/latest/parametric_umap.html`. [Accessed 11-03-2024]. 7.6.4

[333] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, 2001. 7.7

[334] Z. Li, E. Wallace, S. Shen, K. Lin, K. Keutzer, D. Klein, and J. Gonzalez, "Train big, then compress: Rethinking model size for efficient training and inference of transformers," in *International Conference on machine learning*, pp. 5958–5968, PMLR, 2020. 7.7

[335] B. Zhuang, J. Liu, Z. Pan, H. He, Y. Weng, and C. Shen, “A survey on efficient training of transformers,” *arXiv.org*, vol. abs/2302.01107, 2023. 7.7

[336] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, pp. 226–231, 1996. 7.7

[337] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019. 8.1

[338] S. Khan and M. Safyan, “Semantic matching in hierarchical ontologies,” *Journal of King Saud University - Computer and Information Sciences*, vol. 26, no. 3, pp. 247–257, 2014. 8.2

[339] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, “Visualizing high-dimensional data: Advances in the past decade,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1249–1268, 2017. 8.4

APPENDIX A

Participant Information Form

General Information

The aim of this research is to evaluate the effectiveness of a visualisation system developed by this project, quantitative data including system user log and 1 survey response will be collected, based on your feedback. This will consist of your feedback to set questions in an online form, based on your experiences when using the medical literature analysis software.

We appreciate your interest in participating in this questionnaire. You have been invited to participate as you are a medical expert. Please read through this information before agreeing to participate (if you wish to) by ticking the 'yes' box below and put your signature.

You may ask any questions before deciding to take part by contacting the researcher (details below).

The Principal Researcher is Dr Jingyun Wang (*jingyun.wang@durham.ac.uk*), who is an assistant professor in the Department of Computer Science at Durham University. The team member of this project is her Ph.D. student Ryan Hodgson (*ryan.hodgson@durham.ac.uk*). This research is in collaboration with Dr Marie Shigematsu Locatelli from Kochi Hospital.

Do I have to take part?

Please note that participation is voluntary. If you do decide to take part, you may withdraw at any point for any reason.

We have included a 'Prefer not to say' option for each set of questions should you prefer not to answer a particular question.

What kind of data will be collected?

We will provide you a username and the corresponding password to access our literature review system developed by the research team of Dr Jingyun Wang. We will only collect where you click on the pages.

Your personal information data will **not** be collected. We will **not** collect any data that could directly identify you. Also, your IP address will **not** be stored.

After working with our system, you will be asked to complete a survey based on your experience when working with our system. The survey should only take about a few minutes.

How will my data be used?

The data collected in this survey will contribute to identifying how the literature review system can be improved, and findings of this study may be published as part of an academic Journal.

We will take all reasonable measures to ensure that data remain confidential.

The survey responses you provide will be stored in a password-protected, encrypted, electronic file on Durham University secure servers and may be used in academic publications. Research data will be stored for 1 year after publication or public release of the work of the research.

Who will have access to my data?

Durham University is the data controller with respect to the data and, as such, will determine how the data is used in the research. Further information about your rights with respect to the data is available from https://www.durham.ac.uk/about-us/governance/information-governance/privacy-notices/generic-privacy-notice/.

The data you provide may be shared with team member of this study, and may be published in an anonymised format as part of an academic work.

The results will be written up for an academic publication, and subsequent PhD thesis.

Who has reviewed this research?

This research has been reviewed by, and received ethics clearance through, Durham University , Department of Computer Science Ethics Committee.

Who do I contact if I have a concern, or I wish to complain?

If you have a concern about any aspect of this research, please speak to Dr Jingyun Wang (*jingyun.wang@durham.ac.uk*), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with.

Anyone wishing to raise a concern regarding matters of Research Integrity at the University should write in confidence to the PVC Research (pvc.research@durham.ac.uk).

Please note that you may only participate in this survey if you are 18 years of age or over.

□ I certify that I am 18 years of age or over.

If you have read the information above and agree to participate with the understanding that the data (including any personal data) you submit will be processed accordingly, please tick the box below to start.

□ Yes, I agree to take part

Signature:

Date:

APPENDIX B

Questions Regarding Cognitive Load and Technology Acceptance

Table B.1: Questions addressing cognitive load.

	Question
1.	How much effort was required to understand the purpose of using this system?
2.	How much effort was required to label topics, based on the information provided by the system?
3.	When viewing the graphs on the topic details page, how much does the system distract you from finding relevant information?
4.	When viewing the overall topic details page for labelling topics, how discouraged, stressed or irritated did you feel?

Table B.2: Questions addressing technology acceptance.

	Question
1.	The following feature graphs were easy to use. (This question was followed with Figure 6.9.)
2.	The above feature graphs were useful.
3.	The labelling of the publication list is simple to use. (This question was followed with Figure 6.10.)
4.	The labelling based on publication list was useful.
5.	The following relations map which can display the relevant papers was easy to use. (This question was followed with Figure 6.5.)
6.	The above relation map which can display the relevant papers was useful.