University of Southampton

Faculty of Physical Sciences and Engineering

School of Electronics and Computer Science

Insights from Heterogeneous Data

Through Transitive Semantic Relationships and Text Analytics

by

David Ralph

ORCID ID 0000-0003-3385-9295

Thesis for the degree of Doctor of Philosophy

May 2022

University of Southampton

Abstract

Faculty of Physical Sciences and Engineering

School of Electronics and Computer Science

Doctor of Philosophy

Insights from Heterogeneous Data

Through Transitive Semantic Relationships and Text Analytics

by

David Ralph

Many organisations are finding that the volume of information they need to analyse to make effective decisions is increasing. An important element in effective decision making is the ability to prioritise information quickly and accurately from a variety of sources. Technology tools are widely used to aid decision making through analysis and visualisation of numeric data, leveraging structured knowledge as in expert systems, or identifying items based on known existing relationships and content information as in recommender systems. However, producing similar insights from unstructured text documents of varying formats, intents, and domains, with little prior knowledge or labelling, remains an open problem.

This thesis takes the approach of using machine understanding of natural language text and the semantic content of documents as the basis for downstream tasks of recommendation, visualisation, summarisation, clustering, and topic naming to highlight key areas of interest in large heterogeneous datasets. The approach builds on both traditional techniques and recent advances in machine learning and natural language processing and combines and supplements them to address issues including sparse labelling, the cold-start problem, and the explainability of results. A novel recommendation algorithm, Transitive Semantic Relationships (TSR) is proposed to address challenging cases of the cold-start problem and is demonstrated as an effective tool for identifying supply chain relationships using company descriptions and a small number of known relationships. For the more general problem of finding meaning in large collections of unstructured text, this thesis proposes and demonstrates a methodology for combining several existing text analytics techniques to produce an overview of the distribution and typical content of key topics present in the data. This method is demonstrated for varied examples including a survey of experts concerns regarding the COVID-19 pandemic in the United Kingdom, the descriptions of businesses on the Isle of Wight, and the descriptions of 2500 TED talks. A web-based tool, the Text Insights Pipeline (TIP) is presented enabling non-experts to make use of this approach for analysis of other collections of unstructured text.

This thesis concludes that semantic understanding of text through deep learning coupled with explainable downstream algorithms is an effective basis for producing explainable insights and representative overviews of large unstructured text datasets. The contributions of this thesis have already seen adoption in industry, government, and research, and have the potential for making previously indigestible datasets open to analysis by aiding in the presentation and organisation of unstructured text data.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

Research Thesis: Declaration of Authorship

Print name:

David Ralph

Title of thesis:

Insights from Heterogeneous Data Through Transitive Semantic Relationships and Text Analytics

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Ralph, D., Li, Y., Wills, G., & Green, N. G. (2020). Recommendations from cold starts in big data. *Computing*, *102*(6), 1323–1344. https://doi.org/10.1007/s00607-020-00792-y

Ralph, D., Li, Y., Wills, G., & Green, N. G. (2019). Recommendations from Cold Starts in Big Data. In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security* (pp. 185–194). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0007798801850194

Signature: ..... Date: 17/05/2022

Acknowledgements

I thank the following people for helping with this research project:

Project supervisors Dr Gary B. Wills and Dr Nicolas G. Green (University of Southampton),

Industrial collaborator Dr Yunjia Li (Launch International LTD),

Parliamentary collaborator Dr Rowena Bermingham (POST, UK Parliament),

Various contacts that helped in the research proposal and commencement of the project: David Patterson, Dr Phil Jewell, Sally Thompson,

All of the friends and family that have supported me throughout the project, most prominently: Eunice Ralph, Ryan Thickett, Thomas Hewett

Definitions and Abbreviations

Natural Language Processing (NLP)	An area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, such as written and spoken English.
Syntax/Syntactic	Relating to written form.
Semantic	Relating to meaning.
Word/Sentence/Document - Vectors/Embeddings	Fixed length numeric vector representations of words, sentences, or documents. The terms 'vectors' and 'embeddings' are used interchangeably in the literature.
Semantic space/distributional semantics	A many-dimensional space for plotting word, sentence, or document embeddings.
Neural language model	A natural language model based on machine learning using neural networks.
Deep learning model	A neural network with a hierarchy of multiple hidden layers for direct learning of features.
Recurrent Neural Network (RNN), Sequence-to-sequence model	A deep learning model with an encoder-decoder architecture for learning important features in the sequence.
Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU)	Techniques used in RNNs to enable the network to remember parameters learned earlier in the sequence.
Transformer model	A deep learning model that uses attention to weight significance of parts of the input, similar to a RNN but without requiring processing of sequences in order.
Parliamentary Office of Science and Technology (POST)	A bicameral body within the UK Parliament. POST produces impartial, non-partisan, and peer-reviewed briefings, designed to make scientific research accessible to the UK Parliament.
COVID-19 Pandemic	The COVID-19 pandemic in the United Kingdom was part of the worldwide pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus reached the UK in late January 2020.
Standard Deviation (SD)	The square root of the variance. Low values indicate values tend to be close to the mean (expected value).

Chapter 1 **Introduction**

Informed decision-making benefits from the consideration of a large pool of information from varied sources. In many areas the amount of data available to inform decisions is increasing, however, the ability of human decision-makers and analysts to consume, interpret, and organise this data has not scaled with the amount available and required for effective decision making.

Computation and information systems have long been used to record, store, and collate data. From early examples of census data processing to the modern prevalence of big data analytics, data mining, and applied artificial intelligence, a consistent theme is that the value of data is in producing information and insights to inform human decisions. Computer systems can retain and accurately reproduce vast quantities of data, far exceeding what can be held in memory and considered by human analysts. For this reason, it is necessary to present not just data, but information and insights automatically distilled from data so that it can be effectively and expediently interpreted by humans.

Computer systems excel at processing quantitative and structured data, but qualitative, informal, and unstructured data such as natural language text poses additional challenges. For natural texts, such as articles, blogs, descriptions, and free-text survey responses, there is no simple objective method to identify key properties of the data, such as equivalent, typical, outlier, or significant values. As such, it is necessary to generate quantitative representations of text which capture its semantic meaning in order to produce information and insights through computational reasoning, distillation, and visualisation.

Quantitative representations of the semantics of text have been made possible through advances in natural language processing (NLP), machine learning, and text analytics. These representations can be used for a variety of downstream tasks, however, how to best make use of these techniques and present the results in a way that is not just interpretable and useful to humans, but also transparent and trustworthy, and robust to a range of challenging scenarios, such as limited prior knowledge (as in the cold start problem) and uncertainty (as in implicit feedback), remains a challenging research problem.

1.1 Problem Statement

Information contained in large collections of unstructured text is highly valuable for informed decision making, however, the size and lack of organisation and structure in these datasets are prohibitive to humans thoroughly examining and reasoning about the entire content. As such, automated approaches are desirable for identifying key information, structuring the data, and generating insights. To produce these results, an automated system must understand the meaning and significance of the text, such as through natural language understanding and text analytics. These results must be presented and evidenced in a way that is interpretable by humans so that the findings can be easily understood and communicated and so that decision-makers can have confidence in the validity of the results.

1.2 Research Question

The problem stated in section 1.1 leads to the following research question:

Research Question ..... How can machine understanding of text be used to produce insights from large collections of unstructured text to inform decision-makers, analysts, and organisations?

This can be broken down into the following Sub-Research-Questions (SRQ)s:

SRQ1..... How can machine understanding of text be used to identify relationships between documents in large collections of unstructured text?

SRQ2..... How can machine understanding of text be used to produce an interpretable overview of large collections of unstructured text?

SRQ3..... How can the results of text analysis be effectively presented and used to inform decision-makers, analysts, and organisations?

SRQ1 concerns the discovery of potential relationships between documents in a collection. This can be treated as a search or recommendation problem, where inferences can be made combining existing knowledge with machine understanding of text to infer new relationships or identify documents of interest. This question is examined in detail in Chapter 3.

SRQ2 concerns methods of distilling large datasets into summaries, visualisations, and categories so that the data can be presented in a structured way that provides a concise but representative overview of the data. This question is examined in detail in Chapter 4.

SRQ3 concerns the application of the results from the other two sub-questions, including how the results of automated analysis can be effectively communicated, explained, used as a starting point or aid to human analysis, and be compared with the results of traditional analysis methods. This question is relevant to the entire thesis but is discussed in particular regarding provenance and explainability in Chapter 3, presentation and comparison in Chapter 4, and application in Chapter 4 and Chapter 5.

1.3 Thesis Structure

Chapter 1 has introduced the motivation and aims of this research.

Chapter 2 reviews the literature on computational modelling of natural language, as well as search and recommender systems, and examines the types of data and challenges in these areas, and some prominent models and techniques.

Chapter 3 directly examines SRQ1, introducing the problem in detail, identifying the limitations of existing search and recommender systems regarding unstructured and unlabelled data and issues with explainability (SRQ3). The theory, implementation, and experimental results for a novel solution, the Transitive Semantic Relationships (TSR) model, are presented and followed by a further investigation into refinements to the model.

Chapter 4 focuses on SRQ2 and SRQ3, examining techniques for distilling, structuring, and presenting text data. A methodology is demonstrated that combines several models and techniques to produce interpretable analysis results including visualisations, summaries, and generated categories. This method is demonstrated on several datasets and is compared with an independent human analysis. This chapter also presents the Text Insights Pipeline (TIP), an automated analysis tool enabling non-experts to apply this methodology to other datasets.

Chapter 5 concludes the thesis and examines the contributions, impact, and future work resulting from this research.

Chapter 2 **Background**

2.1 Chapter Overview

This chapter sets out the essential background information and literature necessary to understand and contextualise the work presented in this thesis, in particular focusing on the topics of natural language data (2.2), how it can be understood and used by computer systems (2.3), existing approaches to identifying key items and relationships in large datasets (2.4), and the challenges with generating and presenting explainable results (2.5).

The techniques, algorithms, models, and challenges specifically related to approaches used in Chapter 3 and Chapter 4 which are less broadly relevant to the rest of this thesis are discussed in those respective chapters.

While the broader topics of machine learning and deep learning feature throughout this thesis, a detailed understanding of the fundamentals and model architectures should not be necessary to appreciate the work presented, although the interested reader may refer to textbooks on these topics for additional background (Goodfellow et al., 2016).

2.1.1 Scope and Limitations

As this project covers a wide range of areas, each with extensive background literature, a focus is given here on examining the particular challenges, approaches, and developments relevant to the theory and experimental work in subsequent chapters. In particular, some topics are mentioned but not discussed in detail such as natural language processing for languages other than English, semantic representations of media other than text, and approaches to relationship modelling that do not align with recommender systems. Each of these are large research areas tangentially related to some aspects of this project but are not explored in detail in this thesis in favour of more thoroughly examining a set of specific research areas including the challenges of modelling English language text from various sources, the cold start problem, and provenance and explainability of automated insights.

2.2 Natural Language Data

Natural language text presents a variety of challenges for computational analysis. These include syntactic issues with interpreting the written form of the text, such as spelling, handling of diacritics, and tokenisation; semantic issues with understanding the meaning of the text such as ambiguity, context, and intent; and social issues with judging the value and implication of text, such as factual correctness, bias, and search engine optimisation.

This section looks at various sources of data used in training and evaluating natural language models, some of the challenges they present, and how these are addressed. NLP is a large and highly active subject area, so this overview is not meant as an exhaustive list but instead focuses on representative examples of the types of data, challenges, and solutions, in particular where they are relevant to the methods used in this thesis.

2.2.1 Types and Sources of Data

Natural texts vary in length, from standalone sentences to books containing many hundreds of successive related sentences. For the purposes of NLP tasks, these can be divided into 'short-texts' and 'long-texts'. Short-texts include individual words and sentences, sentence fragments, nouns, descriptors (titles, etc.), and search terms (Wang et al., 2016; Yan et al., 2013). Whereas long-texts such as documents, articles, books, etc. are comprised of multiple ordered, related sentences, typically arranged into paragraphs and often sections, chapters, and volumes, depending on the size of the collection (Kiros et al., 2015).

Features can be learned from observing different properties of text, which vary in prominence between corpora. Natural language text from books, articles, and product descriptions, for example, consists of many ordered sentences typically structured as paragraphs but is typically unlabelled other than publication meta-data. Corpora of such texts are often the focus of research on text summarisation, question answering, content-based recommender systems, and information retrieval. They are also used to train or pre-train language models, discussed further in section 2.3.

Alternatively, a corpus of reviews typically consists of many items with differing lengths, often accompanied by a review score (either numeric or binary positive/negative). These corpora may feature less extensive text structures (due to shorter length in comparison to books or articles) but have quantitative labels in the form of user ratings/review scores. Corpora of user reviews are generally the focus of recommender systems, especially collaborative and hybrid recommender systems which require user ratings, discussed further in section 2.4.

Tasks such as sentiment analysis are commonly based on labelled data, such as reviews, where a positive score would indicate that the text is likely to have a positive sentiment. Such labelled data is sometimes community-sourced, such as reviews (Hu & Liu, 2004; Pang & Lee, 2005) and folksonomies (social tagging datasets) (Harper & Konstan, 2015; Hotho et al., 2006; Sen et al., 2006), but must otherwise be commissioned. In either case, as the data requires human annotation, collections are often smaller and/or more expensive to produce (per word or sentence) than unlabelled data from books and articles.

For short texts, particularly in problem spaces such as sentiment analysis, increasingly large amounts of labelled data are being generated due to social media and the proliferation of community-sourced reviews on the web (Google reviews, Amazon product reviews, etc.).

Long texts typically contain large numbers of both implicit and explicit relations to other concepts. These are sometimes linked explicitly at time of writing, for example, using hyperlinks, or following semantic web standards such as the Resource Description Framework (*RDF 1.1 Concepts and Abstract Syntax*, 2014) or Web Ontology Language (*OWL 2 Web Ontology Language Document Overview (Second Edition)*, 2012). Some datasets provide additional meta-data; Scientometric datasets for example associate items with citation counts and authorship which can be used as scoring mechanisms (Beel et al., 2016). Some types of data, such as books and articles, may include a title along with the long-text content.

2.2.2 Types and Cost of Labelling

As described in the previous section, many collections of natural language data are labelled as a usual part of their creation, such as reviews being accompanied by a numeric or binary (positive/negative) score and social media posts featuring tags. In these cases, the volume of user-generated labels is abundant. However, outside of these areas, it can be necessary to commission the labelling of data items.

The cost of labelling is highly dependent on the complexity of the task, specifically the time needed per human annotation and the expertise required. Snow et al., (2008) find that for tasks such as textual entailment and word sense disambiguation approximately four non-expert labels have similar quality to one expert label. Grady and Lease (2010) investigate crowdsourcing binary relevance labelling tasks and find that tasks where annotators must use item descriptions achieve poorer accuracy and require greater time per judgement than tasks using titles.

In some cases, datasets may be too large for comprehensive manual labelling and may only be viable to label by observing the behaviours of users of a service, such as a search engine or

recommender system. These observations may consist of click-through data, viewed items, or other types of implicit indicators that a result is relevant (Huang et al., 2013; Yin et al., 2016). Implicit labels are inherently uncertain as the interactions a user makes are situationally dependent, meaning users may investigate items that are not optimal results, and a user's search objective may change at any time, so it cannot be assumed that all user actions are related (Kong et al., 2015). However, aggregation of many implicit labels can be used to identify statistically significant trends. When evaluating using implicit labels, specialised methods such as those examined in Chapter 3 are often employed to account for the uncertainty in the evaluation cases.

2.2.3 Examples of NLP Datasets

NLP models are typically evaluated using several common benchmark datasets. Most of these datasets were created for, or are well-tailored to, specific problems in NLP. Labelled data allows for training and evaluation by calculating errors or misclassifications. When used for evaluation of language models (see section 2.3) these are commonly referred to as 'downstream tasks', in contrast to the learning task used to train the model (Cer et al., 2018; Conneau & Kiela, 2018).

Supervised models and models which use fine-tuning are typically trained on one of these datasets. Unsupervised and semi-supervised models are typically trained (or pre-trained) on a larger unlabelled corpus such as the 985 million words Toronto Book Corpus (Zhu et al., 2015), or the multi-billion word Wikipedia corpus and Common Crawl datasets (many versions exist for each), or various web news datasets.

Labelled datasets are sometimes provided with split training, validation (sometimes called development), and test sets (Bowman et al., 2015; Marelli et al., 2014). For datasets where this is not provided, items may be split randomly by selecting, for example, 80% of items for training and 10% each for validation and testing (Le & Mikolov, 2014).

Common quantitative evaluation metrics for labelled datasets include average error and number of incorrect classifications, this is commonly reported using the percentage of correct classifications on the test set. This may employ methods such as 10-fold cross-validation with randomised initial network weights (Cer et al., 2018; Kiros et al., 2015). Tools for automated evaluation of some types of models on multiple downstream tasks have been released, such as SentEval (Conneau & Kiela, 2018).

It is also common to give selected examples of output, for example, demonstrating the types of items that are considered most similar or interesting predictions the model has made (Kiros et al.,

2015; Mikolov, Chen, et al., 2013). This is usually given as the basis for discussion or analysis of the kinds of features the model is sensitive to.

Some of the common evaluation datasets are listed in Table 2.1. This list is not exhaustive but shows the variety and scale of datasets used in these tasks and their applications. For some of these datasets, variations may exist with a different number of items, either due to separate releases of the data, or reformulations of the data for use in some studies.

Table 2.1 - Common NLP evaluation datasets. Alternate versions with different numbers of items may exist for some datasets, the versions given here are examples used in the literature reviewed in this chapter.

Name / Source	Description	Usage
Stanford Sentiment Treebank (SST) (Socher et al., 2013)	70k phrases labelled with binary sentiment	Sentiment analysis
Stanford Natural Language Inference (SNLI) (Bowman et al., 2015)	570k pairs of phrases labelled either entailment, contradiction, or neutral	Natural Language Inference (NLI), Transfer learning
Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014)	10k pairs of sentences labelled with scores out of 5 for relatedness	Semantic relatedness, Entailment
Movie Reviews (MR) (Pang & Lee, 2005)	11k snippets from movie reviews labelled with ratings out of 5	Sentiment analysis, Recommender systems
Movie Lens (ML) (Harper & Konstan, 2015)	20M movie ratings and 465k tag applications applied to 27k movies by 138k users	Tag prediction, Recommender systems
Customer Reviews (CR) (Hu & Liu, 2004)	4k sentences from Amazon reviews labelled with sentiment	Sentiment analysis, Recommender systems
Text Retrieval Conference data (TREC) (X. Li & Roth, 2002)	6k questions labelled by type	Question-type classification
SUBJ (Pang & Lee, 2004)	10k sentences from movie reviews and summaries labelled with subjectivity scores	Subjectivity/objectivity classification
Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005)	5.8k pairs of sentences from news sources with binary labels for paraphrasing	Paraphrase detection
MPQA (Wiebe et al., 2005)	11k phrases labelled with opinion polarity (positive or negative)	Sentiment analysis

2.2.4 Information Correctness and Bad Actors

The quality of any machine-learning based solution is dependent on the quality of the data it is trained on. Many systems simply assume the correctness of the training corpus. In natural language, incorrect usage of terms in text is a form of noise that a robust solution must account for. An additional issue when mining a public corpus is that documents may be accidentally or deliberately misleading; this includes issues such as false or exaggerated claims, search engine exploitation ("gaming the system"), poor factual correctness, and subjectivity.

In many scenarios, particularly commercial markets and services, document authors have an interest in promoting their content over others, even if the recommendation may be suboptimal for the user. In the context of search and recommender systems, this is often exhibited as Search Engine Optimisation (SEO) but can in some cases be deliberately used to deceive the algorithm into ranking a document highly for unrelated searches, for example, by adding false keywords.

Much research has focused on methods of mitigating this issue, such as using scores for validity and authoritativeness, typically based on concrete relationships with other documents such as hyperlinks (Brin & Page, 1998) or Scientometric data (Ibrahim et al., 2017).

Approaches based on natural language understanding of full-text content are less subject to SEO (as meta-data is typically less important) but are in turn vulnerable to deceptive use of language, such as false claims and advertising. These approaches are also likely to deceive humans but can be overcome by verifying information through reasoning and by corroborating information from additional sources. Detection and mitigation of false claims are beyond the scope of this thesis but are an important consideration when selecting datasets and evaluating results.

2.3 Language Modelling

One of the principal difficulties in NLP is the effective modelling of natural syntax and semantics. To accurately derive meaning from natural text, it is necessary to have a sophisticated method of processing the language's syntax and the meanings of words, and potentially also phrases.

The vocabularies of natural languages typically contain words with subtle, complex, or contextual meanings such as synonyms (words with similar meanings), homographs (words spelt the same with different meanings), words with multiple connotations, words with variable spelling (including common misspellings of words), and various inflectional forms for different tenses or subjects.

This section looks at some of the techniques used in normalising syntactic differences in text through pre-processing and generating semantic representations of text through language modelling and text embedding, as well as some of the challenges and limitations of these approaches.

2.3.1 Pre-Processing

Natural language text commonly features symbols and terms that may not be of interest for a model, such as punctuation, styling and markup tags, and words deemed to have little semantic meaning. The number of unique symbols and the vocabulary size of a language model can significantly affect performance (Gowda & May, 2020), so it is common to pre-process input text to remove extraneous features and normalise text to address variations in syntax, as outlined in the following sections.

The following sub-sections are not an exhaustive guide to pre-processing text data but demonstrate some examples of the types of features that may be removed from text before use in training or prediction, which may be important to interpreting a model's understanding of language and the features available for it to learn. This section focuses on English language text and some but not all techniques may apply to other languages; the discussion of tokenisation and canonisation techniques for instance is not applicable to languages that use logograms such as Mandarin hanzi or Japanese kanji, where individual characters represent words or morphemes instead of sub-word units as in alphabetic languages.

2.3.1.1 Tokenisation

The first pre-processing step commonly applied is tokenization, where the input text is converted into a sequence of tokens. This conversion is typically based on whitespace and punctuation,

where each word and punctuation symbol becomes a token (Russell & Norvig, 2020). This can result in words containing punctuation, such as apostrophes, being split into multiple tokens, so these may sometimes be replaced before or during tokenisation. Some sophisticated tokenizers may identify locutions such as multi-word phrases (also referred to as n-grams) as a single token based on frequent co-occurrence or grammatical rules (Russell & Norvig, 2020). In the example below, item 1 shows the input text, 2 shows the text after tokenisation (where | is the divider between tokens), 3 shows the text after tokenisation with apostrophes removed, and 4 shows the text after tokenisation preserving locutions.

1. $New\ York\ isn't\ in\ Europe$
2. $New \mid York \mid isn \mid ' \mid t \mid in \mid Europe$
3. $New \mid York \mid isnt \mid in \mid Europe$
4. $New\ York \mid isnt \mid in \mid Europe$

Other approaches to tokenisation include character-level tokenization and sub-word tokenization.

Character-level tokenisation forgoes building a vocabulary of words and instead uses single characters as tokens (Kim et al., 2016; Ling et al., 2015), removing the problem of vocabulary size and allowing handling of out-of-vocabulary words, but resulting in text being mapped to a large number of tokens, which may not individually carry meaning (i.e. individual letters typically do not have semantic meaning).

Sub-word tokenisation allows words to be broken into multiple tokens such that prefixes, suffixes, and other meaningful character sequences (e.g., morphemes) can be used as features. Like character-level tokens, these can be composed to increase the coverage of the vocabulary, but unlike character tokens, these have meaningful semantics. How words should be split is sometimes learned, such as with WordPiece (Sennrich et al., 2016; Wu et al., 2016) which uses a variation of Byte Pair Encoding to select tokens based on frequent pairs of consecutive sequences.

2.3.1.2 Canonization

Some aspects of a vocabulary can be normalised based on grammatical rules and by looking at the roots of words using processes such as stemming and lemmatisation, for which several solutions exist, such as Porter's algorithm (Porter, 1980). These algorithms substitute words with their canonical (dictionary) forms by removing inflections (tense, case, voice, aspect, person, number, gender, and mood), thus reducing the size of the vocabulary (e.g., "playing", "plays", "played" all become "play"). However, some contextual information (e.g., verb tense) may be lost as words sharing the same lemma are considered to have identical meanings.

Some approaches also choose to remove letter casing and diacritics to further reduce the number of unique tokens. This assumes that alternative spellings add little semantic information at the

cost of greater vocabulary size. It has been demonstrated that preserving case can benefit some tasks such as Named Entity Recognition and Part-of-Speech tagging, but for other tasks removing case and diacritics can improve model performance (Devlin et al., 2019).

In the following example, 1 shows an input sentence, and 2 shows its canonical form after stemming and character substitution.

1. $I\ worked\ at\ a\ café\ before\ working\ here$
2. $I \mid work \mid at \mid a \mid cafe \mid before \mid work \mid here$

2.3.1.3 Stop Words

A commonly used approach is to filter the text against a stop list, a list of stop words, which are terms assumed to not be of interest, such as connective words like "the", "to", "a", and "an". NLP software libraries, such as NLTK (Bird et al., 2009), often provide generic stop lists, but there is no agreed universal stop list. Different types of models may benefit from different stop lists due to differences in how they process the input. For example, in the following sentences (1 & 2) the words "from" and "to", which are commonly used as stop words, change the meaning of the sentences significantly. Sentence 3 shows how both sentences, which originally had different meanings, are reduced to the same tokens when the stop list ["to", "from"] is applied.

1. $Travelled\ \boldsymbol{from}\ England\ \boldsymbol{to}\ Spain$
2. $Travelled\ \boldsymbol{to}\ England\ \boldsymbol{from}\ Spain$
3. $Travelled \mid England \mid Spain$

If a model considers word order, then the positioning of the words "to" and "from" can be used to identify which location is the origin and which is the destination in sentences 1 and 2, but this is no longer possible after applying the stop list. If using a Bag of Words (BoW) model, the word order is not preserved, so discarding of the stop words does not change the apparent meaning; to a BoW model sentences 1 and 2 are identical and removing stop words (sentence 3) does not change the meaning any further.

2.3.2 Distributional Language Models

Distributional language models attempt to produce representations for words, phrases, or parts of speech in a continuous feature space. Based on the principle "You shall know a word by the company it keeps" (Firth, 1957), these models allow positioning of terms such that more related terms have greater proximity than unrelated terms.

Many natural language and text analysis systems are based on a Bag of Words (BoW) approach, which creates a statistical model of text by counting the occurrences of words within documents. Some examples include Term Frequency-Inverse Document Frequency (TF-IDF) (Sparck Jones,

1972), Latent Semantic Analysis (LSA) (Deerwester et al., 1990), and Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

One property of distributional models is that they can produce numeric representations of their input in a continuous feature space. In LSA and LDA, the representations of a document's content are used to compare the semantics of documents. Representation can also be produced for individual words, commonly referred to as word vectors or word embeddings. These embeddings represent the semantics of the words based on their co-occurrence in the corpus.

These approaches produce reasonable results for tasks such as topic classification and keyword matching but cannot capture context in the form of the meaningful order and structure of text, which is required for more sophisticated tasks such as reasoning, entity extraction, and relationship extraction; for these tasks, it is necessary to model both the semantic and syntactic properties of the text.

To partially address this, statistical language models can make use of n-grams, grouped representations of frequently neighbouring terms. These models estimate conditional probabilities for the next word for a large number of contexts (the preceding words) (Bengio et al., 2001). However, this approach has limited scalability, requiring very small context sizes (typically up to 3 words), due to sparsity introduced by the large dimensionality, commonly referred to as the curse of dimensionality.

2.3.3 Neural Language Models

Neural language models apply neural networks to the task of learning distributed representations of text. While the earliest examples of these models (Bengio et al., 2001) were limited by computational efficiency, particularly regarding corpus size and vocabulary size, these models were able to overcome the curse of dimensionality by allowing each training sample to inform the model about semantically similar samples, allowing for much greater context sizes without dilution due to sparsity.

In 2013 Mikolov, Chen, Corrado, & Dean (2013) presented a neural network model for learning word vectors, word2vec, able to efficiently process a much larger corpus than previous models and produce superior word embeddings, significantly surpassing state-of-the-art models in both semantic and syntactic word relationship benchmarks. Further optimisations in their second paper (Mikolov, Sutskever, et al., 2013) gave even stronger results and performance gains.

Word2vec employs two architectures for learning word vectors on a large corpus:

The Continuous Bag of Words (CBOW) model, based on the conventional BoW model, uses word co-occurrence to learn word vectors. CBOW is distinct from other BoW approaches in that a filter is applied to the input text as a 'sliding window', such that the embedding for the target word is trained using only words which occur close to it in the text. This adds context to the BoW model, but still lacks respect for word order, and ignores the local distance between words so long as it is within the window size.

The Continuous Skip-gram model instead trains word embeddings by using the target word to predict words that occur nearby. The number of training samples featuring each nearby word is dependent not only on the number of contexts in which the words co-occur but also the distance between the words, such that more distant words comprise fewer training examples. This approach maximises the benefit of a large context window while keeping computational complexity manageable.

These architectures are the basis for several successive approaches such as Paragraph Vectors (Le & Mikolov, 2014) and Skip-Thought Vectors (Kiros et al., 2015) which are discussed later in this chapter. Some models such as GloVe (Pennington et al., 2014) further improve performance by combining these predictive tasks with global statistics like discussed in the previous section.

2.3.4 Distributional Representations of Documents

Paragraph Vectors (Le & Mikolov, 2014) extend the word2vec model for word vectors to larger lexical structures such as sentences, paragraphs, and documents. By including a unique paragraph identifier in each training sample, a vector is learnt for the paragraph as an indirect result of the word prediction task. The resulting paragraph vectors capture the semantics of the paragraph and can be used as a distributional semantic representation of the paragraph.

This was shown to be a more meaningful representation than previous document-level fixed-length techniques such as taking the sum or product of the document's word vectors. The explanation given for this is that the paragraph vectors are better able to capture context in the form of the structure of the documents, which is lost when only considering a documents constituent word vectors.

Future research into document representations supports this hypothesis. Various successive models feature enhancements attempting to retain more context. Some notable examples include Skip-Thought Vectors (Kiros et al., 2015) and Dependency Based Word Embeddings (Levy & Goldberg, 2014).

2.3.5 Sequence to Sequence Models

The Skip-Thought Vectors model is a sequence-to-sequence Recurrent Neural Network (RNN) with a similar learning objective to the word2vec skip-gram model but on the sentence level. Sentence-level vectors are learned by predicting sentences occurring before and after the target sentence. This captures a high degree of contextual information, as it is sensitive to both word order and sentence order. The resulting skip-thought vectors have been shown to be highly effective for tasks such as identifying semantically similar sentences and sentiment classification. A skip-thoughts model trained on the BookCorpus dataset (Zhu et al., 2015) achieved state-of-the-art performance on several benchmarks and is used as a baseline by many future models.

As a deep neural network, the skip-thought model requires a very large number of learnable parameters compared to shallow models like word2vec, resulting in much longer training times and a requirement for a very large training corpus for high-quality vectors to be produced. As skip-thoughts are dependent on the order of sentences within documents, they are also less suited to tasks involving short-texts, for which they perform comparably to other models, many of which are less complex.

It has been shown to be the case generally for sequence-to-sequence models that their improved performance is more prominent in tasks where understanding of global/long-range semantics is required, and less in tasks involving keyphrase recognition (Seo et al., 2020).

The FastSent model (Hill et al., 2016) builds on the principles of Skip-thoughts by learning to predict sentence order but uses a simpler log-linear algorithm where sentence vectors are produced by summing their constituent word vectors. This significantly reduces training time while keeping the sentence level semantics of skip-thoughts, but the loss of word order reduces performance in some tasks, such as detecting sentiment, paraphrasing, and subjectivity.

2.3.6 Fine-Tuning and Transfer Learning

Subsequent works including Conneau, Kiela, Schwenk, Barrault, & Bordes (2017) and Cer et al., (2018) have extensively explored the effects of fine-tuning and transfer learning on sentence-level vectors. In this context, transfer learning involves using vectors pre-trained on a large unsupervised corpus, then fine-tuning using a comparatively small amount of high-quality labelled data.

Conneau et al. (2017) examine several model architectures and evaluation techniques and shows significant performance improvements when sentence vectors are pre-trained on natural language inference tasks, which they hypothesise is due to the generality of the high-level

semantic understanding required to solve the task. An interesting observation made by the authors is that models which perform better at the pre-training task do not necessarily produce the best results in other tasks, which they attribute to over-specialisation and focus on features important to the training task rather than more general semantic information. They also find that larger sentence vectors can generalise better to other tasks, even when this does not result in improved performance on the training task, which they also attribute to less specialisation of learnt features.

The paper also introduces a new model, InferSent, a bi-directional LSTM (Long Short-Term Memory) sequence-to-sequence model using transfer learning. This improved on the state-of-the-art for several tasks of various types including semantic relatedness, inference, and sentiment. While this model requires some supervised data in addition to the usual unlabelled corpus, the data required for both datasets is much less than exclusively unsupervised, or supervised models. When trained on 570 thousand ordered sentences, and fine-tuned using labelled data (specifically, from the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015)**)**, their model consistently outperforms the best Skip-Thought model trained on 64 million sentences.

Finally, the paper introduces an automated tool for evaluating the quality of sentence vectors on several common benchmarks, which was later published and released open-source as SentEval (Conneau & Kiela, 2018). This incorporates a diverse set of 17 NLP challenges including opinion polarity, review sentiment, objectivity, paraphrase detection, inference, question-answering, and image-captioning. This encompasses many of the evaluation datasets commonly used in the contemporary literature.

Cer et al., (2018) take a different approach to transfer learning, by re-using vectors learnt by other models, then applying fine-tuning. They also investigate the effect of combining both word and sentence level vectors. They improve on state-of-the-art performance using their model, Universal Sentence Encoder (USE), which combines pre-trained word vectors from a word2vec skip-gram model with sentence vectors trained on a Wikipedia dataset, then fine-tuned using labelled data from the SNLI dataset.

Their best performing model, USE_T uses a transformer architecture. While this model scores highest in most benchmarks, significantly surpassing state of the art, it has $O(n^2)$ compute and memory requirements based on sentence length. Another model introduced in the paper, USE_D, is trained on the same data but uses a Deep Averaging Network (DAN) architecture, which has linear compute and memory requirements. Their evaluation showed the performance of this

model to be only slightly less than USE_T in most tasks, and with a better score on the subjectivity evaluation benchmark.

As with the InferSent model (Conneau et al., 2017), USE's use of transfer learning allows for near state-of-the-art performance even with minimal training data. USE_T was shown to be competitive with models trained on the full 67.3 thousand examples from the SST dataset (the variant from Conneau et al. (2017)) when fine-tuned on only one thousand examples (Cer et al., 2018).

These papers clearly demonstrate the advantages of transfer learning and fine-tuning. This technique not only sets the new state-of-the-art performance benchmarks but does so with far less training data than previous exclusively supervised or unsupervised models. This both results in faster training times and significantly increases the number of potential applications for these techniques in areas where labelled data is limited. Both factors reduce the cost of implementation and experimentation. This is further supported by both models being openly available online for use in other applications, along with detailed instruction for replication being given in both papers.

2.3.7 Addendum: Deep Bidirectional Encoders

Generating superior semantic representations of documents and parts-of-speech continues to be a highly active research area. Recent models trained on very large datasets, particularly those using bidirectional encoding transformer architectures have achieved significant improvements on a range of downstream tasks. Some notable examples include BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and GPT-3 (Brown et al., 2020). Commonalties of these models include considering sentence context when producing text embeddings and using sub-word tokenisation for better handling of out-of-vocabulary words.

An advantage of using models which produce fixed-length semantic vectors is that they can be used interchangeably in most cases not requiring major methodological changes in their use, so future developments in embedding models can easily benefit downstream applications. While some of the models discussed here were not available when the experiments detailed in this thesis were conducted, it is likely the approaches presented could benefit from the use of these, and future, superior embedding models.

2.4 Search and Recommender Systems

Retrieving information from heterogeneous sources is not a new problem and is commonly a consideration in search, recommendation, and expert systems. This section looks at some of the methods used in these areas, their data requirements, and limitations.

2.4.1 Item Similarity Techniques

Effective methods for calculating item similarity are important for information retrieval tasks such as finding items similar to a query, they also have applications in recommender systems that make use of content information.

Statistical similarity techniques, such as Term Frequency – Inverse Document Frequency (TFIDF) (Sparck Jones, 1972) are widely used in search and recommendation systems (Beel et al., 2016). While these approaches provide the benefits of determinism, comparative simplicity, and understandability (contrasted to learning models with complex parameters), they are syntactically variant meaning that different phraseology, for example, use of different terminology will prevent matching, limiting their usefulness for matching documents from different domains or writing styles (formal versus informal, specialist versus general-audience, etc.).

Language models (detailed in section 2.3) can be used to produce semantic representations of documents (or parts of documents) which can be compared, for example, using cosine similarity or angular distance (Cer et al., 2018; Reimers & Gurevych, 2020). These models overcome syntactic variance and match documents with different phraseology by modelling semantics (meaning) and considering context.

Multi-modal embedding models such as Sun, Li, & Zhang, 2018; Sung, Lenz, & Saxena, (2017) can be used to determine the similarity between various types of media such as images or audio, enabling these to also be considered for a search query or item comparison.

While ranking by similarity is suitable for retrieving documents matching a query, or comparing pairs of items, it alone is not sufficient for matching documents on relationships other than content similarity, such as when documents that might share a relationship are meaningfully dissimilar, for example, in supply chain companies buy from/sell to companies that are different, not companies that are similar (i.e., their competitors).

Distributional semantic techniques such as document embeddings have also been shown to be effective for other tasks such as analogous reasoning, paraphrase identification, inference, question-answering, and machine translation (Cer et al., 2018; Conneau et al., 2017; Conneau &

Kiela, 2018; Hill et al., 2016; Kiros et al., 2015), which may have other applications in information retrieval and recommender systems.

2.4.2 User Behaviour Techniques

Observing patterns in user interactions is the basis of many recommender systems, particularly in e-commerce where this data is abundant, and users commonly have extensive interaction histories. For relationship inference, it is significant that these approaches are not based on the direct similarity between the user and the target documents, but on the implicit relationship of a user's need for what the documents describe.

Collaborative recommender systems aggregate user interactions to find similar users and recommend the items they liked. Common techniques include collaborative filtering, where matrix factorisation is used to reduce the dimensionality of the sparse matrix of user-item interactions. The resulting dense matrix can be used to recommend items based given a user's past interactions (Koren et al., 2009).

However, due to dependence on user interactions, collaborative approaches present issues when items are time-sensitive or competitive as items may not remain valid long enough to accumulate a significant user record (Shalaby et al., 2018; Yuan et al., 2016). Further, this approach can result in positive feedback loops where a document being frequently recommended results in more interactions, resulting in more recommendations and therefore more interactions; this virality effect can result in a few generic or broadly applicable documents being disproportionately recommended, while newer and more niche documents are not promoted due to less existent user behaviour data (Deldjoo et al., 2019; Yuan et al., 2016).

In contrast, content-based recommender systems recommend items based on similarity to a query or the user's past interactions. In some works, such as Suglia et al. (2017) and Ferro et al. (2016), item description embeddings are used for this comparison. A common approach used in both papers is to generate a representation of the user by averaging the description embeddings of items the user has interacted with previously. Some works consider additional auxiliary information about users and items such as user search contexts (Liebling et al., 2012), or additional content or meta-data information retrieved from web sources (Musto, Semeraro, de Gemmis, & Lops, 2016), knowledge-bases (F. Zhang et al., 2016), ontologies (Suglia et al., 2017), or folksonomies (Chen et al., 2010).

These content-based systems are less dependent on items having detailed interaction histories as they can recommend new items based purely on content similarity, but they still require the user to have known past interactions to generate a representation of the user.

Ontological approaches can be used to create user profiles for bootstrapping recommender systems and to generalise observed interactions by using ontological relationships such as parent-topic, sub-topic (Middleton et al., 2004). Ontologies are often hand-crafted and domain specific, however, the ability to make use of existing ontologies can alleviate this issue in domains that are well covered, although it is still necessary that a judgement is made about what ontologies are suitable, and some domains may not have existing comprehensive high-quality ontologies. Ontology learning offers a solution to these issues, producing an ontology through association rules, classification, clustering, or other content based techniques (Maedche & Staab, 2001), or extending ontologies to less covered domains through transfer learning (Xie et al., 2021).

Association Rule Mining can be used to identify rules of association between items, scored by confidence (the proportion of relationships for an item where it co-occurs with the other) and support (the proportion of transactions involving the items) (Agrawal et al., 1993). This has been applied to produce recommendations by creating personalised rules for each user (Adomavicius & Tuzhilin, 2001), or to address the cold start problem by using rules generated from historic data to bootstrap a recommender system (Bendakir & Aïmeur, 2006). This technique has the advantages of being intuitively understandable and easily fitting business cases such as relationships between retail products or departments (Nakhaeizadeh et al., 2000). The use of intuitively understandable rules also allows these approaches to be more explainable. However, these approaches still rely on a quantity of historic data and do not address double cold starts, where both the user and item are new, as these users/items do not appear in any rulesets.

Hybrid recommender systems combine aspects of the techniques discussed previously, such as making use of both content and collaborative data to learn joint user-item embeddings as in neural collaborative filtering (He et al., 2017). Many hybrid models make use of deep learning to learn the relationship between user and item features, however, this results in these techniques requiring a large amount of training data and present issues with interpretability (S. Zhang et al., 2019).

A limitation of behaviour driven approaches is that they depend on users seeking documents similar to previous searches, which may not be true between sessions; that is, if a user changes their search objective the data may no longer be relevant because the nature of the relationship (the user's need) has changed (Kong et al., 2015). Additionally, these approaches are constrained by the availability and quality of relevant behaviour data.

2.4.3 Sparsity, Partial Labelling, and Cold-Starts

Datasets that use user interactions, user ratings, or known item relationships (such as those given by experts or based on modelling real-world relationships such as supply-chain) as labels are often highly sparse. That is, most possible user-item or item-item pairs do not have a known rating/relevance score/relationship class. For example, in the case of review data, it is unlikely that any user has reviewed all items in the dataset; or in supply-chain data, not all potential relationships between businesses will be known or necessarily existent in the real world.

In sparse datasets, labels may not be evenly distributed across all items. Some items in a dataset may have many labels (for example older or more popular products, users who have been using a service longer), while others may have very few or no labels (such as newly added items, newly registered users). In the context of recommender systems, an item (or user) for which no labels are known is referred to as a cold-start.

The cold-start problem can be divided into the two sub-problems of item-wise (new item) and user-wise (new user) cold starts (Lops et al., 2013). The item-wise case is commonly addressed by content-based and hybrid recommender systems (S. Zhang et al., 2019); however, the user-wise case has received less attention, even in scenarios where content information for the user is available.

Content-based and hybrid recommender systems reduce the requirement for item labels by making use of item content, such as descriptions. Many such systems rely on either knowledge bases and ontologies (Middleton et al., 2004; F. Zhang et al., 2016), which do not avert the requirement of experts for new or commercially guarded domains, or tags and categorisation (Shalaby et al., 2018; Xu et al., 2016), which requires either many labels or distinct groupings in the data.

Yuan et al., (2016) examine the real-world data problem of matching users to job postings, where items are time-sensitive and new items are very frequent. They make the case that high-performance techniques that require item labels can be generalised to cold-start items by pairing labelled and unlabelled items based on the similarity of their content.

2.5 Provenance and Explainability

Machine learning solutions and some statistical methods may output only a series of ranked items or confidence scores in response to a query, and the rationale behind these decisions is unknown.

In the case of scores from neural networks or other learning models, the reasoning behind the algorithm's decision is unknowable as it is derived from the free parameters of the model which only have meaning inside the model and cannot be used to meaningfully explain results, these are often referred to as "Black Box" systems. Much research has been done into approaches for understanding the internals of deep learning models via visualisation, particularly in the areas of text summarisation (See et al., 2017) and computer vision (Yosinski et al., 2015; Zintgraf et al., 2017), and some works have looked at understanding the upstream neural language models discussed in section 2.3 (J. Li et al., 2016). However, while these visualisation techniques offer some insights into the factors the model considers important, they cannot produce a reasoned explanation for the response to a particular query in any way similar to how a human decision-maker might.

In contrast to this are "White Box" systems which produce meaningful provenance that can be used to explain results and study the operation of the model, these typically include rule-based models and expert systems. Herlocker, Konstan, & Riedl (2000) discuss how in user-facing scenarios some techniques such as collaborative filtering can be presented as either a white box or black box model, by giving feedback to the users based on either the operational steps of the model (white box) or the inputs and outputs of the system such as user evaluations of the quality of results (black box).

Detailed provenance data, such as lists of decision-making steps, inferences, rules, knowledge, and items considered when evaluating a query can be used to produce visualisations such as graphical plots or flow diagrams to help users understand the reasoning behind a result, increasing their confidence in the decision, or highlighting potential flaws in the model. This makes provenance highly desirable both during development, to debug and improve the model, and for user-facing systems as users have greater trust in answers that are explainable and can make more informed decisions based on the results (Herlocker et al., 2000).

2.6 Conclusions

The extensive literature on natural language processing demonstrates several effective techniques for learning semantic representations of text which can be used to learn or reason about the text's properties for a large variety of applications. Further, recent models show excellent ability to generalise and perform high-level reasoning tasks such as question answering and natural language inference. The introduction of transfer learning and fine-tuning has opened many possibilities for investigation, such as what linguistic tasks and types of training examples promote the best general models of language and why.

Major advances in this problem area have been made by taking conceptually simple techniques from early models and applying them to new problems or in new ways. Now that several techniques have been identified that are highly effective and generalisable, new problems can be approached such as those set out in the research questions of this project (section 1.2).

The next two chapters look in more detail at the challenges, resources, and techniques relevant to this project. The approaches used take inspiration from the techniques applied to similar problems in the literature and extend the current work by using neural language models to address the challenges described in section 2.2, allowing effective downstream models and combinations of techniques to be produced to address the specific challenges and research questions of this project.

Chapter 3 **Inferring Relationships from Few Labels**

3.1 Chapter Overview

This chapter details the approach employed in answering SRQ1: “How can machine understanding of text be used to identify relationships between documents in large collections of unstructured text?”. Section 3.2 introduces in more detail the specific problem examined. Section 3.3 looks at data resources and challenges relevant to this task, section 3.4 presents the Isle of Wight Supply Chain (IWSC) dataset which exemplifies two challenging scenarios. Section 3.5 discusses appropriate methods of evaluation, and section 3.6 introduces the Transitive Semantic Relationships (TSR) model, a novel solution addressing the shortcomings in this area, particularly in concern to SRQ1 and SRQ3. The implementation and experimental setup are detailed in section 3.7, the main results are given and discussed in section 3.8, and subsequent variations on the algorithm explored to improve performance are examined in section 3.9.

3.2 Introduction

Much research has been done into matching short text queries to documents, including keyword queries and natural language questions. These approaches require either an approximate knowledge of the target documents (for example, probable keywords) or a targeted query. While good accuracy has been achieved for these tasks in controlled tests (matching and ranking in test datasets), in real-world applications it is often necessary for users to repeatedly rephrase, modify, and refine their query to find the results they are looking for (versus what is technically a good match for the query). An effective query must describe the target documents, which is problematic when the user has no knowledge of the target documents.

An issue with these approaches is that the query depends on the relationship between the domain of the user's knowledge and the codomain of information contained within retrievable documents. This is problematic if the user does not have knowledge of the information in the codomain and therefore cannot form a query that precisely describes the desired documents. This results in users needing to start with broad queries and progressively narrowing their search as they learn more about the contents of the codomain (i.e., what information is available). This requires more time from users, results in many low-value queries, and may result in the exclusion of good results as users attempt to tailor their queries based on the non-optimal results of earlier queries.

One solution to this problem would be to make direct use of the user's knowledge rather than relying on knowledge gained from unsuccessful queries. In many scenarios, the user is attempting to find a suitable match or matches for an entity that they could describe, which shares a relationship with the target documents. Some examples are listed in Table 3.1.

Table 3.1 - Examples of capability relationships

Domain (user knowledge)	**Relationship**	**Co-domain (retrievable documents)**
Project description	Fulfils criteria	Grant description
Resume or CV	Role suitability	Job advert
Company description	Collaboration / Supply Chain	Company description
Researcher description	Expertise / Consultancy	Company description

Notably, in the cases above the relationships between the documents are not based on similarity. For example, in the case of supply-chain, a company would be seeking a supplier or consumer with a different but related specialisation to themselves. In the case of matching job or grant funding, the format and terminology used in the domain and codomain are likely to differ. As such, searching using a description of a domain entity is unlikely to produce good results using a similarity-based search. It is therefore necessary for an algorithm to consider the nature of the relationship between the domain and codomain documents. The relationships listed above are all examples of types of capability relationships, which will be the focus of this chapter.

Most existing search and recommendation techniques outlined in Chapter 2 do not directly model the relationships between documents, but instead rely on measures of similarity or behaviour trends. While ontological and rule-based approaches can leverage some other relationship types (such as hierarchical is-a relationships), they depend respectively on good domain coverage or many historic examples from which to derive rules, and do not predict relationships like those in Table 3.1 for unseen items. Previously, some of the flaws of existing techniques have been identified, as well as their strengths. SRQ1 seeks to determine if these solutions can be improved by modelling the semantics of such relationships, or if new solutions based on this approach could surpass them in the scenarios in which they perform poorly.

3.3 Data Requirements

New Big Data recommendation systems face a high barrier to entry due to the large labelled data requirement of most existing recommendation techniques such as collaborative filtering and bespoke deep learning models such as Suglia et al., (2017). Obtaining this labelled data, such as user interactions or human judgements, is particularly problematic in highly specialised or commercially competitive domains where this labelling may not yet exist or not be freely available, often requiring expensive expert or crowd-sourced labelling. As such, techniques that function well with few labels are highly desirable.

Constructing a high-quality model for relationship discovery is likely to require a large volume of suitable training data, including examples of existing relationships for learning and evaluation. Machine learning and deep learning techniques can require datasets on the order of hundreds of millions of words to create effective models, particularly when the number of parameters is large. Recent literature on neural language models has particularly shown that more diverse training corpora produce better and more generalisable results (Conneau et al., 2017).

Content-based and hybrid recommender systems reduce the requirement for user-item interaction labels by making use of item content, such as descriptions. Many such systems rely on either knowledge bases and ontologies (Zhang, Yuan, Lian, Xie, and Ma, 2016), which do not avert the requirement of experts for new or commercially guarded domains, or tags and categorisation (Xu, Chen, Lukasiewicz, Miao, and Meng, 2016), which requires either many labels or distinct groupings in the data.

Some machine learning architectures could use transfer learning to benefit from word or sentence vectors pre-trained on a large unrelated corpus. Studies (Cer et al., 2018; Conneau et al., 2017) have shown that transfer learning making use of as few as 1000 labelled examples can produce competitive results on several benchmarks.

Alternatively, statistical and reasoning-based techniques can have less requirement for labelled data as they draw inferences directly from the corpus rather than using it to train many free parameters to produce a generalisable model as in neural networks.

Due to the multitude of applications for relationship discovery, various datasets are available pertaining to different applications. The requirements for a dataset are that historic or current relationships can be extracted (the ground truth), and that entities have sufficiently detailed descriptions. The dataset should ideally contain minimal false or misleading information (such as exaggerated advertising) for the reasons detailed in section 2.2.4.

Table 3.2 - Potential Data Sources

Data Source/Type	Potential Usage	Notes
Innovate UK funding award history (GOV.UK, 2018)	Existing capability and collaboration relationships	Includes collaboration data in the form of multiple participant projects
Gateway to Research funding award history (GtR, 2018)	Existing capability and collaboration relationships	Large dataset of more than 82,000 projects, 68,000 people, and 36,000 organisations
Commercial data	Existing capability and collaboration relationships	Often very sparse or partial coverage/incomplete data
Websites of public institutions and businesses	Textual descriptions of entities	Likely to be very noisy and highly heterogeneous
Pre-trained language models (word embeddings, etc.)	Data analysis / pre-training	

As well as complete datasets, many resources exist which could be used to provide or enhance descriptions of entities to enable better matching. Some candidate datasets and data resources identified are detailed in Table 3.2.

3.4 Isle of Wight Supply Chain Dataset

In collaboration with the project's industrial sponsor, a dataset on the Isle of Wight Supply Chain (IWSC) has been produced. The data consists of varying length text descriptions of 630 companies on the Isle of Wight taken via web-scraping of websites promoting local businesses (*IWChamber*, 2018; *IWTechnology*, 2018; *Marine Southeast*, 2018).

HTML tags and formatting have been removed, but the descriptions are otherwise unaltered and are provided untokenized, without substitutions, and complete with punctuation. Some descriptions contain product codes, proper nouns, and other non-dictionary words.

Most of the descriptions are a few sentences describing the market role of the company or a general description of the company's activities or products. Several but not all the descriptions also contain a list of keywords, but this is included as part of the descriptive text and not as an isolated feature. The mean description length is 61 words, or 412 characters (including whitespace). The distribution of description lengths is shown in Figure 3.1.

The IWSC dataset is provided with two discrete sets of labels intended to evaluate algorithmic performance in different scenarios. In both cases, the labels are binary, directed, expert judgements of market relatedness based on the company descriptions. The number and distribution of labels are shown in Table 3.3. These labels are speculative potential relationships, not necessarily real existing relationships. Binary labelling was used as real-world supply chain relationships are typically multi-class binary relationships. i.e., any two companies either are or are not in each possible type of supply chain relationship.

The first label set, 'IWSC-SL', is comprised of the labels 'SL_consumers', 'SL_not_consumers', 'SL_suppliers', 'SL_not_suppliers', 'SL_competitors, and 'SL_not_competitors. These labels are concentrated on a small number of labelled items, relating them to a random distribution of other items (both labelled and unlabelled). These labels are intended for evaluation in the case that only records for a small subset of items are known and it is necessary to extrapolate from this to perform inferences on many unseen items. This scenario is termed "Subset Labelling" (SL).

The second label set, 'IWSC-ES', is comprised of the labels 'ES_suppliers', 'ES_consumers', 'ES_competitors', and 'ES_unrelated'. The labels are randomly distributed across all labelled items with no intentional patterns (random pairs were selected for labelling). These labels are intended for evaluation in the case that known items have very few labels and many are entirely unlabelled, in contrast to common recommender system datasets such as Movie Reviews (MR) (Pang and Lee, 2004), Customer Reviews (CR) (Hu and Liu, 2004), and MovieLens (Harper and Konstan, 2015), where most items have many recorded interactions. While in those examples the

labels are sparse as most possible item pairs are unlabelled, in this scenario, there is the additional condition that most items in the dataset do not occur in any of these pairs, as such this is termed “Extremely Sparse” (ES) labelling.

In addressing SRQ1, relating to relationship discovery, the four following tasks are of interest:

1. Prediction of “SL_consumers” labels using IWSC-SL labels and item descriptions
2. Prediction of “SL_suppliers” labels using IWSC-SL labels and item descriptions
3. Prediction of “ES_consumers” labels using IWSC-ES labels and item descriptions
4. Prediction of “ES_suppliers” labels using IWSC-ES labels and item descriptions

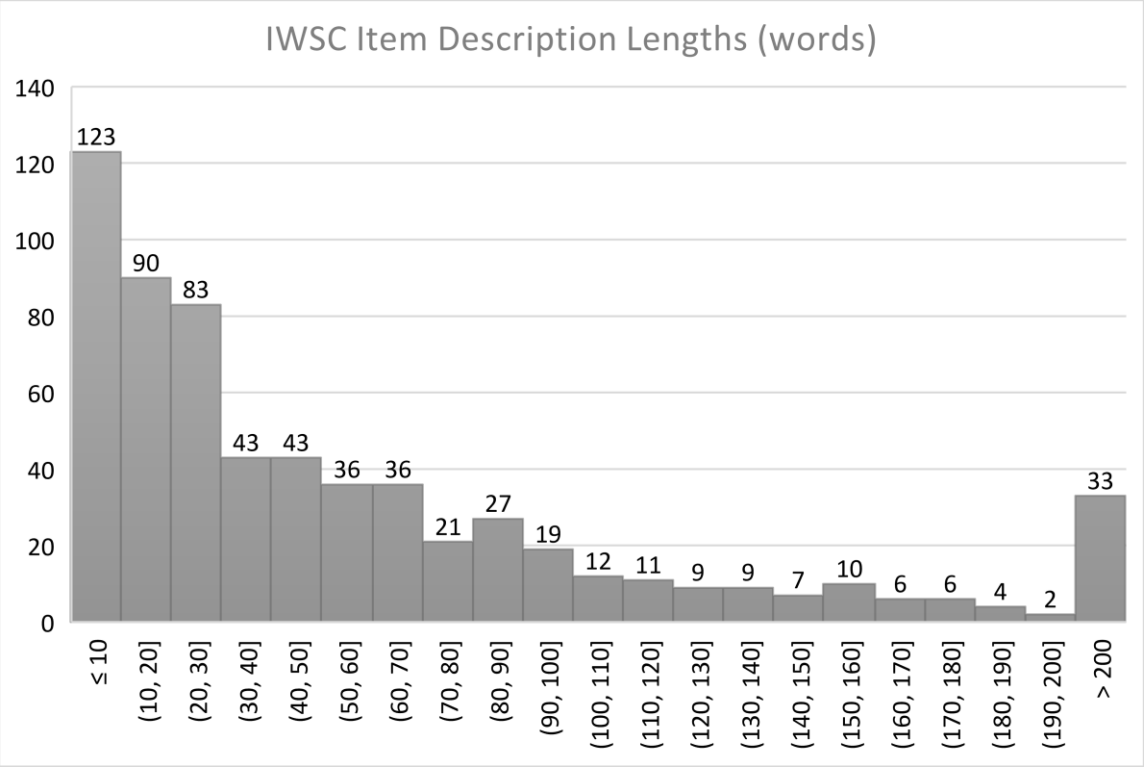


Figure 3.1 - Histogram of item description lengths in the IWSC dataset

Table 3.3 - Labels in the IWSC dataset. Labels are directed, such that "Labelled Items" is the number of items that known relationships are "from", and "Unique Targets" is the number of items relationships are "to".

Label Name	Total Labels	Labelled Items	Unique Targets
SL_suppliers	142	15	75
SL_not_suppliers	563	16	120
SL_consumers	376	17	117
SL_not_consumers	712	16	157
SL_competitors	82	15	49
SL_not_competitors	396	17	99
ES_suppliers	92	48	76
ES_consumers	207	51	171
ES_competitors	95	53	82
ES_unrelated	431	75	299

3.5 Evaluation Methods

Various evaluation metrics are used in recommender systems and information retrieval literature. As the IWSC dataset uses binary labels, and the total number of labels is small, various evaluation techniques have been investigated to determine the most suitable.

Normalised Discounted Cumulative Gain (NDCG) (Järvelin & Kekäläinen, 2002) is a common evaluation metric in information retrieval literature. This is a graded relevance metric which rewards good results occurring sooner in the results list, however, it does not penalise highly ranked negative items. As binary labels have no ideal order for positive items, this metric is unsuitable.

Quantitative error metrics such as Root Mean Squared (RMS) error or Median Absolute Error are also common. Error metrics naturally favour scoring systems optimised to minimise loss such as learning-to-rank algorithms and require scores to fit the same range as the label values. For the IWSC dataset, as the labels are binary, the range is 0 to 1.

For a binary labelled dataset, it is intuitive to set some threshold on the rankings and produce a confusion matrix and take precision (P), recall (R), and f1 scores. As scores are not evenly distributed, there is no obvious score value to use as a threshold for predicted positives and negatives, so instead some number of the top-ranked items must be considered predicted positives.

Due to the sparsity of labels in the dataset, the number and ratio of known positives and known negatives varies significantly between items and in many cases, the number of known positives is smaller than typical values of K used for Precision at K. As an alternative, R-Precision can be used, setting the threshold at R, the number of known positives, and taking the R most highly-rated items to be predicted positive and all remaining to be predicted negative; at this threshold P, R, and f1 are equal. In the results section, scores taken at this threshold are denoted as @R. A drawback of this approach is that only labelled pairs (known positives and known negatives) can be used for evaluation, which is a minority of possible pairs in a sparse dataset. The difficulty of this evaluation task also varies with the ratio of known positives and negatives which is undesirable when evaluating datasets such as IWSC where the ratio varies greatly between items.

Finally, techniques from the literature on implicit feedback are considered. Techniques for implicit feedback have the desirable property of allowing expansion of the number of unique evaluation cases by including the use of unlabelled pairs of items (which for a sparse dataset is most possible item pairs) as implicit negative feedback. The chosen evaluation technique is the common evaluation framework used by He et al. (2017) and Koren (2008), where leave-one-out cross-

validation is performed by, for each item, taking one known positive and 100 randomly selected other items (excluding known positives) and judging the ranking algorithm by the ability to rank the known positive highly. The typical threshold used is that the known positive must be in the top 10 results, this Hit Ratio (HR) metric is denoted as HR@10. HR@5 refers to the known positive being in the top 5, and HR@1 as it being the highest rated item. Other metrics used include the mean and median values for the ranks of the known positives across all test cases.

It is notable that due to the random selection of negative items results may vary between runs. To ensure the results are representative each known positive is tested against multiple random pools of implicit negatives. This significantly increases the compute time required for evaluation but minimises variation in scores between runs.

Having a fixed number of items in each evaluation and repeating with different random sets of items makes this metric well suited to datasets with uneven label distribution such as IWSC. Additionally, the values can be understood intuitively as the random-algorithm performance for any HR@n is approximately n%, with ideal performance always being 100%. Mean and median positive label rank is in the range of 0 to 100, and for a random-algorithm would tend towards 50.

3.6 Transitive Semantic Relationships

To investigate SRQ1: "How can machine understanding of text be used to identify relationships between documents in large collections of unstructured text", a new model has been developed to solve the four IWSC prediction tasks pertaining to extremely sparse labelling and subset labelling (section 3.4) and additionally looks at cold starts. As these scenarios present difficulties arising from having very small numbers of labels, the use of document features is essential for good performance, and in the case of cold starts, is required for better than random performance.

The new model is named "Transitive Semantic Relationships" (TSR) and uses item content information for unsupervised comparison of items to expand the coverage of the few available labels. This is conceptually similar to other embedding based hybrid recommenders such as Vuurens et al. (2016) and He et al. (2017) but uses a novel approach that combines item content embeddings with inferential logic instead of learned or averaged user embeddings, making it suitable for datasets with fewer labels and producing provenance that is both intuitively understandable and easy to visualise.

3.6.1 Theory

Transitive Semantic Relationships are based on an apparent transitivity property of many types of data items, where it is the case that items which are described similarly are likely to have similar relationships to other items. Take, for example, supply-chain: if company A, a steel mill and company B, a construction firm are known to have the relationship A supplies (sells to) B, it may be inferred that some other companies C, another steel mill, and D, another construction firm, might have a similar relationship. Given content information about each company, such as a text description of their product or market role, and the example relationship $A \rightarrow B$, we can infer the potential relationships $C \rightarrow D$; $A \rightarrow D$; $C \rightarrow B$. This is illustrated in Figure 3.2.

It follows that the greater the similarity between an item of interest and an item in a known relationship, the greater the confidence that the relationship is applicable. Given some fixed-length vector representation of the content information about each item, cosine similarity can be used to measure similarity between the items. The vector representation should ideally capture semantic features of the content information that indicate whether the items they describe are similar in function in terms of the known relationship. If the vector representations fulfil this criterion, then the cosine similarity between two items is their semantic similarity. It then follows that the confidence that some query item and some target item share a relationship can be determined by measuring the cosine similarity of the content vectors for the query and the target with another pair of items that are known to share a relationship of the type of interest.

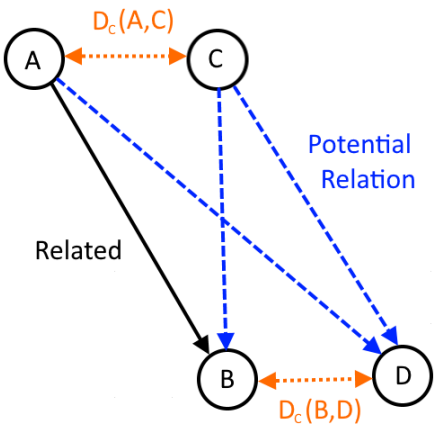


Figure 3.2 - Illustration of Transitive Semantic Relationships. The dotted lines labelled $D_C(A, C)$ and $D_C(B, D)$ represent the cosine distance between the content embeddings of items A and C, and B and D respectively

Herein cosine distance (equation 3.1), (where u and v are the content embeddings) is used rather than the similarity as it is easier to interpret when results are visualised and when distance values are weighted; other distance metrics could be substituted if suitable for the content embeddings.

$$distance(u, v) = 1 - \frac{u \cdot v}{||u||_2 \, ||v||_2} \qquad 3.1$$

To keep scores in the same range as the distance function when combining the two distances of the query and the target from the labelled pair, take the sum of the distances over 2, this value is the combined-semantic distance, shown in equation 3.2, where $D_C(Q, S)$ is the cosine distance of the query item Q and an item S, which shares a relationship with another item R, which is distance $D_C(R, T)$ from the target T.

$$combined\ semantic\ distance = \frac{D_C(Q, S) + D_C(R, T)}{2} \qquad 3.2$$

To obtain a confidence value where 1 is full confidence of applicability and 0 is no confidence, subtract the combined-semantic-distance from 1, this value is the TSR Confidence score for the route, given in equation 3.3. For embeddings supporting negative values, cosine similarity and the resulting TSR confidence is in the range -1 to 1, where negative values suggests confidence against (as opposed to 0 meaning they are orthogonal).

$$TSR\ Confidence = 1 - \frac{D_C(Q, S) + D_C(R, T)}{2} \qquad 3.3$$

Continuing from the prior example illustrated in Figure 3.2, if the cosine distance of A and C is $D_C(A, C)$, and the distance of B and D is $D_C(B, D)$, the confidence for each inferred relationship can be calculated as shown in equations 3.4, 3.5, and 3.6. In equations 3.4 and 3.5 a $+0$ is

included to represent $D_C(A, A)$ and $D_C(B, B)$ respectively, because the cosine distance between an item and itself is always 0. In this example, $A \rightarrow D$ is an item-wise cold-start (D has no known labels), $C \rightarrow B$ is a user-wise cold-start (C has no known labels), and $C \rightarrow D$ is a double cold-start (both C and D have no known labels).

$$A \rightarrow D = 1 - \frac{0 + D_C(B, D)}{2} \qquad 3.4$$

$$C \rightarrow B = 1 - \frac{D_C(A, C) + 0}{2} \qquad 3.5$$

$$C \rightarrow D = 1 - \frac{D_C(A, C) + D_C(B, D)}{2} \qquad 3.6$$

To further illustrate this, if C is very similar to A, for example, $let\ D_C(A, C) = 0.2$, but D was only slightly similar to B, $let\ D_C(B, D) = 0.8$ then calculation shows that: $A \rightarrow D = 0.6$; $C \rightarrow B = 0.9$; $C \rightarrow D = 0.5$ indicating that there is high confidence that C could share a similar relationship with B as A does, but other new relations have low confidence. In another example, if C remains similar to A, $let\ D_C(A, C) = 0.2$, but D is made more similar to B, $let\ D_C(B, D) = 0.3$, then the calculation gives: $A \rightarrow D = 0.85$; $C \rightarrow B = 0.9$; $C \rightarrow D = 0.75$, showing that while all relationships have high confidence, higher confidence scores are awarded when there is greater similarity to the labelled pair.

3.6.2 TSR as a Recommender System

The previous scenarios suppose that the items of interest for comparison are already pre-determined. However, the principle of TSR can be extended to the selection of items for comparison, given an input item to use as a query. This query is not a written question or search term as in traditional search engines but is instead content information for an item for which we want to find relations (e.g., an item description). This approach can be used as a recommender system to produce a ranked list of recommended items for the query (which in recommender system terminology would be the “user”). The approach described in this section for applying TSR as a recommender system is illustrated in Figure 3.3 and pseudocode is given in Figure 3.4.

There is a distinction between cases where relationships map from one space to some other non-overlapping space, for example, separate document collections, and the alternative case where items on either side of the relationship co-exist in the same space. A practical example of the

former might be a collection of resumes and a collection of job adverts, while an example of the latter might be descriptions of companies looking for supply chain opportunities, as in the IWSC dataset. The TSR scoring does not differentiate between these two dataset types, but in the former case, with separate item collections, it is only necessary to make distance calculations between items in the same collection and irrespective of the total number of collections, it is only necessary to examine the collections featuring items on either end of at least one example of the relationship type of interest; this may be a useful filtering criterion in datasets featuring many types of relationships across many non-overlapping collections.

Having identified the collections that are of interest, additional filtering of items can be applied before distance calculation, such as by using item meta-data or additional auxiliary information, for example, only considering recent information, or limiting by language or region. This filtering could be done to the list of known relationships, if, for example, historical trends are not of interest, or could be applied to potential targets, for example, ignoring content in a different language to the query item.

The next stage is to calculate the distances between the query item and other items in the same collection which are members of relationships of the type to be inferred, items not in such relationships are not of interest. The distance between the query and each of these "similar nodes" is then calculated, the distance of each is referred to as $D1$. If the number of similar nodes is large, a limit $L1$ can be applied to truncate the list of similar nodes, preferring the least distant.

Next, all items pointed to by a known relationship of a similar node are examined, referred to collectively as the "related nodes". The distance between each related node and every other eligible "target node" in that space is then calculated, the distance of each is referred to as $D2$. An item can be both a related node and a target node ($D2 = 0$), but an item cannot be both the query and a target node. If the number of target nodes is large, the number of comparisons in the next stage can be controlled by imposing a limit $L2$ on the maximum number of target nodes for each related node, preferring the least distant.

Alternative scoring approaches are discussed in section 3.9, but a simple scoring metric equivalent to the examples in the previous section is to determine the score for each target node by finding the route for each with the greatest TSR Confidence (equation 3.3) $1 - (D1 + D2) \,/\, 2$ that creates a path to it from the query item, where $D1$ is the distance between the query and an item in the query's space (the similar node), which shares a relationship with an item in the target's space (the related node) which is of distance $D2$ to the target node. This scoring system ranks items by the least combined semantic distance from a known relationship of the desired type, that is, by the best route as measured by TSR Confidence.

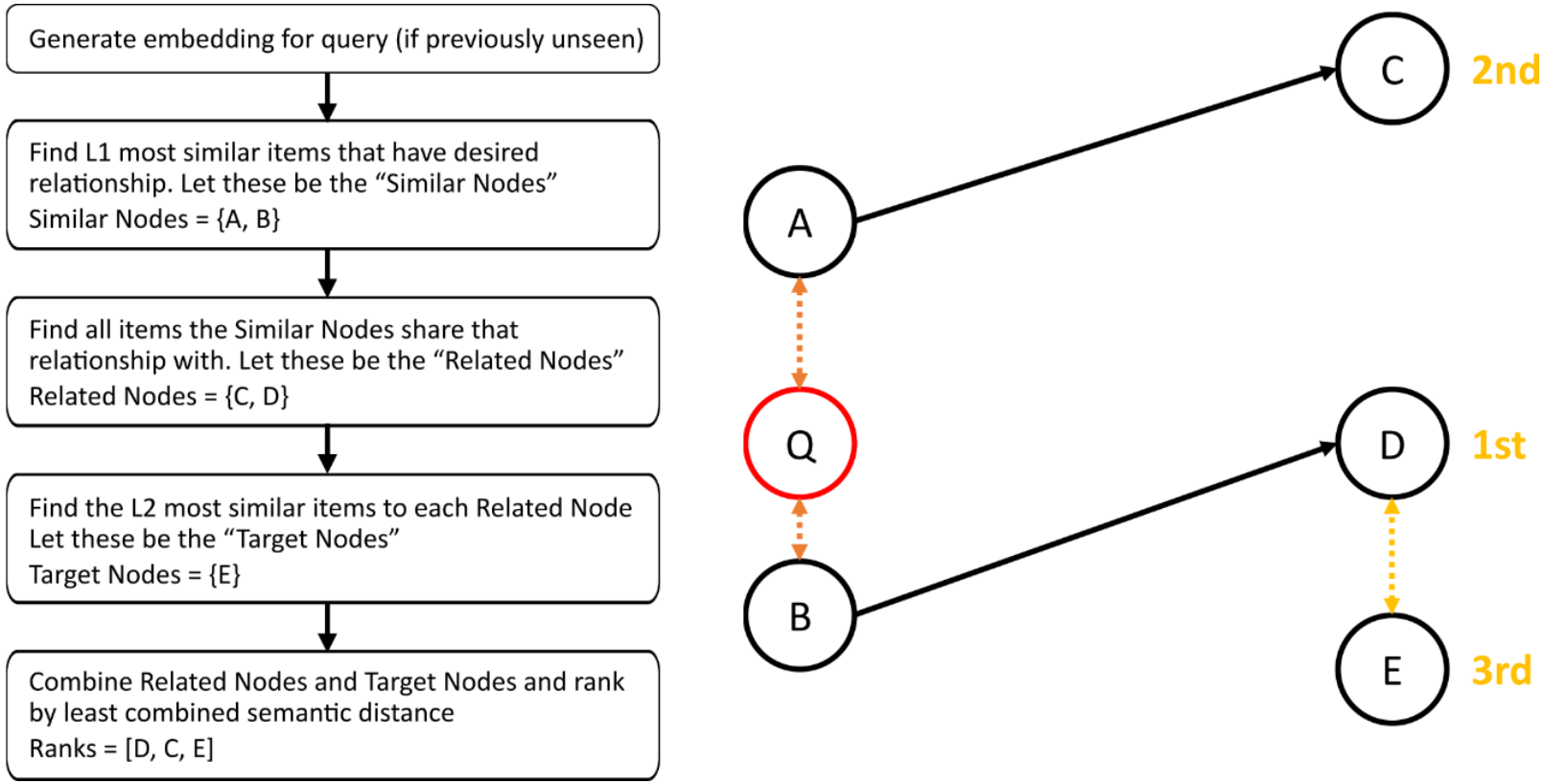


Figure 3.3 - An illustrated example showing steps in the TSR recommendation algorithm

```
let L1 be the number of similar nodes to check
let L2 be the number of related nodes to check for each similar node
let Q be the input document
let SIMILAR_LIST be an array of the L1 labelled documents least distant to Q
let OUTPUT be an empty list of tuples
for each node SIMILAR_NODE in SIMILAR_LIST
   let D1 be the distance between Q and SIMILAR_NODE
   let RELATED_LIST be all the nodes sharing a relationship with SIMILAR_NODE
   for each node RELATED_NODE in RELATED_LIST
      let RELATED_SCORE = 1 - D1  / 2
      add to OUTPUT the tuple ( RELATED_NODE , RELATED_SCORE )
      let TARGET_LIST be an array of the L2 documents least distant to RELATED_NODE
      for each node TARGET_NODE in TARGET_LIST
         let D2 be the distance between RELATED_NODE and TARGET_NODE
         let TARGET_SCORE = 1 - (D1 + D2) / 2
         add to OUTPUT the tuple ( TARGET_NODE , TARGET_SCORE )
      end for
   end for
end for
for all duplicate first values in OUTPUT, keep only the tuple with the greatest second value
sort OUTPUT by the second value of each tuple in descending order
return OUTPUT
```

Figure 3.4 - Pseudocode for using TSR as a recommender system. Outputs a list of (item, score) tuples in descending order of score where higher scores are more strongly recommended. These scores are the TSR Confidence of the route with the least combined-semantic-distance for the item.

3.7 Development and Experiments

Development of TSR began during a one-month industrial placement in October 2018. During the placement, it was agreed with the industrial partner that the task of supply chain relationship discovery for businesses on the Isle of Wight would be the initial testing scenario for the proposed method as this is data for which they could provide expert labels and good results would have the potential for real-world applications within the business.

The data was received from the industrial partner in two stages; first, annotated with the labels which would form the extra-sparse (ES) tasks, and second, annotated with the labels for the subset-labelled (SL) task. A more detailed description of the dataset is given in section 3.4.

3.7.1 Validation of Assumptions

Early development focused primarily on finding support for the conjecture that semantic embeddings of the descriptions of companies can indicate whether they are potential competitors. For this purpose, a Python script was prepared that would, using a pre-trained distribution of Universal Sentence Encoder (TensorFlow Hub, 2018), generate embeddings for the descriptions of all companies in the dataset, and then calculate the average pairwise cosine distance for items sharing each relationship.

The results, shown in Table 3.4, show a significantly lower average distance for known competitors compared to all other known relationships, demonstrating that the conjecture is correct, that cosine similarity of description embeddings can be used as an indicator of companies being competitors.

The results also show that the average distance for items known to share any supply chain relationship (consumers or suppliers) is slightly less than the average for items known not to, however, the difference is much smaller than that of competitors and non-competitors. This shows that textual similarity is a weak indicator of supply chain relationships other than competitors but is a strong indicator for whether companies are potential competitors.

This script was then extended to visualise the label distributions, shown in Figure 3.5. USE embeddings are 512-dimensional vectors, so require dimensionality reduction to be plotted for visualisation. Initially, Principal Component Analysis (PCA) (Hotelling, 1933) was used, but further experimentation found that t-SNE (Maaten & Hinton, 2008) produced more pronounced clusters. This result is not unexpected, as t-SNE is well suited to reducing very high dimensional data.

Figure 3.5 is a distribution plot that shows visible, although not clearly separable, clusters in the embeddings, with competitor labels more often staying within the same neighbourhood but consumer and supplier labels often connecting to more distant regions, which explains the distance values in Table 3.4. This supports the hypothesis that a relationship transitivity method could be used to 'bridge' between these areas, by using semantic similarity as an indicator of supply chain similarity (i.e., likelihood of being competitors).

Subfigures A and B also illustrate the difference between the labelling sets in the IWSC dataset, where labels in the Subset-Labelled (SL) set offer less coverage of the dataset but better describe particular items, and labels in the Extremely-Sparse (ES) set provide greater coverage but offer little information for each labelled item.

A similar approach is used for visualisation later in the project in Chapter 4, where it is discussed in more detail. The distribution of items in the IWSC dataset, and particularly the clusters within the data are revisited in Chapter 4, section 4.9.

Table 3.4 - Cosine distance of labels. Lower values indicate items in the relationship have more similar descriptions.

Label Name	Mean Cosine Distance	Median Cosine Distance
SL_competitors	**0.40**	**0.39**
SL_not_competitors	0.44	0.43
SL_suppliers	0.47	0.47
SL_not_suppliers	0.46	0.47
SL_consumers	0.47	0.45
SL_not_consumers	0.49	0.48
ES_competitors	**0.33**	**0.30**
ES_suppliers	0.56	0.58
ES_consumers	0.54	0.56
ES_unrelated	0.58	0.60

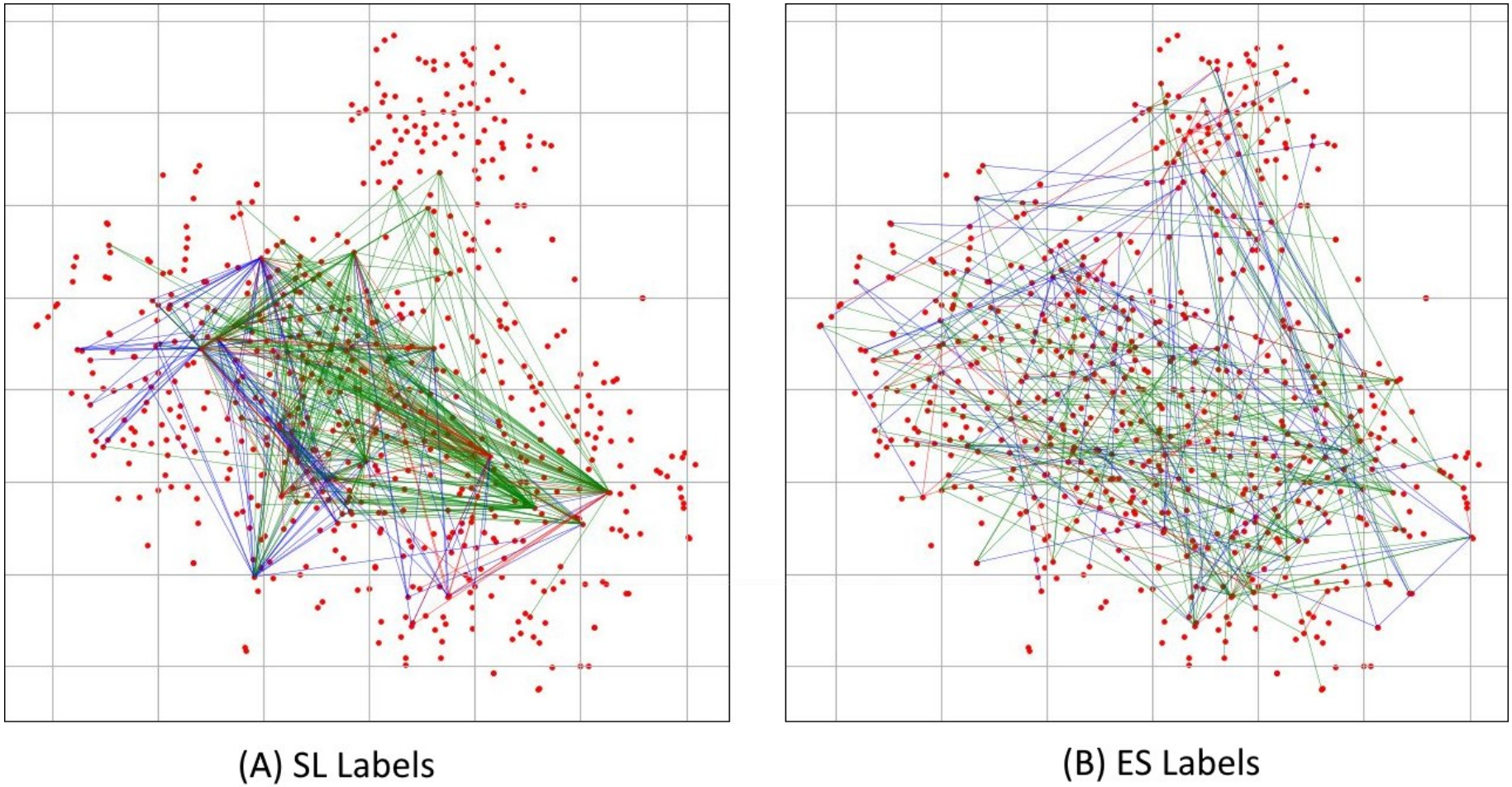


Figure 3.5 - A 2D t-SNE plot of ISWC item description embeddings showing known relationship labels for competitors (red), consumers (green), and suppliers (blue). Subfigure A shows the SL labelling set. Subfigure B shows the ES labelling set.

3.7.2 Implementing TSR

Next, an initial implementation of the relationship transitivity method would be produced, named the Transitive Semantic Relationships (TSR) model. Pseudo-code for this had been drafted while waiting for data from the industrial partner (as shown before in Figure 3.4). At this stage, two different implementations were being considered: a clustering model; and a distance model, which would rank results as follows:

The clustering model would: “Rank targets by the number of relationships with one member sharing a cluster with the query item and one member sharing a cluster with the target”.

The distance model would: “Rank targets by the semantic distance from the input to the target passing through exactly one relationship which is given as a distance of 0”.

A limitation of the clustering model is that without an additional scoring metric, for a given query item, target items within the same cluster would have the same number of routes, and therefore the ranking algorithm cannot meaningfully order them as they would all have the same score. Additionally, the method used for clustering would significantly affect the quality and interpretability of results.

The distance model avoids these issues but requires evaluating a large number of routes per possible target (potentially as many routes as there are known relationships), and additionally would not be able to assign different scores to targets that share a relationship with the same competitor of the query item, as their distances would all equal the distance between the query and the competitor +0.

The solution implemented is a combination of these two methods, with targets ranked by distance, but selected using limits similar to the clustering model (although using nearest neighbours rather than clustering). This reduces the number of routes that must be evaluated per target and ensures decidability of rank in most cases. The alternative scoring algorithms outlined in section 3.9 are different combinations of these two methods, where a trade-off is made between prioritising distance or number of routes.

The initial implementation of TSR was a command-line tool that would allow the selection of an item from the dataset as a query, or input of custom text. In both cases, this was treated as an unseen item with no existing relationships. The output would be a ranked list of targets with routes and scores. A later variant of this tool would also output 2D and 3D t-SNE plots visualising the routes for the top-ranking items, as shown in Figure 3.6 and Figure 3.7. The 3D plots can be

manipulated in a web browser and support interactive inspection of routes and nodes, allowing easy exploration of the provenance and results.

3.7.3 Visualisation and Provenance

Chapter 2, section 2.5 discussed the benefits of transparency and provenance for improving accountability and user confidence in automated systems, particularly “black box” learning-based systems which can be subject to bias and omission (Caliskan et al., 2017; Nadeem et al., 2021). This closely relates to SRQ3: “How can the results of text analysis be effectively presented and used to inform decision-makers, analysts, and organisations?”.

TSR makes use of a ‘black box’ upstream embedding model to produce ‘white box’ recommendations. While it is not possible to plainly describe why any two items are considered similar, the working of the algorithm in all later stages, such as items and known relationships considered, and the weighting of each, are fully transparent. This way, the reasoning behind a recommendation can be simply explained and visualised by showing the key items and relationships that informed it.

Figure 3.6 and Figure 3.7 show visualised examples of TSR routes for the top-ranking items for a query. The evaluation software can also produce interactive plots (viewable with a web browser) which allow inspection of individual routes and the relevant items and labels, allowing some insight into the behaviour of the scoring algorithm. The output of TSR also includes the full list of routes considered in evaluating the query, ordered by their TSR Confidence scores.

3.7.4 Optimisations

To prepare for a more extensive investigation of the TSR approach, the previously described tools were refactored into separate program modules, specifically: data preparation, the TSR ranking algorithm, and output/visualisation of results.

As the subject of the investigation was the TSR ranking algorithm, a pre-processing module was added which generates embeddings for all item descriptions and pre-calculates the cosine distance between all possible pairs and stores them as a lookup table for each item. This allows the TSR ranking algorithm to be rapidly tested on many different subsets of the data, such as for cross-validation, without needing to re-calculate item similarity between runs. This significantly reduces compute and memory requirements for running TSR evaluations without affecting performance.

3.7.5 Evaluation Toolkit

The results presented in the next section were produced using the TSR evaluation toolkit. This is a series of modules created for rapid evaluation of the TSR ranking algorithm using a variety of techniques from the literature. Section 3.5 provides a detailed discussion of the evaluation techniques employed.

The toolkit is a command-line tool that takes several configuration options, including both evaluation conditions, such as repeat count for random pools, and TSR parameters, including values for L1 and L2 (see section 3.6), and selection of scoring algorithm (section 3.9). The output is a CSV file containing the experimental parameters, dataset statistics, and evaluation scores.

3.7.6 Hyperparameters

TSR supports two hyperparameters: L1 is the number of nearest neighbours of the query node for which to score routes, and L2 is the same but for each related node (for details see section 3.6.2). These values are limits imposed to prevent excessive computation beyond what is needed to score the best routes. For recommender systems usually only the top scoring items are of interest, as is reflected in the evaluation metrics typically used (see section 3.5), in this case it is therefore usually unnecessary to exhaustively score all items in a dataset if the algorithm is capable of early stopping after identifying the highest scoring items, as is the case with TSR.

When using the least-combined-distance scoring method described previously, it is only necessary to consider a small number of routes if items are inspected in ascending order of semantic distance (for query node to similar node, and for related node to target node) as it is necessarily the case that less distant items score higher than more distant ones. Some alternate scoring algorithms discussed in section 3.9 consider multiple routes for each item, so less optimal routes may have some impact on results, but generally the impact of less favourable routes is small and a small number of very good routes dictate the top scoring items.

Experimentation shows that with values of $L1 \geq 5$ and $L2 \geq 10$ there is no change in the ranking of the 10 highest scoring items in the IWSC dataset for several randomly chosen queries; these are the values used throughout this chapter. With these values, the only effect of greater values is that a greater number of items are given scores (although the additional items always score lower than items included by smaller values of L1 and L2), this does not impact the top-ranking items but does affect mean-positive-rank during implicit feedback because sometimes the known-positive may not receive any rank if it would otherwise (with higher values of L1 and L2) receive a poor rank. During implicit feedback, unranked items are considered to have the worst possible

rank, so cases where the known positive is not ranked inflate the mean-positive-rank. For this reason, median-positive-rank is a more reliable (but less granular) metric.

The number of routes TSR will calculate is at most the sum of the number of known relationships for the L1 nearest neighbours of the query (ignoring unlabelled items) multiplied by L2.

E.g., If $L1 = 2$ and $L2 = 10$ and the two labelled items most similar to the query have 2 and 3 known relationships, then the total number of routes calculated is $(2 + 3) \times 10 = 50$.

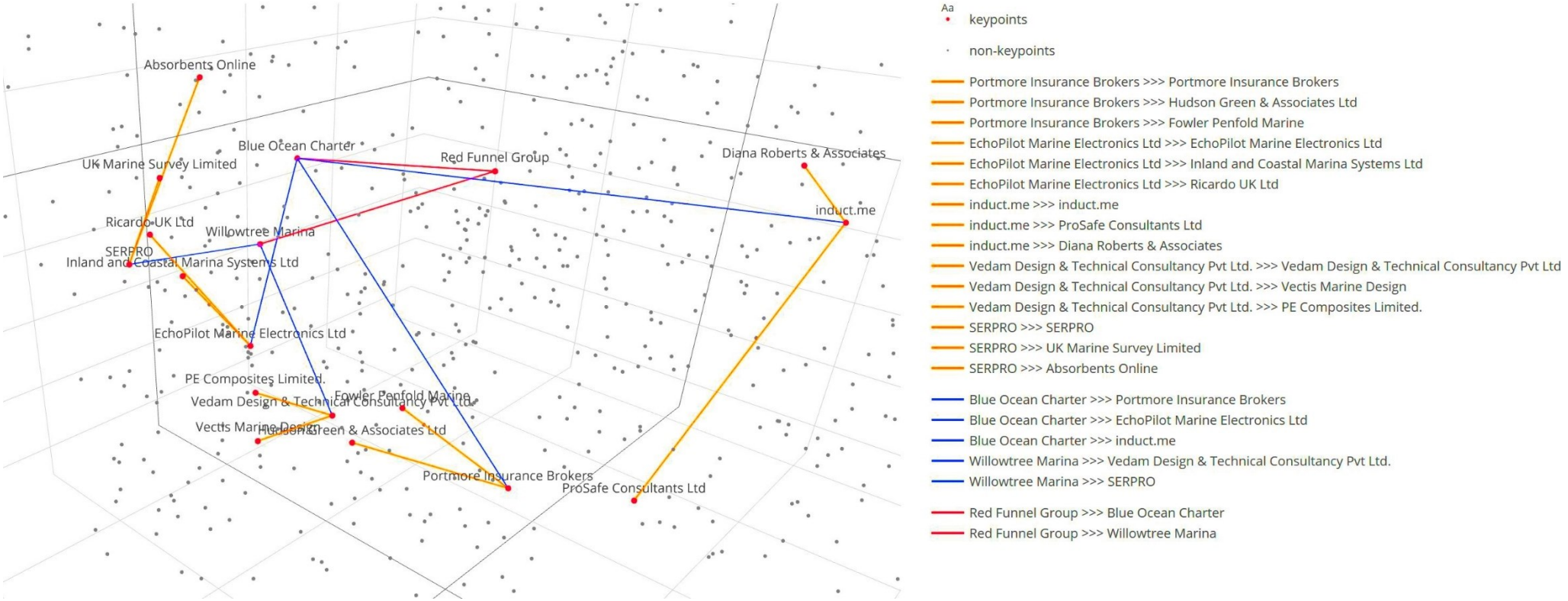


Figure 3.6 - A 3D visualisation of a TSR query showing labelled and inferred relationships considered for the top-ranked items. Each route is comprised of three lines: query node → similar node (red), similar node → related node (blue), related node → target node (yellow).

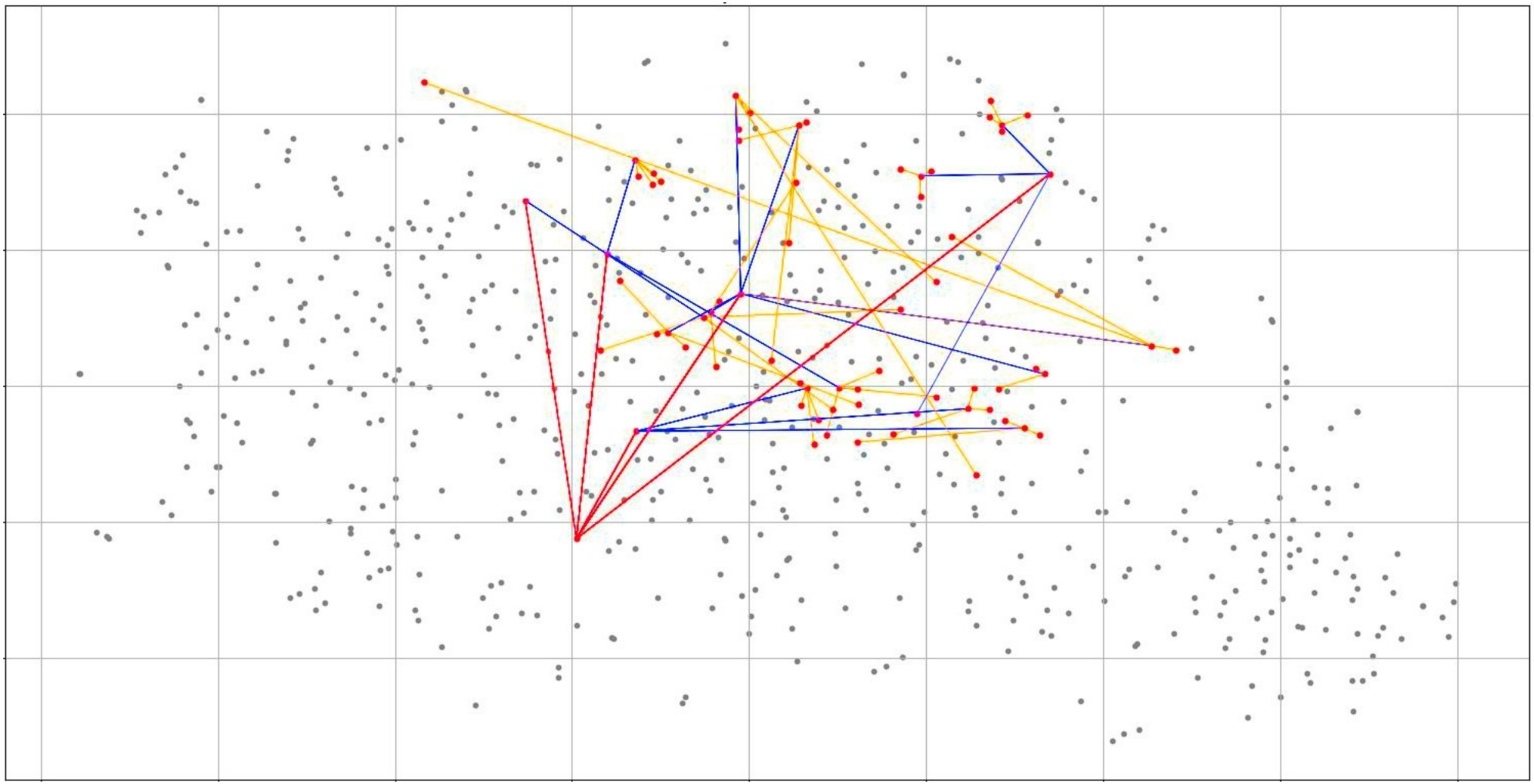


Figure 3.7 - A 2D visualisation of a TSR query showing labelled and inferred relationships considered for the top-ranked items. Each route is comprised of three lines: query node → similar node (red), similar node → related node (blue), related node → target node (yellow).

3.8 Results of TSR on IWSC

This section evaluates the performance of TSR on the IWSC tasks, using the most suitable evaluation metrics discussed in section 3.5, including both explicit and implicit feedback techniques.

When considering error metrics (RMS error and Median absolute error), it is notable that scores awarded by TSR have no guarantee of symmetric distribution over the possible output range and are typically concentrated towards high-middle values due to averaging similarity scores making extreme values uncommon. Figure 3.8 shows the typical score distribution for the standard TSR algorithm TSR-a using the least-combined-semantic-distance metric.

Section 3.9 details some alternative scoring algorithms with unbounded upper values. A scaling function can be applied after scores are calculated to fit them to a specific range, but this still does not guarantee the desired distribution and could be sensitive to outliers, such as unusually high scoring items, distorting error values.

In this evaluation, item similarity is computed using cosine similarity of Universal Sentence Encoder (USE) embeddings of item descriptions. USE was chosen as it shows good performance on a range of existing downstream tasks (Cer et al., 2018). It is also of particular interest that this model was fine-tuned on the SNLI dataset (Bowman et al., 2015), a set of sentence pairs labelled as contradiction, entailment, or unrelated. It seems likely that this may require the model to learn similar linguistic features as are needed for the supply chain inference task as the ability to discern whether pairs of descriptions are entailed or contradictory is essential to human judgements for this task, in particular, in determining if companies serve similar supply chain roles. A detailed investigation of the effects of upstream embedding models is left to future work (see Chapter 5, section 5.4.2).

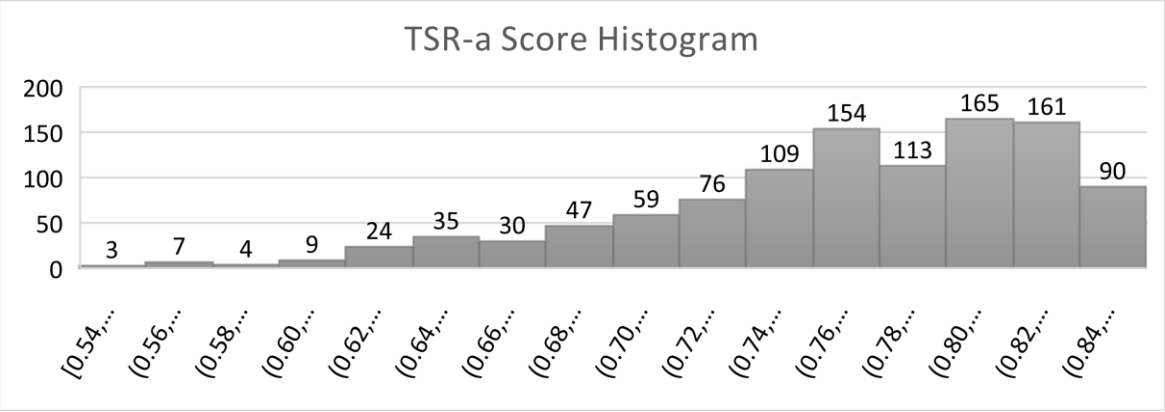


Figure 3.8 - Histogram of item scores produced by TSR-a

3.8.1 Results for Subset Labelling Tasks

Table 3.5 and Table 3.6 show the results of TSR on the two IWSC-SL tasks introduced in section 3.4. In these experiments, the scoring metric used is least-combined-cosine-distance, as described in section 3.6 and the evaluation metrics used are as discussed in section 3.5. All experiments are cold-start scenarios where the input (query) item is treated as unseen, only the USE embedding of its description is known.

The TSR parameters are set as follows: $L1 = 5$ and $L2 = 10$. For this scoring metric the value of these parameters has little impact on performance as only the best routes contribute to scoring, but it is observable that this inflates the mean positive rank as items lacking good routes are excluded from the results. Items missing from the results are given the worst possible rank.

For the implicit feedback evaluations (HR and Positive Rank), one known positive and a random pool of 100 not-known-positive items are used. This process is repeated 10 times for each label, using different random pools, and scores are calculated across all tests. Therefore, the number of test runs is always 10 times the number of positive labels. The number of labelled items and positive labels used in the implicit feedback tests is greater than for explicit feedback, as implicit feedback allows testing of items that lack any known negatives.

The results show good performance on the IWSC-SL tasks, considering how few labels are available, achieving a hit-rate@10 of over 75%. Notably, performance is less than 9% worse on the SL_suppliers test despite having less than half the number of labels, showing that the algorithm can achieve good performance on labelled-subset tasks even when extremely few labels are available (142 labels in a dataset of 630 items). For both IWSC-SL tasks the frequency of the top-ranked item being the known positive (when competing with 100 randomly selected others) HR@1 appears similar and is 14-15 times better than random.

3.8.2 Results for Extra Sparse Labelling Tasks

Table 3.7 and Table 3.8 show the results on the two IWSC-ES tasks introduced in section 3.4. The algorithm and parameters are the same as used for the IWSC-SL tasks. The IWSC-ES tasks each have around half the number of positive labels as IWSC-SL, so poorer scores are expected.

The IWSC-ES results show significantly worse hit-rate, but smaller median absolute error and RMS error. This may be the result of lack of dense regions in the labels, due to the extreme sparsity and random distribution, making identifying a particular known positive more difficult, but the better error values and F1 score indicate that the predicted scores are still effective for discerning good and bad results despite being less effective at a ranking a given good result highly.

Table 3.5 - Explicit feedback evaluation of TSR-a on the IWSC-SL tasks

Positive Label Name	Labelled Items	Positive Labels	Negative Labels	F1 @R	RMS Error	Median Absolute Error
SL_consumers	16	375	712	0.520	0.204	0.688
SL_suppliers	15	142	525	0.477	0.234	0.682

Table 3.6 - Implicit feedback evaluation of TSR-a on the IWSC-SL tasks

Positive Label Name	Labelled Items	Positive Labels	HR @10	HR @5	HR @1	Median Positive Rank	Mean Positive Rank
SL_consumers	17	376	0.752	0.510	0.146	4	7.8
SL_suppliers	15	142	0.663	0.543	0.150	4	14.0

Table 3.7 - Explicit feedback evaluation of TSR-a on the IWSC-ES tasks

Positive Label Name	Labelled Items	Positive Labels	Negative Labels	F1 @R	RMS Error	Median Absolute Error
ES_consumers	39	115	198	0.549	0.167	0.560
ES_suppliers	46	90	259	0.350	0.177	0.572

Table 3.8 - Implicit feedback evaluation of TSR-a on the IWSC-ES tasks

Positive Label Name	Labelled Items	Positive Labels	HR @10	HR @5	HR @1	Median Positive Rank	Mean Positive Rank
ES_consumers	51	207	0.221	0.119	0.018	36	43.0
ES_suppliers	48	92	0.197	0.129	0.055	32	47.7

3.9 Alternative Scoring Algorithms

The least-combined-semantic-distance scoring algorithm introduced in section 3.6 and used in the previous results sections is relatively simple to calculate and is both intuitive and easy to visualise (see Figure 3.5 and Figure 3.6). However, as only the shortest route to a target is considered, it does not factor in supporting evidence. For example, in the case of two targets with highly similar shortest distances from the query, if one had multiple short routes and the other had only one short route, it is intuitive that more confidence can be had in recommending the target with greater supporting evidence.

An illustrated example is given in Figure 3.9, in that example, target nodes A and C are the same semantic distance $D1$ from Q, A is supported by two known relationships whereas C is supported by only one; it is intuitive that in case of otherwise equal scores, the node with greater evidence should be preferred. However, when ranking targets B and D, the combined semantic distances are not the same as the values for $D2$ ($D_C(A,B)$ and $D_C(C,D)$ respectivly) differ. When considering only the best route, target D would be preferred as the semantic distance is less, however, this ignores the fact that B is supported by more routes. If the distance $D_C(A,B)$ was only marginally larger than $D_C(C,D)$ then the target with more routes, B, might be a better recommendation than C due to evidence of additional routes, even though the least-combined-semantic-distance is larger. Neither of these scenarios is covered by the least-combined-semantic-distance scoring algorithm described previously, as additional routes are not considered.

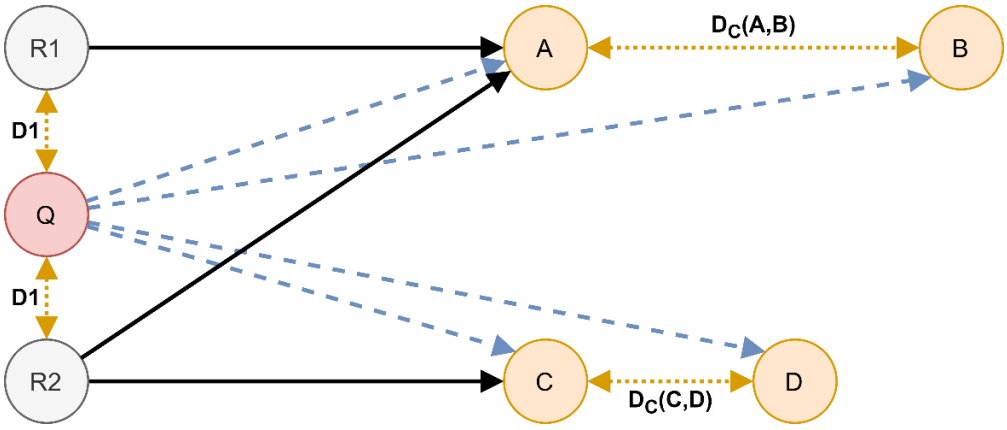


Figure 3.9 – Illustration of a scenario where multiple TSR routes exist for a target. Related nodes $R1$ and $R2$ are an equal distance $D1$ from the query node Q. Nodes A, B, C, D are possible target nodes. There are known relationships $R1 \rightarrow A$; $R2 \rightarrow A$; $R2 \rightarrow C$.

Several variations of the scoring algorithm have been tested which boost the score when multiple good routes to the target are found. These approaches include boosting the score based on the number of routes (TSR b and c), taking the weighted sum of the scores for each route (TSR d, e, f, g, h, k, l, m, o, p, and q), and taking the sum of scores for each route but increasing the significance of distance (e.g., distance squared or cubed) (TSR i, j, and n). The results for some of these tests for the SL_consumers task is shown in Table 3.9 and a comprehensive comparison across all tasks is shown in Figure 3.10.

As these algorithms produce scores with different ranges, a simple scaling algorithm is applied as shown in equation 3.7.

$$f(s_i) = \frac{s_i - \min(s)}{\max(s) - \min(s)} \qquad 3.7$$

The scaling algorithm does not modify the order of results but ensures scores are within the same 0-1 range as the labels to make them suitable for error measurement. TSR-a produces scores in the range 0-1 without scaling for positive vectors, but a scaled version TSR-a* is also included for comparison as TSR-a rarely gives scores close to its bounds (see Figure 3.8).

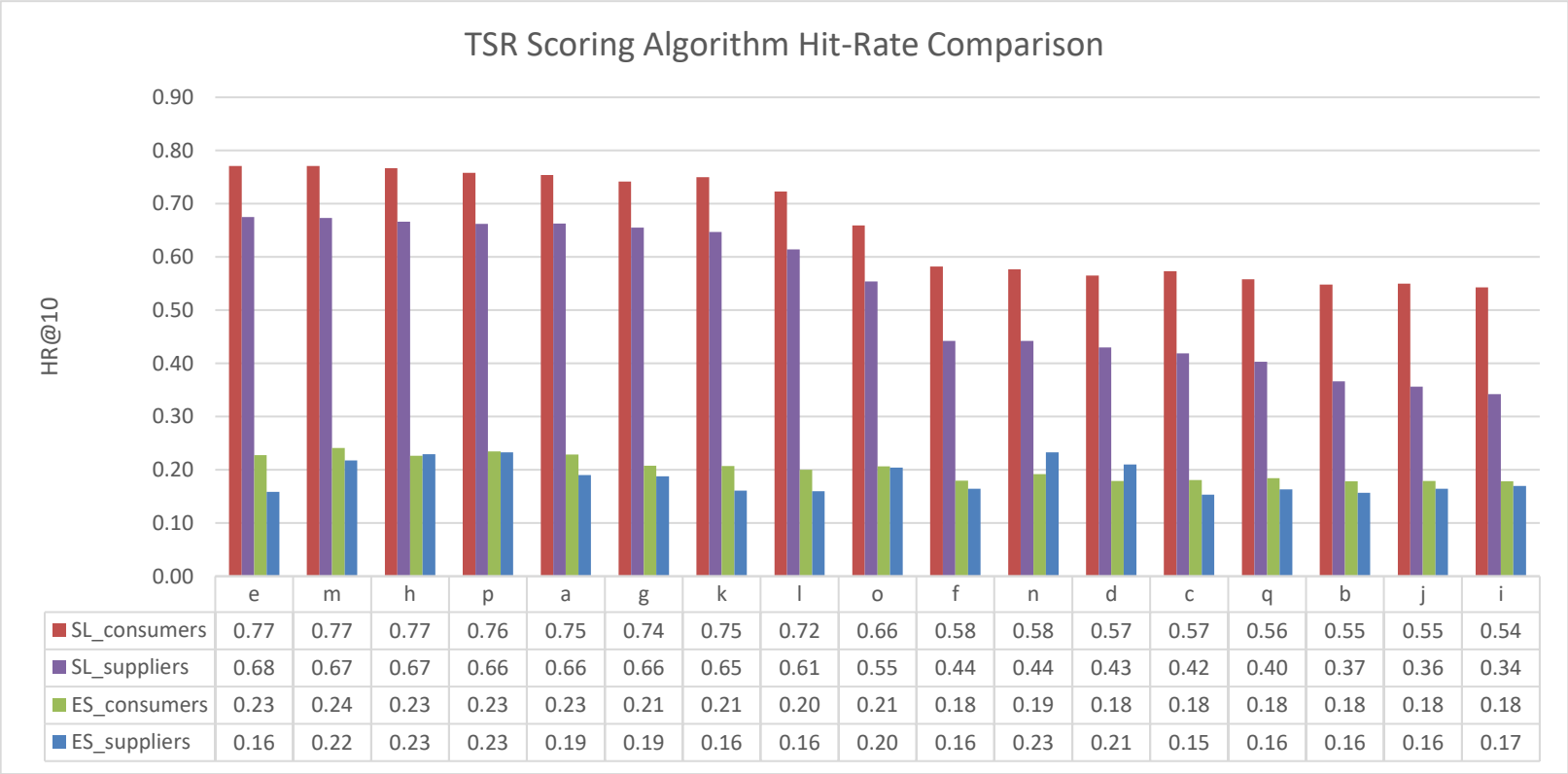
The results show a notable stratification with some algorithms performing similarly to TSR-a, and some significantly worse. The scoring metrics that perform better show a slight improvement in HR@10, but a proportionally larger improvement in HR@5 and HR@1. Examination of the results using the provenance generated by TSR shows that TSR-a is sometimes indecisive in ordering the top-ranking items, with multiple items receiving the same score. This explains TSR-a's proportionally worse performance when looking only at the top-ranked item, as in some cases TSR-a might rank several top-scoring items arbitrarily. Consideration of additional information for each target removes this indecision, allowing the alternative scoring algorithms to order the top-scoring items meaningfully.

The best performing algorithm for the IWSC-SL tests is TSR-e, where the target score is calculated as the sum of the score for the best route and half the score of the second-best route. This produced absolute improvements of 1.7%, 2.3%, and 1.8% for HR@10, HR@5, and HR@1 for the SL_consumers task, which is a relative improvement of 2.3%, 4.5%, and 12%. However, this scoring algorithm has the disadvantage of having a score distribution concentrated towards middle values as extreme values would require either all routes to be very poor or both routes to be very good, which is less common than only the best route being very good or bad. This may account for its comparatively high error values as error measurements will be high even for a correct ordering if values are concentrated towards the mid-range, due to comparison with binary labels.

Another well-performing algorithm is TSR-m, as given in equation 3.8, where r is the number of routes to the target, i is the rank of each route (where $i = 1$ is the route with the least combined-semantic-distance), and d_i is the combined-semantic-distance (equation 3.2) of route i. The scaling function is omitted for clarity as it is already given in equation 3.7. Scaling is applied once all score values have been calculated. This algorithm considers all routes to a target but with significance diminishing by the cube of the route's rank (e.g., the best route adds $1/d$ to the score, the second adds $1/8d$, then $1/27d$, etc.).

$$S = \sum_{i=1}^{r} \left(\frac{1}{d_i i^3} \right) \qquad 3.8$$

The algorithms TSR-o and TSR-p are the same as TSR-m except that the exponent of the route's rank, which the score is divided by, is 1 and 2 respectively; these variations perform significantly worse. It is interesting that when penalising the contribution of additional routes performance is sub-standard when the penalty is small, but above-standard when it is large. This suggests that some ideal penalty function exists where additional routes do not overpower the normal scoring but still provide support in closely scored cases. It is possible that the best scoring penalty is a property of the distribution of the data and labels, and that the ideal penalty function may be dependent on the dataset. Testing of this property on other datasets and alternative penalties for this dataset is left to future research.



	e	m	h	p	a	g	k	l	o	f	n	d	c	q	b	j	i
SL_consumers	0.77	0.77	0.77	0.76	0.75	0.74	0.75	0.72	0.66	0.58	0.58	0.57	0.57	0.56	0.55	0.55	0.54
SL_suppliers	0.68	0.67	0.67	0.66	0.66	0.66	0.65	0.61	0.55	0.44	0.44	0.43	0.42	0.40	0.37	0.36	0.34
ES_consumers	0.23	0.24	0.23	0.23	0.23	0.21	0.21	0.20	0.21	0.18	0.19	0.18	0.18	0.18	0.18	0.18	0.18
ES_suppliers	0.16	0.22	0.23	0.23	0.19	0.19	0.16	0.16	0.20	0.16	0.23	0.21	0.15	0.16	0.16	0.16	0.17

Figure 3.10 - Comparison of Hit rate of alternative TSR algorithms on all four IWSC tasks

Table 3.9 - Evaluation of alternative TSR algorithms on the IWSC SL_consumers task

Scoring Algorithm	HR @10	HR @5	HR @1	Median Positive Rank	Mean Positive Rank	F1 @R	RMS Error	Median Absolute Error
TSR-a	0.754	0.509	0.145	4	7.7	0.520	0.204	0.688
TSR-a*	0.754	0.509	0.145	4	7.7	0.520	0.195	0.481
TSR-b	0.548	0.364	0.115	8	11.5	0.541	0.120	0.319
TSR-c	0.573	0.385	0.133	7	10.9	0.544	0.120	0.309
TSR-d	0.565	0.373	0.124	7	11.1	0.544	0.122	0.322
TSR-e	0.771	0.532	0.163	4	7.6	0.530	0.204	0.584
TSR-f	0.582	0.408	0.158	7	10.5	0.549	0.146	0.456
TSR-g	0.742	0.536	0.185	4	7.8	0.533	0.192	0.523
TSR-h	0.767	0.538	0.152	4	7.5	0.531	0.196	0.508
TSR-i	0.543	0.362	0.112	8	11.5	0.541	0.121	0.320
TSR-j	0.550	0.359	0.117	8	11.6	0.541	0.120	0.318
TSR-k	0.750	0.538	0.179	4	7.9	0.525	0.207	0.605
TSR-l	0.723	0.529	0.189	4	8.1	0.536	0.189	0.525
TSR-m	0.771	0.530	0.151	4	7.5	0.523	0.170	0.433
TSR-n	0.577	0.385	0.135	7	10.7	0.541	0.121	0.320
TSR-o	0.659	0.466	0.181	5	9.2	0.539	0.143	0.452
TSR-p	0.758	0.533	0.158	4	7.5	0.531	0.165	0.456
TSR-q	0.558	0.372	0.119	8	11.2	0.541	0.120	0.325

3.10 Example Query

To help illustrate the behaviour of TSR, Table 3.10 shows the TSR-e results for the query company "Resmar Marine Safety" using the SL_consumers label set. The parameters used in this example are the same as in the empirical evaluation, and the query was treated as a user-wise cold start (only the description was used).

In this case, four out of the top five results are potential consumers, and the other is unknown (not labelled). For TSR-e, the two best routes are considered when scoring targets. Inspecting the provenance output from TSR shows that, except for "Datum Electronics Limited", the recommendations were primarily based on Resmar Maine Safety's similarity to "Superyacht Doc" and "ProSafe Consultants Ltd", which the top four results are all labelled as potential consumers of. Repeating this query using TSR-a produces a different set of top results. TSR-a considers only the best route for each target when scoring. In this case, the similarity between Resmar Maine Safety and Superyacht Doc is the deciding factor for all of the top results.

It can be seen from these examples, that TSR can make good quality recommendations from little labelled data, and these results are easily explainable. Unlike other hybrid recommender systems where it may not be possible to identify the items and relationships impactful on the ranking of results, for TSR it is trivial to interpret as a list of items and relationships considered, and their weightings, is included in the provenance of the results.

Table 3.10 - Example results for an SL_consumers query using TSR-e for the company "Resmar Marine Safety" (highlighted). The description text for each company is taken verbatim from the IWSC dataset and was originally sourced from the websites of (*IWChamber*, 2018; *IWTechnology*, 2018; *Marine Southeast*, 2018).

Name	**Description Text**	**Known Relation**	**TSR-e Rank**
Resmar Marine Safety	"Resmar specialise in boat safety equipment, fire safety apparatus, and industrial safety equipment. The boating safety equipment we supply includes Life Rafts, Life Jackets and Flotation aids."	Query	-
Caversham Boat Services	"Holiday Boat Hire - Narrowboats and Cruisers, Jetty services, Slipway, Engineering and Mooring"	Consumer	1
Burgess Marine Ltd	"Super yacht refit, 900 ton ship lift, steel and aluminium welding and fabrication, All aspects of commercial ship repair Support of WFSV, commercial ferry industry, Royal Navy Surface Fleet and commercial tonnage."	Consumer	2
Green Marine Solutions	"After completing three successful years on the Greater Gabbard wind farm, the Marine Management team contracted by Fluor to plan, initiate and manage the Marine Coordination Centre have formed Green Marine Solutions. Green Marine Solutions offer three packages to the Offshore Renewable industry:, 1) Marine Operations and Coordination. By packaging Marine Coordination, management and equipment procurement under one umbrella, GMS will work with clients to plan, run and continually develop their Marine Co-ordination centre and procedures." *(TRUNCATED DUE TO LENGTH)*	Consumer	3
Motions Charters	"Motion Charters is a family run business based in Hamble near Southampton. We offer a variety of luxury cruising boats, powerboats and race yachts which are all well maintained and fully equipped for your trip, whether you're enjoying a spot of gentle cruising or competing in a sailing event. We pride ourselves on friendly customer service and offer 24/7 support to ensure you have an enjoyable time on the water. Call us for more details and we'll find the best package to suit you and your guests."	Consumer	4
Datum Electronics Limited	"Datum Electronics is a world-leading supplier of marine shaft power meters. Our unique fully modular systems are capable of measuring the on-shaft torque and power of a ship on shafts from 150mm to 1,100mm (and above) diameters. Shaft Power and Torsionmeters, systems suitable for ship trials or permanent installation into ships." *(TRUNCATED DUE TO LENGTH)*	Unknown	5

3.11 Comparison with other approaches

There are conceptual similarities between TSR and other approaches that broaden known or learned relationships based on item content, such as by pairing an unlabelled item with the most semantically similar labelled one (Yuan et al., 2016), or using automated ontological classification to generalise a user's specific interests (Middleton et al., 2004). These approaches offer similar advantages to TSR in providing explainable results and supporting some cold start scenarios. However, these works do not address double cold starts (where no labels are known for both the user and item).

Additionally, the approach taken by Yuan et al., (2016) considers only the relationships of the most similar item, whereas the alternative scoring algorithms for TSR (section 3.9) consider multiple similar items in a weighted fashion. This approach was demonstrated to significantly improve results on the IWSC dataset and may also on other datasets, especially where labels are highly sparse. TSR does introduce additional complexity by considering multiple routes, but compute and memory requirements were not found to be problematic for the IWSC dataset, and methods for effective filtering and optimisation are described in sections 3.6.2 and 3.7.4.

The ontological approach of Middleton et al., (2004) groups items based on content (into topics in the ontology) and these items are then considered to share applicability to the user, as opposed to the TSR approach of weighting the applicability for each item pair. Ontological grouping (or alternatively a clustering based approach) seems sensible for domains with discrete topics, such as for a research paper recommender system like in Middleton et al., (2004). However, for less separable domains, the distance weighted approach of TSR may better capture relationships for items that exist on the edge of multiple categories, do not neatly fit any, or could fit multiple. Such cases might also benefit from an approach where items can be placed in multiple overlapping categories, for example using LDA (Blei et al., 2003). A comparison of these approaches with TSR may give interesting insight into the benefits of categorical versus distance weighted application of relationships, but this is left to future research.

The distance weighted application of relationships used in TSR could potentially be applied outside of recommender systems, for example to learned association rules. To determine what rules to apply to a new item, existing rules could have their rule confidence multiplied by TSR confidence (which is always ≤ 1), such that rules for more distant items are given less confidence.

3.12 Conclusions

TSR has been demonstrated as an effective solution to the challenging problem of cold-start recommendations in datasets with few labels, by making use of unstructured text descriptions of items. This addresses SRQ1: “How can machine understanding of text be used to identify relationships between documents in large collections of unstructured text?” as well as a general problem in the area of recommender systems.

This novel technique has the advantage of producing detailed provenance for the results, including the items and relationships considered and how they are weighted. Unlike some content-based recommender systems TSR is not dependent on similarity between query and target item content, unlike collaborative approaches, it does not require a history of interactions for either the query or target, and unlike ontology and rule-based approaches it does not require existing structured knowledge or many historic examples per item from which to derive rules and is not domain specific.

TSR can be used as a stand-alone recommender system or could be used to support other systems when dealing with cold-starts, without the need for bootstrapping. TSR is particularly suited for applications in high-velocity big data or similar environments where items and relationships may be time-sensitive or for other reasons few relationships are known for each item and/or none are known for many items, but some historic examples of relationships are known. This could include domains such as supply chain, tenders, job postings, or consultancy.

TSR has already seen real-world adoption and use by industrial partner Launch International LTD. Additional details of this application are given in the impacts section of Chapter 5. The full IWSC dataset (section 3.4), TSR implementation, and evaluation toolkit (section 3.7) have been made publicly available for download as open-data/open-source software (Ralph et al., 2019). Parts of this chapter were published as a full paper at the IoTBDS 2019 conference and as a journal article in Springer Computing.

Chapter 4 **Finding Meaning in Survey Data**

4.1 Chapter Overview

This chapter looks at techniques for distilling and presenting large collections of unstructured, unlabelled, text to be more easily interpreted by humans. This particularly addresses SRQ2: “How can machine understanding of text be used to produce an interpretable overview of large collections of unstructured text” and SRQ3: “How can the results of text analysis be effectively presented and used to inform decision-makers, analysts, and organisations”.

A particular focus is given to free-text survey responses and follows a collaborative project with the Parliamentary Office of Science and Technology (POST) analysing a survey of experts concerns regarding the COVID-19 pandemic. Sections 4.2 and 4.3 provide more background on the problem; sections 4.4 to 4.6 follow the iterative steps taken in producing the analysis, and its findings; section 4.7 presents a generalisation of this approach, the Text Insights Pipeline (TIP), and makes comparison to other tools and methods of analysis; sections 4.8 and 4.9 look at applying the generalised approach to other datasets; and section 4.10 discusses other potential applications.

4.2 Introduction

Analysis of free-text responses to surveys by human analysts is a laborious manual process requiring reading potentially vast collections of text and appraising, categorising, or synthesising an overview of its content. In the case of thematic analysis, where key topics are identified in the responses to produce a categorisation scheme or coding (Joffe & Yardley, 2003), it is necessary for the analyst to comprehend both the prominence and the diversity of topics. Large datasets are expensive and time-consuming to analyse, requiring great intellectual labour in the reading and memorisation of responses in order to perform these processes (Braun & Clarke, 2006; Joffe & Yardley, 2003). Large datasets not only make the task more difficult but also risk introducing accidental omissions where less frequently mentioned or harder to define topics may be overlooked. To cope with a large volume of responses, it is therefore highly desirable to have an automated system to either perform or aid in the analysis or presentation of the data to analysts and decision-makers.

The COVID-19 pandemic had wide-reaching implications across many areas of society and experts from many fields have offered their concerns and advice in response to the crisis. The following sections present an analysis of the responses to the COVID-19 Expert Concerns survey conducted by the United Kingdom Parliamentary Office of Science and Technology (POST). Statistical, text-analytics, and visualisation techniques are applied to this new dataset to identify key areas of concern, overlapping areas of concern, and typical responses for each area of concern. The outputs were produced with the aim of assisting POST's analysts in interpreting the data for their own analysis and presenting the findings in a suitable format for distribution to policy makers.

4.3 COVID-19 Expert Concerns Survey Dataset

Between April 3rd and April 30th, 2020, the UK Parliamentary Office of Science and Technology (POST) conducted a survey of experts in "COVID-19 or its likely impacts" to elicit their concerns over a range of timeframes, with the aim of identifying the major areas of concern among experts and identifying any consensus, and to produce a synthesis to inform UK Parliament. Participants were selected by snowball sampling from UK universities and other research institutes by the POST Knowledge Exchange Unit. A complete list of experts and more detail on the survey methodology are available on the POST website (Parliamentary Office of Science and Technology, 2020a, 2020b, 2020c). The data analysed in this survey are the responses received prior to April 19th, 2020; additional survey responses may have been received outside this timeframe but are not included in this dataset.

The survey has a total of 4,096 responses from 1024 participants. Each response is unstructured English language text describing the participant's greatest concern(s) for one of four time-frames, which are (relative to when the survey was conducted): Immediate; Short Term (within 3 months); Medium Term (3-9 months); or Long Term (from 9 months onwards) and is associated with one of 22 categories selected by the respondent from a list previously decided by the survey authors (POST). Table 4.1 shows the full list of categories and for how many responses each category was chosen for each timeframe. For ease of comparison, this is also presented as a bar chart in Figure 4.1 and as a line graph in Figure 4.2.

It can be seen that there is a significant imbalance in the total number of responses for each category (range=633, mean=178, SD=164). With "Physical and Mental Health" being the largest by far (639 total) followed by "Virology, Immunology and Epidemiology of COVID-19" (430 total) and "Communities and Populations" (416 total). The prominence of each class also varies significantly between timeframes. In the immediate timeframe "Virology, Immunology and Epidemiology of COVID-19" has the most responses, but this falls off in later timeframes while other areas of concern such as "Economic and Financial Affairs" become more prominent in later timeframes. Section 4.4 examines the changing patterns in the distribution of responses between timeframes to gain insight into how the types of concerns in each area change over time in addition to the changes in number shown here.

Table 4.1 - Number of responses in each category for each timeframe, selected by the respondent

Category Name	**Immediate**	**Short Term**	**Medium Term**	**Long Term**	**Total**
Brexit	4	4	3	8	19
Business and Trade	23	30	52	38	143
Communities and Populations	105	119	96	96	416
Crime, Justice, and Policing	16	31	23	24	94
Culture, Arts, and Leisure	7	8	10	9	34
Devolution and Devolved Matters	0	1	1	4	6
Economic and Financial Affairs	50	72	114	98	334
Education and Training	50	63	73	69	255
Environment, Agriculture, and Food	25	30	32	33	120
Housing	8	9	11	11	39
I did not answer this question	51	38	63	86	238
Inequalities and vulnerabilities	94	89	68	73	324
Infrastructure and Energy	7	6	11	8	32
International Affairs and Foreign Policy	8	11	13	28	60
Manufacturing and Industry	19	15	14	17	65
Media and Communications	24	28	19	13	84
Parliament, Government and Constitution	21	15	12	36	84
Physical and Mental Health	167	183	153	136	639
Scientific Research/Methods and Emerging Technologies	78	67	64	88	297
Social Care	21	22	16	12	71
Transport	11	6	9	7	33
Virology, Immunology and Epidemiology of COVID-19	177	105	73	75	430
Work and Employment	58	72	94	55	279

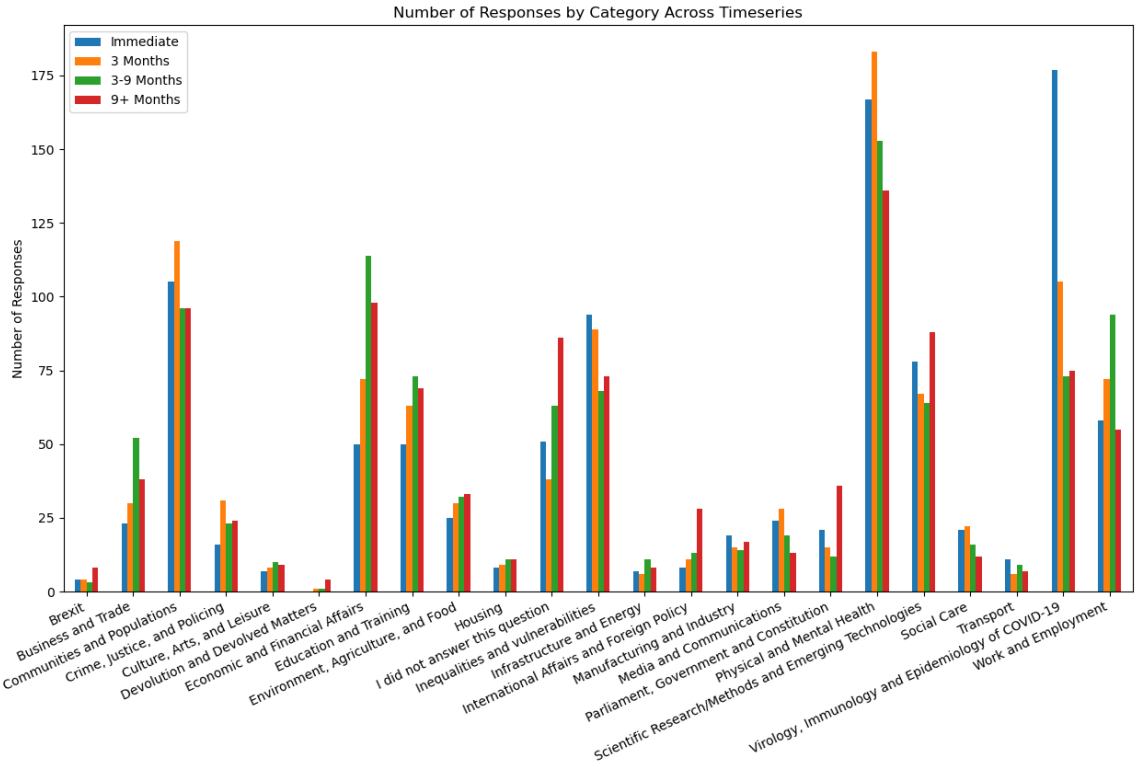


Figure 4.1 - Bar chart of responses in each category for each timeframe, selected by the respondent.

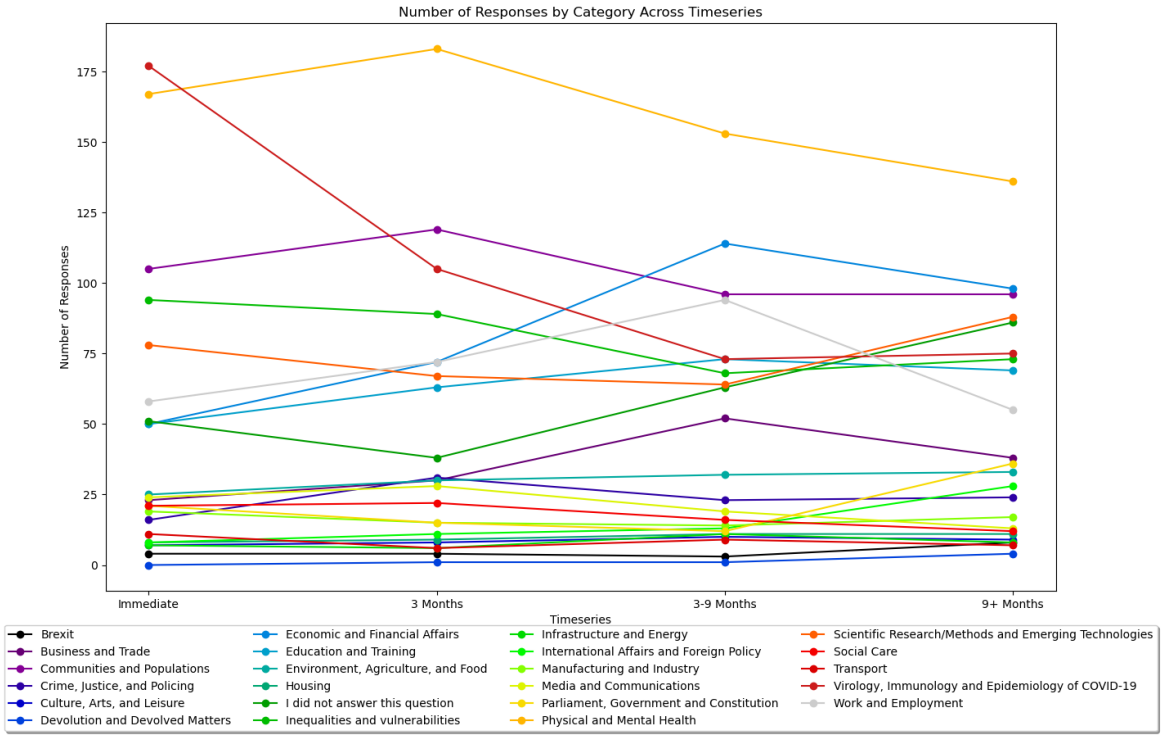


Figure 4.2 - Line graph of responses in each category for each timeframe, selected by the respondent.

4.4 Response Distribution

The high variance of the number of responses for each category across timeframes implies a changing focus of the distribution of the concerns. While statistics alone show the change in focus, the reason is unclear, a few likely causes are:

1) The importance of each category changing between timeframes
2) The types of concerns changing between timeframes
3) The category each concern falls into changing between timeframes

The influence of these and other factors can be examined by looking at the content of the responses. Another area of interest would be determining whether categories with more responses represent areas of greatest concern or are instead very broad and simply encompass a greater range of unrelated concerns.

This section presents a series of visualisations that show the distribution of the response content in semantic space, where responses with similar semantic meanings are plotted more closely. This enables examination of how the focus of concerns in each category changes, which categories have overlapping or similar concerns, how diverse the responses in each category are, and also provides a visual indication of how prominent each category is. Section 4.6 goes on to examine clusters in the data and proposes an alternative categorisation scheme based on automated clustering.

The content of the survey responses is in the form of natural language (English) text of varying lengths. The text is unstructured other than being split by respondent and timeframe. For the following experiments, Universal Sentence Encoder (USE) (Cer et al., 2018) is used to generate semantic text-embeddings for each response. These embeddings are fixed-length vector representations of the semantic features of the text in a shared semantic space (512 dimensions in the case of USE). Embeddings in this space can be compared using cosine similarity, where items with greater cosine similarity have more similar semantic meaning, and the inverse, cosine distance, given previously in Chapter 3 Equation 3.1.

To visualise the high-dimensional text embeddings, the dimensionality must be reduced, for this, these experiments use t-SNE (Maaten & Hinton, 2008) with pairwise cosine distance as the metric. t-SNE has been shown to be effective for producing visualisations of high dimensional data, including text embeddings (Kiros et al., 2015; J. Li & Jurafsky, 2015) (also see section 3.7.3). In contrast to some other dimensionality reduction methods such as Principle Component Analysis (PCA) (Hotelling, 1933), t-SNE relies less on preserving distances between widely separated points and so better captures local neighbourhoods of points (Maaten & Hinton, 2008),

this suits the purpose of visualising survey responses as semantically similar responses are grouped more effectively, which is of greater interest than the global position of responses in semantic space (i.e., knowing which responses are similar is more useful than knowing which are dissimilar). t-SNE provides a tuneable parameter, "perplexity", which allows the focus to be put on local or global features; for this dataset, a perplexity of 30 was found to produces a result where, upon qualitative inspection, tightly clustered items appear similar in non-trivial ways and a minimum of duplicate clusters exist. All visualisations in Figure 4.3 and Figure 4.4 use the same semantic space and t-SNE model (such that Figure 4.3 (A) is a composite of all subfigures in Figure 4.4 superimposed). For all t-SNE visualisations presented in this chapter, multiple random seeds were used to generate visualisations and the result most clearly showing the features of interest was chosen. The nature of the features does not generally change, only their arrangement.

Examination of Figure 4.3 (A) shows several interesting features. Firstly, some natural clusters are present in the survey responses, particularly around the areas of "Education and Training", and "Work and Employment", these clusters are mostly separate from the rest of the dataset, suggesting that these are well-defined categories that have a focused set of related concerns (in that they have lower semantic variance compared to other groupings of data). There are also dense regions of responses mostly belonging to the same category for the "Physical and Mental Health", "Media and Communications", "Crime, Justice and Policing", and "Economic and Financial Affairs" categories, which suggests there are a set of common concerns at the centre of each, which are common to responses for these categories but also some items of other categories, or that there are concerns in other categories that partially overlap with those of the primary category present in the cluster.

There is also a distinct region dominated by the overlap of the categories "Brexit" and "International Affairs and Foreign Policy", and many smaller regions where other categories strongly overlap, this suggests that these categories may share similar concerns. As the number of items for both the "Brexit" and "International Affairs and Foreign Policy" categories is small and they are shown to have strongly overlapping concerns, one interpretation is that the categories might be combined ("spliced") to create what could be considered a more powerful coding (Joffe & Yardley, 2003). Conversely, some highly distributed categories with many items, such as "Physical and Mental Health" and "Virology, Immunology and Epidemiology of COVID-19" may be better represented as multiple smaller categories split into sub-areas (e.g., separating Physical Health and Mental Health).

Also notable are several apparent clusters which do not have a clearly dominant category and/or feature items from many different categories. These suggest latent topics which are not captured

well by the categorisation scheme and cannot be seen by the prior statistical analysis in section 4.3. Figure 4.3 (B) shows a copy of Figure 4.3 (A) with several such clusters annotated with labels for the topic identified by a human (not affiliated with POST) by examining the text content of the responses within the cluster. Section 4.6 looks at alternative categorisation schemes produced through automated clustering, which provides a systematic approach to making similar types of observations and allows inspection of the data under a new lens.

Examination of Figure 4.4 gives an overview of how the prominence of different areas of concern changes between timeframes. As discussed in section 4.3, the number of concerns in each category changes significantly, but it can also be seen that their distribution changes. At the beginning of this section, three scenarios were proposed which may be responsible for these changes, which Figure 4.4 may provide support for. For each respective scenario, some expected observations would be as follows:

1) Changes in the number of items within the typical area covered by each category, representing its change in prominence.
2) Drift in the concentrations of points between timeframes, reflecting the change in focus of the content of the responses.
3) Change in the colour coding of each area between timeframes, showing how similar concerns are interpreted as belonging to different categories at different times.

Observing the figures shows some evidence of all three of these scenarios, suggesting that a combination of these factors explains the changes seen in the statistics in section 4.3. Some specific observations are given in Table 4.2.

Table 4.2 - Observations of changes in distribution and categorisation of responses between timeframes

Observation	**Implication**
Concerns for the category "Work and Employment" remain focused in the same region but vary in number across timeframes.	Change in quantity supports scenario 1.
Concerns for the category "Economic and Financial Affairs" remain focused in the same region but vary in number across timeframes.	Change in quantity supports scenario 1.
Concerns for the category "Virology, Immunology and Epidemiology of COVID-19" are most numerous and diverse in the immediate timeframe and become more focused in later timeframes.	Change in quantity supports scenario 1. Change in distribution supports scenario 2.
Concerns for the category "Media and Communications" are highly focused in early timeframes but become more diverse in later timeframes, while remaining similar in number.	Change distribution supports scenario 2.
The area identified as "Future Preparations" is far more densely populated in the long-term timeframe than any other.	Change in distribution supports scenario 2.
The area identified as "Exit plans for lockdown" is far more densely populated in the immediate timeframe than any other.	Change in distribution supports scenario 2.
Responses in the area identified as "NHS and healthcare funding and resilience" primarily belong to the "Virology, Immunology and Epidemiology of COVID-19" in earlier timeframes but change to "Physical and Mental Health" in later timeframes.	Change in category supports scenario 3.

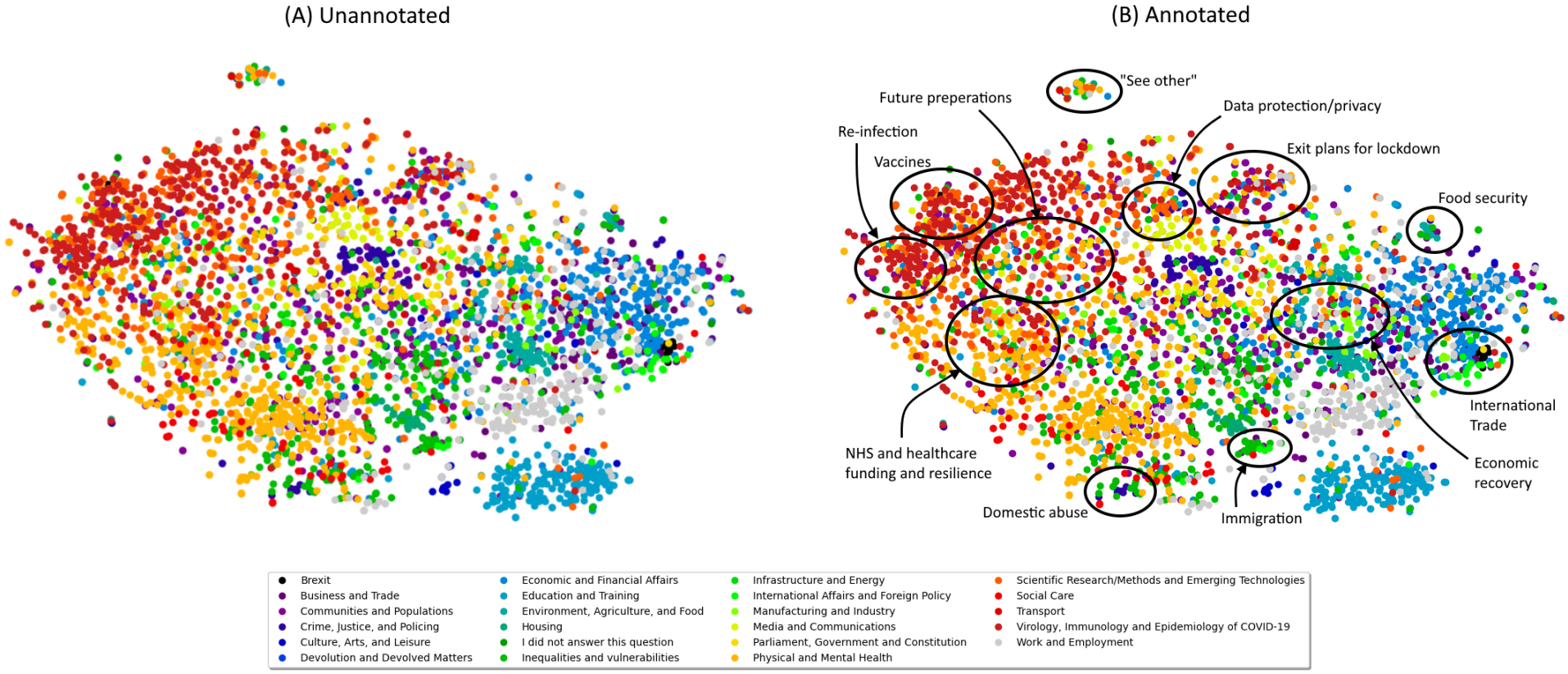


Figure 4.3 - 2D t-SNE plots of USE text embeddings for responses in the Expert Concerns dataset colour coded by the category selected by the respondent. The figure shows responses for all timeframes. The figure shows the prominence and distribution of each category and overlap between categories. Subfigure B is a copy of subfigure A but annotated with human observations of latent topics which form multi-category clusters

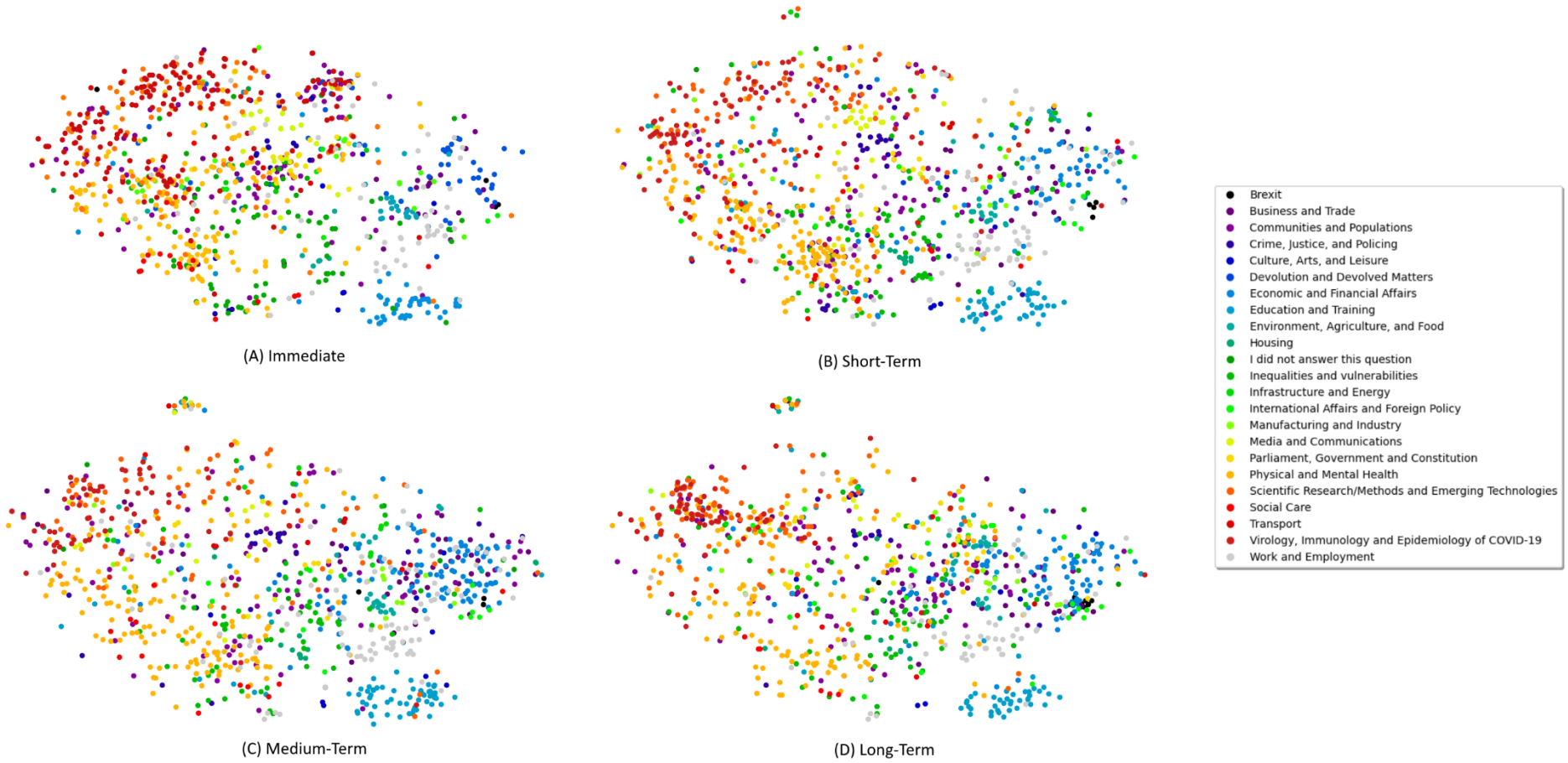


Figure 4.4 - 2D t-SNE plots of USE text embeddings for responses in the Expert Concerns dataset colour coded by the category selected by the respondent. Subfigures A, B, C, and D show the responses for each timeframe, Immediate, Short-Term, Medium-Term, and Long-Term concerns, respectively. The figure shows how the distribution of responses and the prominence of each category changes between timeframes

4.5 Response Summarisation

As each category contains many responses and a variety of different concerns, it is desirable to generate an overview of the key concerns for each category. For this, text summarisation can be used with the aim of capturing a diverse set of common talking points. For the result to be representative, it is important that the summary has both good coverage of the range of concerns (sensitivity) and prioritises the most frequently mentioned concerns (specificity), as well as being unbiased both in terms of the subject matter and the wording of the concerns. These summaries are analogous to the selected extracts for each theme in thematic analysis, although differ in that they are selected only to best represent the theme rather than to also address research questions specific to a study (Braun & Clarke, 2006).

There are two types of summarisation to consider; abstractive summarisation aims to generate new text which captures the key points of the source text, whereas extractive summarisation aims to select the most representative or important sentences from the source text. Much research has been done in both areas, with some significant recent advances being made using deep-learning and text embeddings. However, concerns have been raised over the biases in pre-trained models and abstractive models generally, as they can be prone to misattribution, factual inaccuracy, and repetition, and struggle with out-of-vocabulary terms (Caliskan et al., 2017; Nadeem et al., 2021; See et al., 2017). Further, deep-learning-based summarisation algorithms generally operate as “black-box” models which are difficult to interrogate for a meaningful explanation of why a particular result is given and the evidence it is derived from.

As the output of abstractive summarisation is novel text rather than strictly sampling from the input text, it is also more difficult to identify the provenance of each output statement, which may be desirable when the input text is comprised of many discrete items from different authors.

Conversely, extractive summarisation algorithms which are not pre-trained and quote directly from the source text may preserve existing biases within the source text but should not introduce new biases and mitigate the possibility of misattribution (when sampling complete sentences), although still risk presenting sentences out of context. For extractive summarisation, the parts of the input text sampled from can be easily identified as each segment is unmodified.

For the task of producing a representative but focused overview of the key concerns for each category, based on a large and diverse set of responses from many different authors, extractive summarisation is more desirable. This can be used to identify a small set of representative responses (or parts of responses) that are indicative of the large collection of source sentences.

For this aim, sampling a series of statements is more important than relating the ideas together in a flowing narrative, and provenance of the origin of each statement (the item it is sampled from) is also desirable.

Summaries were generated for each (category, timeframe) pair (e.g., Long-term concerns for Work and Employment). As sentence length and the number of source sentences varies significantly for each grouping, using a desired word count is most effective for producing summaries of consistent length. For the purpose of providing a concise overview of each grouping, a word count of 100 was chosen (this is a soft limit due to sampling complete sentences). The complete results cannot be included due to concerns of confidentiality and personal information; however, a few representative output samples are given in Table 4.3.

A variation of the TextRank algorithm from (Barrios et al., 2015; Rehurek & Sojka, 2010) is used for summarisation of the concatenated text responses for each category. TextRank is an extractive summarisation algorithm that uses a graph-based model for selecting the most important sentences from a text corpus, similar to how early search engine and information retrieval algorithms such as PageRank (Brin & Page, 1998) select results.

This graph-based approach has the advantage of being invariant to the language and terminology used as it is entirely unsupervised and does not rely on a defined vocabulary (Mihalcea, 2004; Mihalcea & Tarau, 2004); this allows the summarisation to recognise the significance of domain-specific terminology mentioned frequently in the source text whereas a model with a finite pre-learned vocabulary (such as text embedding models) may undervalue their significance due to being out-of-vocabulary. As the TextRank model is not trained on any corpus other than the text to be summarised, the model has no previous bias to the significance of topics discussed in the text and determines this exclusively using the text analysed (Mihalcea & Tarau, 2004).

Graph-based extractive summarisation also aligns with the aim of selecting the most representative sentences from the source text, as the graph is generated based on sentences with overlapping themes “recommending” each other so as to maximise coverage of the most prominent themes in the text graph (Mihalcea & Tarau, 2004). While this can produce summaries lacking narrative flow (as no attempt is made to select sentences that lead on to each other), it is not required for this task, which instead desires the most representative sentences regardless of how they relate to each other aside from avoiding redundancy (e.g., a bullet-point list of key concerns). A caveat of this approach is that the model does not account for synonyms and semantically similar terms, although lemmatization is employed (i.e., inflected forms of words are substituted with their dictionary form). The inability to recognise semantically similar words may result in some redundancy in the generated summary where different terminology is used as the

model is unable to recognise the equivalence. It is also possible that some topics will be under or over valued depending on how diverse and consistently used terminology is for that topic.

Table 4.3 shows some example output from the extractive TextRank summarisation. When considering the quality of the summaries, the key criteria are that the summaries:

- Consist of meaningful sentences not lacking essential context
- Are sufficiently sensitive to the range of concerns in that category
- Are sufficiently specific in capturing the key concerns of that category

Examining the results shows that the generated summaries generally perform well in satisfying these criteria. In particular, all the generated summaries are sensible and relevant, selecting a variety of issues most of which strongly relate to the category but also with little redundancy. Qualitative inspection of items in each category shows that the summaries are generally representative of typical responses. There is no obvious bias in the summaries and the selected sentences include a variety of factual statements, questions, and opinions. This indicates the approach is not overly discriminating based on the formatting and writing style of the responses.

In the examples shown in Table 4.3, summaries are included for the same categories over multiple timeframes. It is apparent from the content of the summaries how the focus of responses changes between timeframes for the same category, which further supports scenario 2 discussed in section 4.4.

Table 4.3 - Examples of extractive summaries for (category, timeframe) pairs generated using a variation of TextRank on the concatenated text of all responses for that category and timeframe. Examples have been chosen in accordance with confidentiality and anonymity requirements of the data while being representative of the typical output.

Category - Timeframe	**Generated Summary**
"Physical and Mental Health" – Short Term	"Both the direct and indirect impact on children's and young people's mental wellbeing through real health risks, negative media, self-isolation, lack of safety, and loss of protective factors such as schools and social activities." "I am worried about the impact of Covid-19 on the mental wellbeing of healthcare workers in the NHS, and of working people more generally, how are they coping with their health worries, their finances job security and working remotely from home" "I am concerned about the impact that social distancing, self-isolation and lockdown measures will have on people's wellbeing."
"Physical and Mental Health" – Long Term	"The potential effects on physical and mental health (e.g. sedentary behaviour and chronic disease, obesity, anxiety, depression) due to social distancing measures (i.e. restricting movements and access to facilities, key services and education) and wider anxieties about the pandemic, particularly for the most vulnerable/shielded groups (e.g., diabetes, arthritis, immunosuppressed)." "Longer-term wellbeing of patients/families - burdens on individual as they potentially lose job/independence, find it impossible to access appropriate neuropsychological help as likely to fall between physical/mental health services.", "People with diabetes are at increased risk of long-term adverse physical and mental health outcomes as a result of the COVID epidemic."
"Education and Training" - Short Term	"My main concern over the next three months is whether universities have the systems in place, and are able to mobilise rapidly enough, to remotely support the mental health of current university staff and students so that the academic year can be successfully brought to a close." "How will Government ensure that schools are advised and supported to use technology effectively, providing technology resources so that learners at pivotal points in their education pathway, specifically those in Y10, Y11, Y12 and Y13 will be adequately equipped for the continuation of their studies later in the year?"
"Education and Training" - Long Term	"How to minimise the impact of the outbreak on the education and training of students and trainee teachers whose learning and professional development has been disrupted during the academic year 2019-2020." "Future workforce provision within healthcare - without careful, considered & proactive management, whilst we may meet the immediate challenge, the requirement for clinical experience to ensure protection of the public could lead to significant challenge providing this experience for all health & social care students." "How will Government ensure that the schools system is digitally resilient, able to utilise technology effectively to reduce disadvantage and quickly switch into remote learning to deal with future crisis of a similar scale."

4.6 Alternate Categorisation

As discussed in section 4.3, the categorisation scheme chosen for the survey results in high variance of the number of items in each category, and the analysis in section 4.4 shows that some larger categories have a very diverse range of concerns, whereas other more focused groupings of concerns are not well covered by any category. This section demonstrates an alternative categorisation system based on clustering of response texts, with the aim of better capturing the key areas of concern described in the data.

This section presents the results for categorisation of long-term concerns as these were of most interest in the collaboration with POST, who then contributed human naming for the generated categories (see Table 4.6) and used them as the basis for their Areas of Research Interest (ARIs) for COVID-19 (see section 5.3.2).

4.6.1 Categorisation by Clustering

As the data contains clear groupings when the text responses are visualised, as in section 4.4, it follows that automated clustering may be an effective method of categorising the data. To achieve this, K-Means clustering (Lloyd, 1982; Pedregosa et al., 2011; Sculley, 2010) is applied to the semantic text embeddings (Cer et al., 2018) of the responses.

Qualitive criteria for judging a categorisation scheme might be that categories should have internal homogeneity and external heterogeneity (Patton, 2015), such that items within a category are similar but categories are dissimilar to each other. Using distinct clusters of items in sematic space as categories seems likely to satisfy these criteria if the clustering is meaningful.

For this dataset, a number of categories between 10 and 20 were found to produce the most intuitively meaningful results, where qualitative inspection of the items within each cluster shows each cluster to have a well-defined focus that is neither overly specific nor broad. For each number of categories, experiments were repeated with different random seeds to confirm that the results were consistent. The ideal number of clusters may depend on the desired usage, for example, to identify more specific issues within common topics. Figure 4.6 presents a comparison of the results for long-term concerns under the original human categorisation scheme and as categorised by automated clustering with 20 categories.

Several metrics exist for the quantitative evaluation of clustering algorithms; however, many require ground truth labels. While a complete set of labels does exist for the dataset in the form of the human selected categories, they do not align with the generated categories (as the approach taken produces a new categorisation scheme as opposed to fitting items to the existing

Table 4.4 - Comparison of effectiveness of clustering methods and human categorisation

Method	**Number of Clusters**	**Mean Silhouette Coefficient (Higher is better)**	**Variance Ratio Criterion (Higher is better)**	**Davies Bouldin Score (Lower is better)**
Human Categorisation	23	-0.08067	7.242	5.000
USE K-Means	23	**0.06522**	**18.310**	**3.306**
USE K-Medoids	23	0.02050	11.874	3.924
USE K-Means	20	**0.07708**	**20.177**	**3.226**
USE K-Medoids	20	0.00849	12.317	4.179

categories). For this reason, it is necessary to use unsupervised evaluation metrics. Table 4.4 compares the performance of the clustering approaches investigated and the scores for the original human categorisation. In addition to K-Means, the same methodology was tested but using K-Medoids (Park & Jun, 2009), which is similar to K-Means but seeks to minimise the distance between items and a real item at the cluster's centre (the medoid), instead of a virtual point which is the average of the cluster (the centroid) as in K-Means. Basing distances on the medoid instead of the centroid means K-Medoids is generally more robust to outliers with extreme values (Park & Jun, 2009), however, as extreme values do not exist in the normalised space of the text embeddings used this provides little improvement. When comparing algorithmic performance to human labelling it should be noted that these metrics measure a distance-based criterion similar to what these models aim to optimise, which is not the case for human labelling, so these scores are indicative of the nature of the results (particularly, how well defined the clusters are) but do not alone prove their validity.

Figure 4.6 shows a comparison of the number of items assigned to each category for human labelling and automated categorisation. It can be seen that in the human categorisation scheme, some categories have far more responses than others (range=138, mean=46.2, SD=40.0), whereas the categories generated through automated clustering of the responses have much more even applicability (range=62, mean=50.8, SD=15.1), except for two categories which have significantly fewer responses.

A more even category distribution is desirable as oversaturated categories are likely too broad in scope or are vaguely defined, and rare categories may be too specific or otherwise not represent a common area of concern. In both cases, the descriptive power of the categories is diminished.

Therefore, for a given number of categories, it would be advantageous to eliminate very small categories and subdivide the very large ones, so as to maximise the descriptive power of the categories and the coherence of the resulting analysis (Joffe & Yardley, 2003).

The generated categories should be expected to have a more even distribution as they are fit to the data, whereas the list of categories for human labelling was decided before conducting the survey. As K-Means clustering aims to find cluster centroids that keep clusters as small as possible (minimising the within-cluster sum-of-squares) (Arthur & Vassilvitskii, 2007), it would be expected to see similarly sized clusters for data with a continuous distribution and uniform density. Variation in the number of responses per cluster can then be explained by variations in the density of their positions in the embedding space, that is, clusters with more responses indicate that the responses are more similar, whereas clusters with fewer responses are more diverse, or otherwise occupy less dense regions of the embedding space (i.e., they have few similar responses).

In the results, the emergence of two clusters that have significantly fewer items (see Figure 4.6 (B)) may be the result of them containing outliers and highly unusual responses. Examination of the responses assigned to these clusters supports this, as these responses are mostly statements lacking context, such as “what is the new normal?” and “R&D with regard to COVID-19”. The semantics of these statements requires context that would be absent in a general language model as it pertains to a specific current event (the COVID-19 pandemic). It can be seen in Figure 4.5 (B) that the responses in these categories are not tightly grouped and are on the edges of other clusters, supporting the hypothesis that the model considers them outliers.

4.6.2 Cluster Summarisation

To investigate the results of the algorithmic categorisation scheme produced, text summarisation can be applied following the same methodology as used for the original human categorisation scheme in section 4.5, Table 4.5 shows some sample results.

As with the summaries of the human selected categories, the summaries for the generated clusters are generally coherent and focused. Inspection of the items in each cluster confirms that the selected sentences are generally representative of common concerns of the responses in that cluster. It is notable that the cluster summaries contain some responses describing more specific concerns than are present in the summaries of human categories. This demonstrates how the clusters capture a more focused (i.e., better fit) set of responses than the often either very broad or niche categories in the human categorisation (as is also indicated by the variance in the number of responses in each). This reflects how the clustering captures specific areas of concern

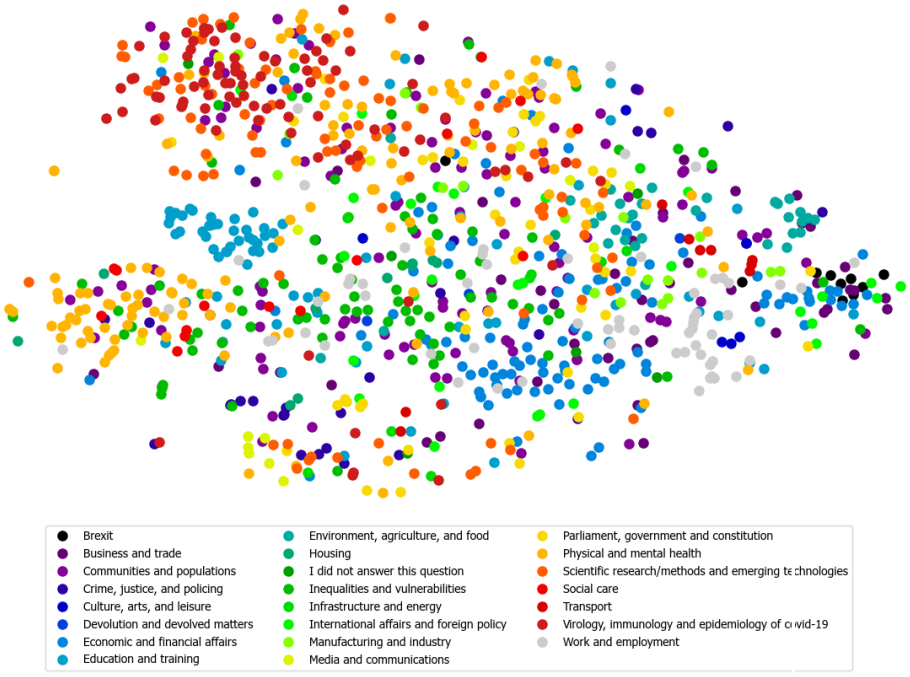
prominent in the data (e.g., “Social, economic and health inequalities”) rather than the mix of broad (e.g., "Physical and Mental Health") and specific (e.g., “Brexit”) themes which make up the human categories, which result in less focused summaries as the responses within each are often less focused or very few in number.

The example summaries include the cluster with the smallest number of responses “Resilience of society to future shocks”. In the previous section, it was hypothesised that the clusters with significantly fewer items may contain responses the model considers outliers. By looking at the generated summary it can be seen that while there is still an apparent general theme, the responses are highly diverse and generally more lacking in context than are seen in the other clusters. As the summarisation aims to select the key themes, this indicates that the key themes of these categories are less clear to the model, although a reasonable result is still achieved.

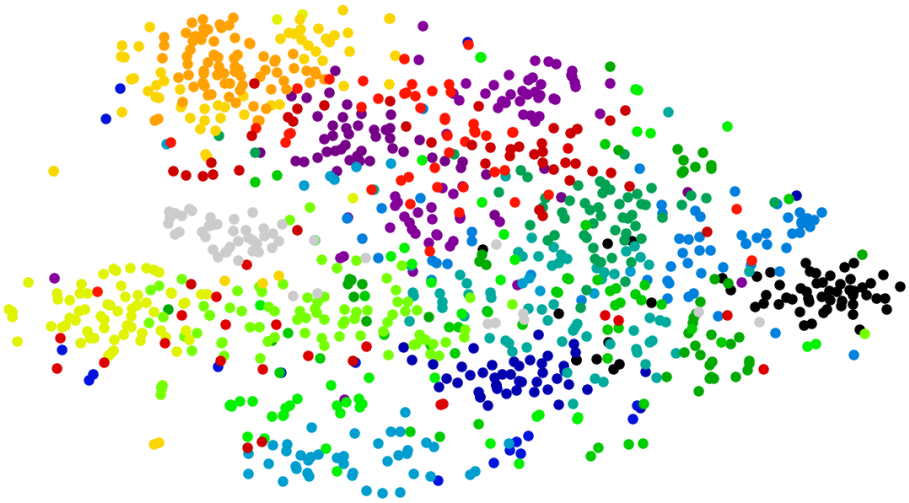
Table 4.5 - Examples of extractive summaries for clusters. Summaries are generated using a variation of TextRank on the concatenated text of all responses for that cluster. Examples have been chosen in accordance with confidentiality and anonymity requirements of the dataset while being representative of the typical output. Spelling and grammar are presented verbatim to show any possible effects on the algorithm.

Cluster	**Generated Summary**
“Social, economic and health inequalities”	"Increased time spent confined to homes will reinforce a cycle of poverty caused by inhabiting poor quality housing tipping the most vulnerable into long term ill health, greater poverty and having a knock on effect on adult employability and children's educational attainment." "In the longer-term, with shrinking economy, job-market and possibly support-services, the outbreak's negative impact will disproportionately fall on the more disadvantaged social-groups, including minority-ethnic people, older people, and those in precarious and low-paid employment." "Impact of Covid-19 on mental health and well-being of the population, including long term socioeconomic inequalities likely to arise from the economic impact of Covid-19 which will affect poorest communities hardest"
“Lessons learned from the COVID-19 outbreak”	"We must ensure that pandemic preparedness plans are in place to prevent such an economic impact of future pandemics." "What lessons can we learn from the current outbreak regarding the control of future epidemics/ pandemic?", "How are you going to ensure that pandemic outbreak planning and preparedness is enforced and effective, and not ignored and scaled back, as happened in the last years, leading to the current disaster?" "Preventing a second outbreak and preparing measures for a potentially different pandemic in the future" "What have we learned from Covid-19 modelling and strategy decision making for future preparedness planning?"
“Strategy for vaccine development, production and distribution”	"Development and deployment of a vaccine to general population; research to know if the length of immunity; how to prepare ahead of the new COVID-20 such that vaccines will be develop faster; Monitoring communities for cases of infection, using long-term-use fever screening at key locations." "Given high mutation rate of Covid19, long-term vaccine may not be possible; ongoing vaccine development will be needed." "Long term cardio-respiratory consequences and morbidities of former COVID-19 patients, based on 'lessons learnt' implementation of a streamlined research framework for any future pandemics and candidate vaccine manufacturing for future virus cycles / occurrences in the UK and across the globe."

Cluster	Generated Summary
“Changes to the role of education and the future of learning”	"Many years we support blended learning approaches in the UK Universities, but the level of technology integration is mainly related to members of staff workload and skills" "The role of education in general and universities in particular to play a role in the economic recovery of the country - ensuring better skills and vital social inclusion measures." "How to minimise the impact of the outbreak on the education and training of students and trainee teachers whose learning and professional development has been disrupted during the academic year 2019-2020."
“Resilience of society to future shocks”	“R&D with regard to COVID-19” “what is the new normal?” “Maintaining the system 'upgrade' and avoiding slipping back into 'business-as-usual'” “Unable to react efficiently and quickly to similar scenarios in future” “To help business recover from COVID-19 and prepare for the next negative shock.” “Again, resilience, recovery and coalescing around the possibilty of new norms of working patterns” “A full proof pro-active, rather than reactive plan, in place for similar event in the future.” “The readication of COVID-19 will leave scars on the history of man” “Can things get back to the old normal?”

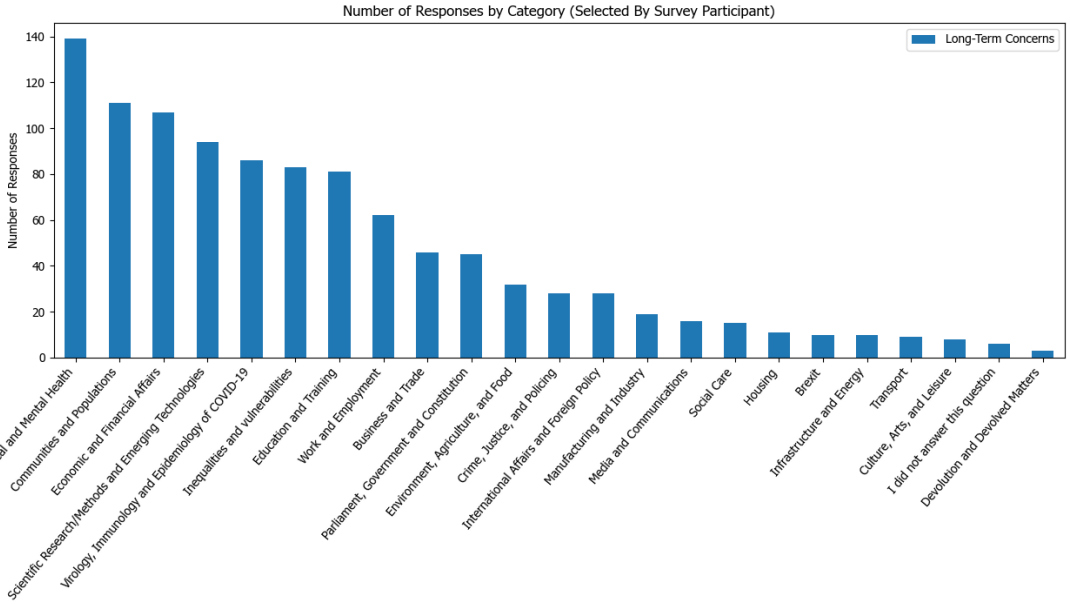


(A) Human Categorisation

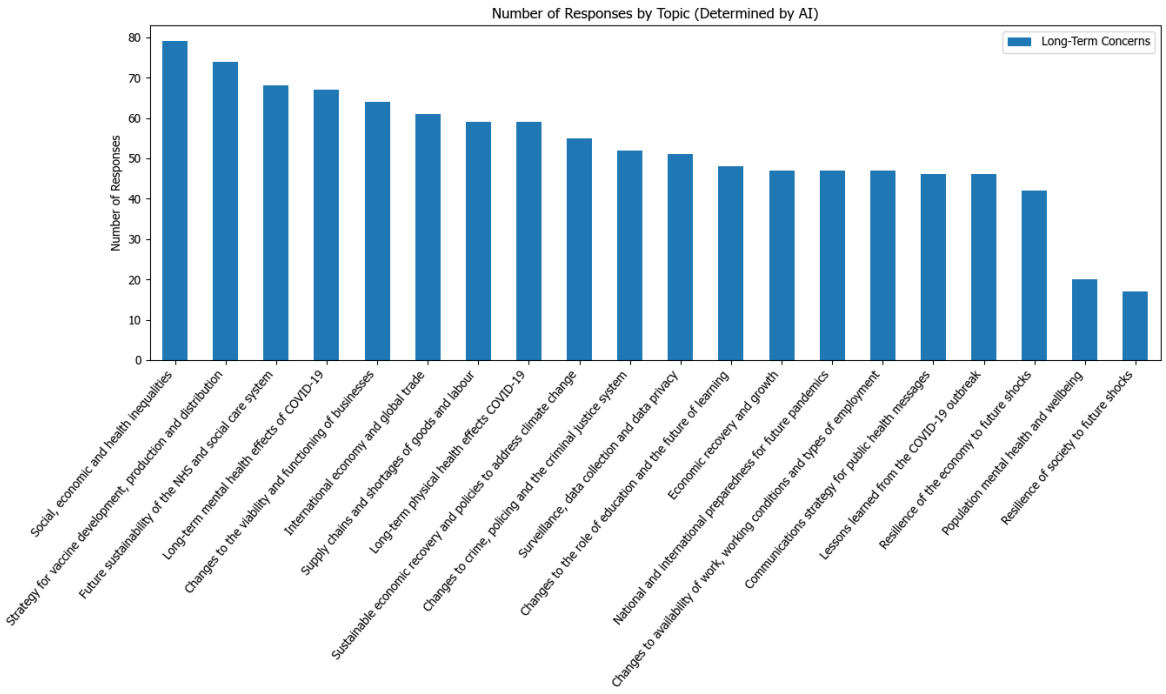


(B) Algorithmic Categorisation

Figure 4.5 - 2D t-SNE plots of USE text embeddings for responses for long-term concerns in the Expert Concerns dataset, colour coded by the category. Subfigure A shows the category selected by the respondent. Subfigure B shows categories generated by clustering.



(A) Respondent Selected Categories



(B) Automated Categorisation

Figure 4.6 - Bar charts of the number of responses in each category for the long-term timeframe. Subfigure A shows categories selected by the respondents. Subfigure B shows the results of automated categorisation

4.6.3 Topic Name Synthesis

This section has so far demonstrated an effective automated clustering approach for categorising the responses with the aim of producing a categorisation scheme that better reflects the areas of concern discussed in the response text content. The previous sections show the distribution of the new categories and look at summaries generated for those categories. While this information combined with an inspection of the results may be sufficient for a human analyst to assign category names, it is also possible to suggest category names based on machine analysis of the text content of each category's responses. This section shows the results for two alternative approaches for generating category names through keyword analysis, these are then compared to human-produced category names. All results discussed in this section, which concern naming the categories generated in section 4.6.1 for long-term concerns, are presented in Table 4.6.

The first approach investigated is applying TextRank (Barrios et al., 2015) for keyword extraction. This approach follows a similar methodology as for the generation of summaries, where the text analysed for each category is the concatenation of all of that category's responses. For generating keywords, TextRank uses co-occurrence of terms as its metric for "recommending" the terms most important in connecting together the content. The merits of this approach are similar to those for extractive summarisation, in that the graph-based model should select a collection of terms that provide good coverage of the web of terms as they occur in the source text, without bias to any prior training (Mihalcea & Tarau, 2004). It is notable, however, that as TextRank is not trainable and produces results exclusively based on the sample this approach does not consider the global distribution of terms over the entire dataset when used to generate keywords for each category. This means it is likely to produce similar keywords for each category if their content is similar as opposed to highlighting the differences between each one. For example, for the dataset in question, which is concerned with the impacts of COVID-19, many categories will likely be given the keyword "COVID", whereas a human would likely not include that term for each category as it is implied by the context of the dataset.

The next approach applies TF-IDF (Sparck Jones, 1972). A model is trained on the concatenated text of all responses in the dataset, then keywords are generated for each category using the concatenated text from only that category. By training on the full dataset, the TF-IDF model can consider the relative frequency of terms across the whole dataset (Term Frequency) as compared to within each category (Document Frequency). As a result, the keywords generated better reflect what makes each unique, as the model gives less weight to terms that are similarly frequent across all categories. As TF-IDF is a bag-of-words type model in which word order is unimportant, it is simple to inject additional terms into each document representing bigrams and multi-word

phrases which occur frequently in the dataset (e.g., if the word “public” is often followed by the word “health”, the n-gram “public_health” can be added to the list of terms for each document containing that pattern. It seems intuitive that including these common n-grams should produce more meaningful category names as these are commonly used in human-made category names.

The results of the keyword analysis using both techniques are shown in Table 4.6, as are the final human labelled names for each of the categories (which are the clusters from section 4.6.1). The human naming was contributed by an expert human analyst from POST (the survey authors) and their naming was based on a combination of their own analysis of the survey results (see the following section on their Topical Analysis), and the generated summarises (see section 4.5), TextRank keywords, and TF-IDF keywords generated for each category. It can be seen from the results that in most cases, the human analyst uses many terms matching or synonymous with the generated keywords, especially the TF-IDF keywords and multi-word n-grams.

While the adoption of many of the generated keywords by the human analyst in producing the final names for each category does evidence their usefulness, it remains that the generated keywords alone do not provide an easily readable category name without rewording by a human. This is a fundamental limitation of the keyword extraction approach as no concern is given to ordering the keywords for readability. However, abstractive summarisation has been shown to be effective in producing readable, human-like, titles for documents such as news articles (Chopra et al., 2016; Takase et al., 2016). While the text to be summarised for each category is much longer than is typical in news article summarisation, and the structure is very different (being a concatenation of individual texts rather than a single article with ordered paragraphs), this approach seems promising as a potential method for producing more readable category names as part of future research.

4.6.4 Topical Analysis

The human categorisation scheme that the results have so far been compared to was, as described in section 4.3, schematically defined before the survey was conducted and so is blind to the results of the survey. Independent of the computational analysis conducted in this work, POST also conducted their own topical analysis, the results of which are published on their website (Parliamentary Office of Science and Technology, 2020b). Like the analysis presented here, their study also looks at the results of the COVID-19 Expert Concerns Survey, however, the set of responses which they analyse does not exactly match those examined here as they select responses from a different range of dates, although there is significant overlap with the data used

here such that the prominent themes should be similar, but with minor quantitative differences. They also choose to produce a different number of categories.

Figure 4.7 shows a visualisation of the categorisation scheme they produced overlayed on items present in both versions of the dataset (the intersection). Comparison of Figure 4.7 and Figure 4.5 (B) shows that many of the same clusters are identified as distinct topics, and the naming assigned in each case is similar. Notably the human identified topics “Research and Innovation”, “International Affairs”, “Economy and Finance”, “Environment”, “Education”, “Society and Community”, and “Health and Social Care” all directly correspond to clusters produced by the computational analysis, being very similar in both distribution and assigned name. As this human categorisation was performed independently of the automated analysis, this strong consensus evidences the efficacy of the automated approach in terms of producing meaningful categories in a way similar to an expert human analyst.

The only contribution made by the author of this thesis to POST’s topical analysis was the “typical responses” presented for each category, which were produced through extractive summarisation following the same methodology as for the original categorisation (section 4.5) and the automated clustering (section 4.6.2). Like with the visualisation in Figure 4.7, this was done after the authors of the topical analysis produced their categorisation scheme so does not bias the comparison of their categorisation scheme with the one presented here. The results from sections 4.4 and 4.5, which do not concern alternative categorisation, were made available to them to aid in their preliminary analysis.

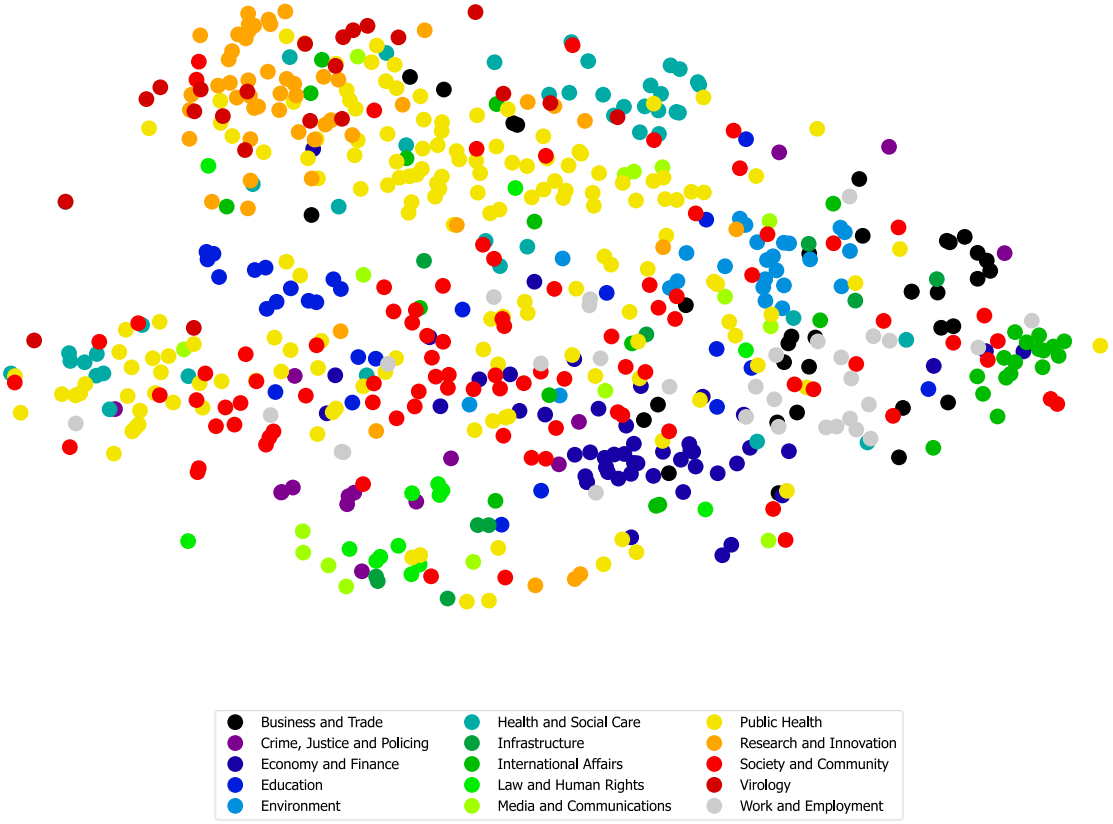


Figure 4.7 - A 2D t-SNE plot of USE text embeddings for responses in the Expert Concerns dataset colour coded by the category assigned in the topical analysis. Only responses present in both the subset of the data used by the topical analysis and the subset presented in this work are shown.

Table 4.6 - Clusters identified by K-Means clustering of USE text embeddings for responses for long-term concerns in the Expert Concerns dataset. Keywords were generated using Text-Rank and TF-IDF as provisional topic names which were then presented along with a synthesis of responses to a human analyst who contributed the human naming scheme

ID	**Human Naming (Contributed by POST)**	**Keywords (Text-Rank)**	**Keywords (TF-IDF)**
1	Changes to availability of work, working conditions and types of employment	working, employment, economic, needs, worker	work, employment, working_home, key_worker, job
2	Changes to crime, policing and the criminal justice system	news, covid, communicate, police, justice	criminal_justice, police, community, policing, crime
3	Changes to the role of education and the future of learning	teacher, educators, student, university, impact	education, university, school, student, teacher
4	Changes to the viability and functioning of businesses	economic, terms, needed, businesses, jobs	economy, business, job, recovery, lost
5	Communications strategy for public health messages	researches, governments, public, future, people	research, public_health, future, strategy, risk
6	Economic recovery and growth	economical, economies, social, recovery, maintaining	economy, recovery, monetary, financial, impact
7	Future sustainability of the NHS and social care system	health, governments, ensured, nhs, careful	nhs, health, social_care, staff, healthcare
8	International economy and global trade	economic, globally, governing, terms, economies	economy, brexit, international, global, trade
9	Lessons learned from the COVID-19 outbreak	pandemics, future, plan, better, preparedness	future_pandemic, pandemic, outbreak, pandemic_preparedness, lesson_learned
10	Long-term mental health effects of COVID-19	terms, health, socially, people, communication	mental_health, support, family, longer_term, anxiety
11	Long-term physical health effects COVID-19	disease, terms, covid, virus, long	virus, disease, infectious_disease, outbreak, long_term

ID	Human Naming (Contributed by POST)	Keywords (Text-Rank)	Keywords (TF-IDF)
12	National and international preparedness for future pandemics	health, covid, future, pandemics, include	pandemic, outbreak, future, health, future_outbreak
13	Population mental health and wellbeing	people, life, educations, uncertainties	education, lesson, routine, mental_health, learn_lesson
14	Resilience of society to future shocks	normality, failures, business, protective, seasonal	avoiding, failure, normal, resilience, business
15	Resilience of the economy to future shocks	economics, social, managing, risk, terms	economy, risk, management, business, sustainability
16	Social, economic and health inequalities	socially, economically, health, inequality, impacting	inequality, social, economy, mental_health, health
17	Strategy for vaccine development, production and distribution	vaccines, research, covid, developing, like	vaccine, development, vaccination, immunity, research
18	Supply chains and shortages of goods and labour	locally, terms, supplying, new, sectors	food, supply_chain, longer_term, local, city
19	Surveillance, data collection and data privacy	digital, data, health, future, surveillance	data, digital, pandemic, surveillance, public_health
20	Sustainable economic recovery and policies to address climate change	future, climate, globally, covid, economic	climate_change, future, climate, economy, crisis

4.7 Text Insights Pipeline

This chapter has looked at applying several algorithms and models to the task of exploring the Expert Concerns survey dataset and in particular in how the outputs of each of these can be combined to provide greater insight into the nature and key themes of the data.

However, the approach taken is not limited to this particular survey, or necessarily even to survey data. As part of producing the analysis, a suite of software tools was created to automate the generation, combination, and presentation of these results. This tool, which is one of the contributions and potential key impacts of this project, is the Text Insights Pipeline (TIP): a web-based tool for automated analysis of collections of text.

The Text Insights Pipeline is a research tool for visualising and interpreting collections of unstructured text, such as survey responses, item descriptions, or short articles. The tool combines the techniques described in this chapter to group similar items, identify naturally occurring topics, generate names and key sentences for each topic, and visually present the items and their topic groups for inspection by an analyst. The tool provides a means for analysts, such as social scientists and policy advisers, to explore and navigate their data much more efficiently than the traditional approaches of inspecting items in random or arbitrary order or by use of constructed queries, which risk introducing bias or accidental omission (Joffe & Yardley, 2003).

The tool is provided as a website where analysts can upload their data. The analyst may use the topic groupings identified by the tool as a basis for their analysis; combining, splitting, or tweaking groups as necessary, or to identify potentially overlooked topics in other analyses. The tool requires little or no configuration to produce good results in most cases, but advanced options are provided for trained operators to refine their results. The tool can also visualise and summarise according to existing categories, if present in the input data, which can be used to identify overlap or separability of the existing categories, or simply to generate key sentences for each category. The outputs are presented as a report style webpage, as well as downloadable figures and CSV files containing the complete results. The results are intended to be easy to interpret and non-technical explanations of each output are included in the report, however, some training in understanding the visualisation and clustering methods may add value for analysts.

The architecture of the pipeline is shown in Figure 4.8.

The tool takes as input a collection of text items in CSV format (e.g., for survey data, this would be one response per line), and optionally any known categories. The pipeline then performs several analysis steps; embeddings are generated for each text item, the dimensionality is reduced to two dimensions, and the items are visualised as a scatterplot (section 4.4), if existing categories were

provided the items are labelled as such, otherwise automated clustering is used (section 4.6.1) and topic names are generated for each cluster (section 4.6.3). For each category/cluster, a summary is generated (section 4.6.2). The results are presented as a human-readable report, and machine-readable CSV and JSON files with cluster information, including summaries, item mappings, and provenance.

The configurable settings include the desired length of generated summaries, the desired number of keywords, the number of clusters to generate, the t-SNE perplexity parameter (see section 4.4), and the random seeds for clustering and visualisation. There are also options to split items into sentences (where each sentence becomes its own item), and to automatically filter out items with fewer than two words. At the time of writing the tool makes use of the same USE model (Cer et al., 2018) as the experiments in this chapter, but the tool is designed to work with other embedding models, which may be added as options in future.

At the time of writing the tool is undergoing evaluation by the Wessex Institute and James Lind Alliance for use in assisting their analysts with their work on Priority Setting Partnerships, which involves identifying themes in large collections of patient and practitioner survey responses. More detail of expected future collaboration and development of the tool based on feedback is given in Chapter 5.

4.7.1 Comparison to Thematic Analysis

Results from TIP (and the analysis of COVID-19 concerns presented previously) are analogous to human produced thematic analysis but differ in some significant ways.

Firstly, human created themes usually have a description, both for the benefit of the reader and to aid the analyst in assessing their choice of themes (Braun & Clarke, 2006). TIP does not produce descriptions for its categories, only representative summaries and keywords, so it is necessary for a human analyst to interpret these.

Additionally, the data-driven inductive approach used by TIP does not select themes or examples tailored to a study's research question(s) or preconceptions, it instead produces an overview of the data and interpretation is left to the analyst. This may combat bias towards preferred findings, which benefits impartiality but may hurt relevance. The language model may have its own biases and limitations, and these should be considered when selecting a model, diligence should also be exercised when interpreting, communicating, or applying the results of automated analysis.

Thematic analysis is a flexible methodology with variations beyond the scope of TIP, including: theoretical (as opposed to inductive), where data are coded according to an analyst's interest or

research questions; and latent (as opposed to semantic), where data are interpreted to examine underlying ideas, assumptions, and ideologies, rather than their surface meaning (Braun & Clarke, 2006).

Other approaches to thematic analysis may also place items within multiple themes. While TIP is able to automatically split items into sentences and categorise each individually, this differs from how a human may code an item such that it is assigned to multiple themes. Hierarchies of themes are also not directly addressed by TIP, but could be produced by an analyst by running TIP with different subsets of the data or with differing numbers of desired categories as demonstrated in section 4.8 and discussed in more detail in section 4.10.

4.7.2 Existing tools for Thematic Analysis

Other software solutions exist to aid in qualitive analysis, including content and thematic analysis. These tools may aid analysts with tasks such as: creating codes (possibly with some automation such as named entity recognition); applying codes to data through highlighting or tagging; managing and collaboration on collections of data and analyses; and data ingestion from heterogeneous sources such as websites and multi-media (Dupplaw et al., 2012; Jackson Kristi & Bazeley Pat, 2019).

TIP differs from these tools in that it provides an alternative instead of an aid to the coding and assignment of codes to themes approach used in thematic analysis or approaches like word counts used in content analysis. As previously discussed, an analyst may still perform an important active role in using and applying TIP, and the results resemble that of some forms of thematic analysis, but the underlying method differs.

Unlike existing tools, TIP is able fully automatically to produce an initial overview of the data, making it suitable for exploring or giving a broad overview of a large dataset without requiring significant investment of time and effort. While advanced configuration options are available, it requires no special training to produce good results, as demonstrated in the following sections looking at alternative datasets, which each use the tool's default parameters. The work invested by an analyst using TIP would be in fine-tunning and tailoring the results to their research questions, rather than producing, reviewing, and organising codings.

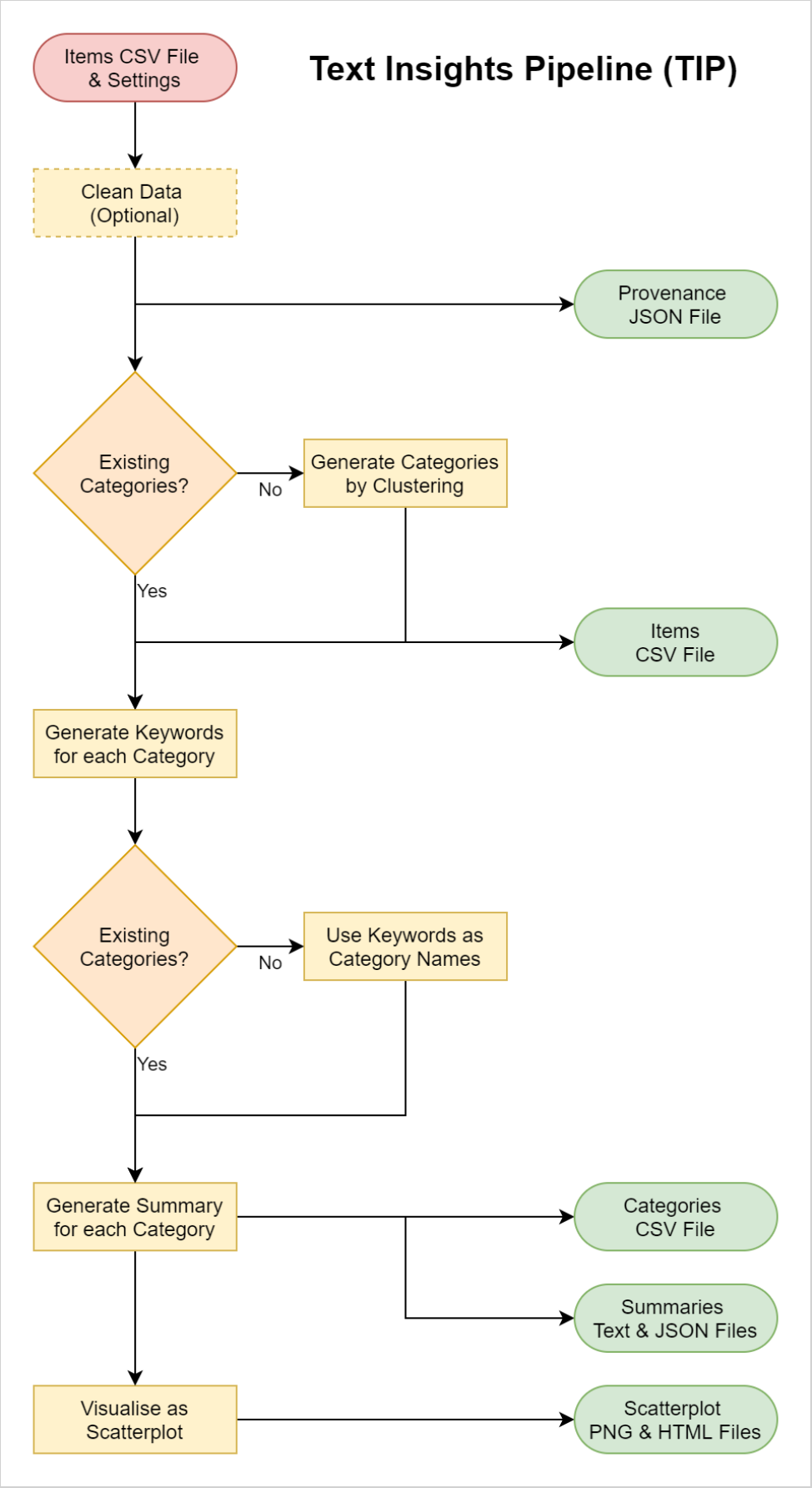


Figure 4.8 – High-level architecture of the Text Insights Pipeline (TIP)

4.8 Alternative Dataset: TED Conferences

The generality of the approach in this chapter can be easily demonstrated on other datasets by using the Text Insights Pipeline. Due to confidentiality requirements, only limited information can be published regarding the real-world projects described in the impacts section, so instead this section gives an example from an open dataset of TED Conferences (Banik, 2017).

TED Conferences (Technology, Entertainment, Design) are a series of short talks (18 minutes or less) by many different speakers across a variety of topics (TED Conferences LLC, n.d.). For this example, a dataset of the titles, descriptions, and tags of 2500 English language talks are used. The descriptions are one or more complete sentences (word count: range=121, mean=52, SD=18) including the topic of the talk and sometimes information about the event or speaker. The tone of the talk is sometimes described (e.g., “in this funny talk”), and the profession of the speaker is usually stated. The descriptions are typically written in a casual, attention-grabbing style, or phrased as a question. The tags are a comma-separated list of words and multi-word phrases reflecting the topic of the talk, sometimes including the names of places and organisations (number of tags: range=31, mean=7.5, SD=4.3).

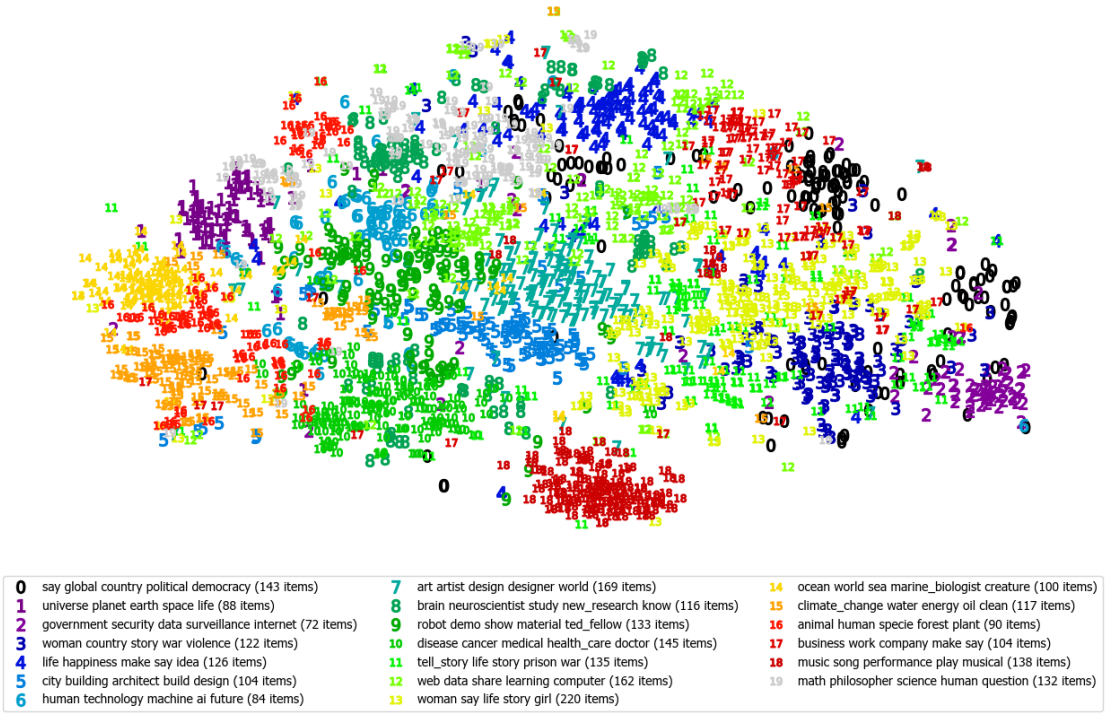
Figure 4.9 shows the results from the Text Insights Pipeline for this dataset, using only the talk descriptions (subfigure A) and tags (subfigure B). Figure 4.10 shows a comparison of results for 10, 20, and 30 generated categories. All other TIP parameters are their default values (the same as used for the Expert Concerns dataset). Interactive versions of these figures where individual items can be inspected, as well as the generated summaries, and provenance and output data files, are available online (Ralph, 2021). The interactive webpage which presents these results is the same as the output format of TIP (a generated HTML report) and is similar to the style of the report provided to the Lords Committee for the COVID-19 project, except for that report also includes a comparison with POST’s categorisation and multiple timeframes.

Whether using the talk descriptions or tags, some distinct clusters are clearly visible, and the generated topic names are a good indicator of their content. The clusters are much more separable when using tags, as would be expected from the discrete nature of the terms used (a list of tags should not generally contain words with little or unclear semantic value). Comparison of the category names shows some strong similarities, showing that this approach is good at selecting appropriate keywords and that the variable formatting of the text (sentences versus a list of tags) does not confuse the model.

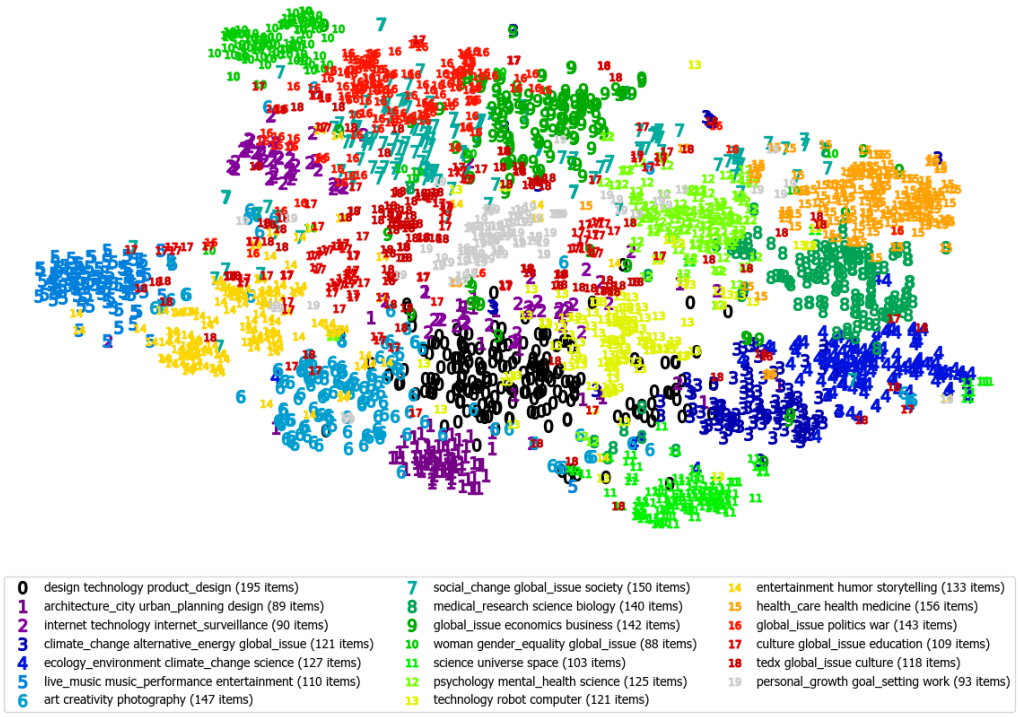
When varying the number of categories, we see that the smaller set of categories have broader topics whereas the larger set of categories are narrower and more focused. Some specific

examples are given in Table 4.7 and Table 4.8 of how increasing the number of categories allows for broad categories to be subdivided into meaningfully distinct but related sub-categories.

It is notable that the time taken to generate these results using the Text Insights Pipeline is approximately one minute on a workstation with no exceptional hardware, whereas an unassisted human analyst would require much longer to complete an analysis of so many items. A typical English language silent reading speed is often given as 300 words per minute, and possibly lower for non-fiction reading (Brysbaert, 2019). Assuming a human analyst was perfectly efficient in categorising items as they read them at a rate of 300 words per minute, for the TED dataset it would take a minimum of 63 minutes just to read the tags, or 440 minutes (7 hours and 20 minutes) just to read the descriptions. While the compute time for TIP does increase with the number of items, it remains orders of magnitude faster than a human analyst for the quantities of data tested.



(A) Descriptions



(B) Tags

Figure 4.9 – Visualisation and categorisation produced by the Text Insights Pipeline of talks in the TED dataset using their descriptions (subfigure A) and tags (subfigure B).

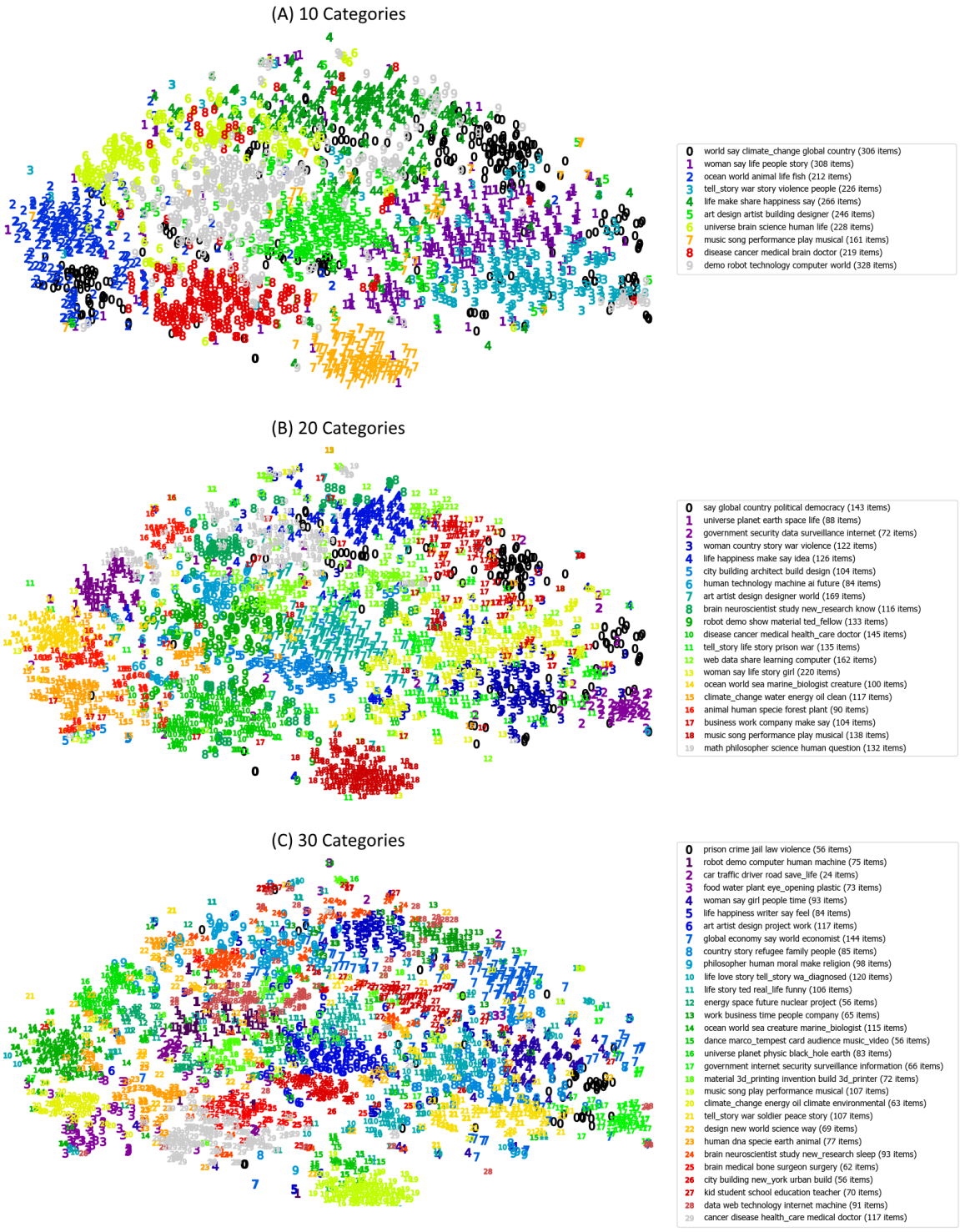


Figure 4.10 - Visualisation and categorisation produced by the Text Insights Pipeline of talks in the TED dataset using their descriptions for 10 categories (subfigure A), 20 categories (subfigure B), and 30 categories (subfigure C).

Table 4.7 – Categorisation of TED science talks for 10 (highlighted) and 30 clusters.

ID	Human Naming	Generated Name	Items
10 Categories Cluster 6	Science	universe brain science human life	228
30 Categories Cluster 16	Physics & Space Science	universe planet physic[s] black_hole earth	83
30 Categories Cluster 22	Scientific Methods	design new world science way	69
30 Categories Cluster 24	Neuroscience	brain neuroscientist study new_research sleep	93

Table 4.8 – Categorisation of TED technology talks for 10 (highlighted) and 30 clusters.

ID	Human Naming	Generated Name	Items
10 Categories Cluster 9	Technology	demo robot technology computer world	328
30 Categories Cluster 1	Robotics	Robot demo computer human machine	75
30 Categories Cluster 2	Automotive	Car traffic driver road save_life	24
30 Categories Cluster 12	Energy	Energy space future nuclear project	56
30 Categories Cluster 18	3d Printing	material 3d_printing invention build 3d_printer	72
30 Categories Cluster 28	Web & Internet	Data web technology internet machine	91

4.9 Alternative Dataset: Isle of Wight Supply Chain

Figure 4.11 shows the visualisation and categories produced by the Text Insights Pipeline for the IWSC dataset detailed in Chapter 3, section 3.4. The generated categories appear coherent and align with what would be expected as the major areas of economic activity on the Isle of Wight. In particular, the model identifies categories relating to marine engineering (clusters 0 and 1), marketing (2), tourism (3), marine survey (4), business services (6), boat/yacht charter (8), and the local service sector (9).

Cluster 5 contains mostly very short texts which may account for why they are more scattered due to each lacking information that might relate them to other items. Most of these relate to culture or art, as the generated category name suggests.

The theme of cluster 7 is not obvious by the generated name but inspection of the items shows that it is engineering-related, particularly concerning instrumentation and sensor systems. Many of these items have long descriptions and include a lot of product names and website links. The useful words “system” and “data” do appear in the generated category name, so it is possible the other less useful keywords are the result of a lack of common terminology between items resulting in the model instead preferring the common terminology found in hyperlinks.

These results show that the model is quite robust to highly varied and noisy data, but also shows some of the limitations, particularly in positioning items and generating category names when some items have very long or short text. Removing markup tags may improve category naming.

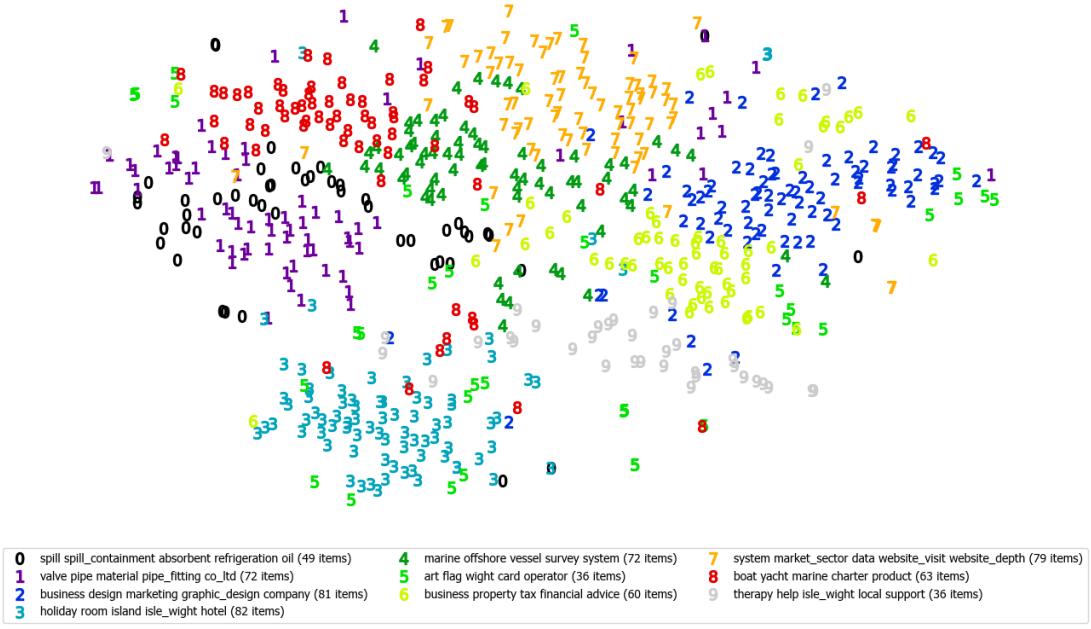


Figure 4.11 – Visualisation and categorisation produced by the Text Insights Pipeline of company descriptions in the IWSC dataset.

4.10 Discussion & Conclusions

The approach presented in this chapter has been shown to be an effective method for the automatic production of thematic analysis of a variety of text datasets from varied domains. Additionally, these are presented in a visual and interactive medium which enables human analysts to interpret and explore the data intuitively. This provides a solution to SRQ2: “How can machine understanding of text be used to produce an interpretable overview of large collections of unstructured text?” and a basis for answering SRQ3: “How can the results of text analysis be effectively presented and used to inform decision-makers, analysts, and organisations?”.

In regard to SRQ3, the analysis produced could be used as a basis for further human analysis. Having an interactive visualisation of the distribution of items, showing similar items close together and the clusters formed by common topics in the data, allows an analyst to inspect the data more efficiently than by inspecting items in random or arbitrary order.

This presents the opportunity for the approach to be used as part of a collaborative human and machine process where the automated analysis and its interactive outputs can expedite the work of human analysts by presenting a general categorisation that they can then refine. Using the Text Insights Pipeline, it is possible to provide input with pre-assigned categories to produce the same visualisations and summaries without performing clustering, which allows for analysts to modify the outputs from a first-pass (including clustering) by performing their own, splitting, splicing, and renaming of categories, and reassigning anomalous items, and then reinputting the data to the pipeline with the modified categorisations to produce the final visualisations and summaries to be included in the finished analysis. This mimics the process used in some thematic analyses where items are initially grouped into general categories before being re-analysed to identify sub-categories (Joffe & Yardley, 2003; Parliamentary Office of Science and Technology, 2020b).

By running multiple rounds of automated analysis set to produce varying numbers of clusters, more general and more specific concerns can be identified, as demonstrated in section 4.8. While using a large number of clusters may produce some duplication, where a human may judge the differences between clusters to be unimportant, it can pick out smaller but distinct sub-topics. These smaller topics may provide greater insight into the data and depending on the objectives of the analysis they may decide some of these smaller topics might be included in the final analysis or used to inform their discussion of the results, as in thematic analysis (Braun & Clarke, 2006).

By automating a highly labour-intensive part of the analysis process, the approach has the potential to massively reduce the time required for analysing large text datasets and make practical the analysis of larger datasets than is otherwise possible without introducing the

complexities of subdividing the data among multiple analysts and additionally may help highlight areas within the data which might otherwise be overlooked.

The potential applications for this approach are broad and include analysis of survey data (as with the COVID-19 Expert Concerns Survey), presenting overviews of large libraries or collections (as with the TED Conferences dataset), or for business intelligence (as with the Isle of Wight Supply Chain dataset). The results of the COVID-19 Expert Concerns Survey analysis were presented to the UK Parliament Select Committee for COVID-19 and were used to create POST's COVID-19 broad areas of research interest (Parliamentary Office of Science and Technology, 2020b). Chapter 5, section 5.3.3 details ongoing collaboration with the Wessex Institute and James Lind Alliance looking at applying this tool to survey data in the healthcare domain.

Chapter 5 **Conclusions and Future Work**

5.1 Conclusions

This thesis has explored several methods for generating insights and human-interpretable overviews from large heterogeneous datasets, employing a variety of techniques including recommender systems, visualisation, clustering, text summarisation, and keyword extraction. These techniques have been demonstrated in a variety of real-world scenarios including supply-chain recommendations and topical analysis of survey data.

The novel Transitive Semantic Relationships (TSR) model introduced in Chapter 3 addresses especially challenging cases of the cold start problem, where recommendations must be made for unlabelled items using the few labels known for a large dataset. The solution is robust to noisy text data collected from web-scraping by making use of text pre-processing and highly generalisable upstream deep learning models for producing semantic representations of text. The solution also generates detailed provenance in the form of a list or graph of items and relationships considered (both ground truth and predictions), which is easy to visualise. This work provides one answer to SRQ1: “How can machine understanding of text be used to identify relationships between documents in large collections of unstructured text?”. TSR has already seen real-world adoption by an industrial partner, details of this application are given in section 5.3.1.

The approach taken in Chapter 4 of combining several traditional text analysis and mining techniques into a Text Insights Pipeline (TIP) addresses the challenge of structuring and presenting data for efficient and effective human analysis and provides a method of automating highly labour-intensive stages of thematic analysis or producing a preliminary analysis for review by experts. Like in Chapter 3, the use of text pre-processing and generalised deep-learning models allows for handling highly heterogeneous text, including free-text survey responses, summaries of conference presentations, and web-scraped descriptions of companies. These features are then used for clustering, visualisation, and summarisation to produce one comprehensive overview of the data, which can be presented as a report. This work offers a solution to SRQ2: “How can machine understanding of text be used to produce an interpretable overview of large collections of unstructured text?”, and SRQ3: “How can the results of text analysis be effectively presented and used to inform decision-makers, analysts, and organisations?”. This approach was used in collaboration with the Parliamentary Office of Science and Technology (POST) to produce an overview of expert’s concerns regarding COVID-19 in the United Kingdom which led to POST’s COVID-19 broad areas of research interest, details of the collaboration and its impacts are

detailed in section 5.3.2. The tool (TIP) is also employed in an ongoing collaboration with the Wessex Institute (see section 5.3.3), and further development and investigation into additional applications are ongoing.

Throughout the thesis, it has been demonstrated that combining the generalisability and deep understanding of language from modern deep learning approaches, such as neural language models, with traditional, well understood, and explainable down-stream algorithms is an effective technique for producing high-quality results on a variety of heterogeneous text data while preserving some degree of explainability and can produce results which are informative, and well evidenced. In combination with the examination of the sub-research-questions given previously, this provides an effective solution to the overall research question of “How can machine understanding of text be used to produce insights from large collections of unstructured text to inform decision-makers, analysts, and organisations”.

5.2 Contributions

This research makes several major contributions:

The Transitive Semantic Relationships (TSR) approach for cold-start recommendations in sparsely labelled data. This novel algorithm addresses a challenging and often overlooked edge case in recommender systems (SRQ1) and also addresses some common criticisms of other recommender systems such as poor explainability (SRQ3). The algorithm has the potential for significant impact in industry, particularly in high-velocity big data such as supply chain recommendations.

The COVID-19 Expert Concerns survey analysis demonstrates an effective methodology for applying and combining a variety of data analysis techniques to produce easily interpretable visualisations, categorisations, and summaries of large text datasets such as survey data (SRQ2 and SRQ3). The results of this analysis helped inform UK parliament in setting research priorities.

The Text Insights Pipeline is a web-based research tool that enables data analysts to make use of the methodology used for the COVID-19 Expert Concerns survey analysis without requiring specialist or technical knowledge of the techniques or algorithms involved (SRQ3). This tool has attracted attention from external organisations with interest in using it for various real-world applications and further research is ongoing with their support.

5.3 Impact

5.3.1 Impact in Industry

The Transitive Semantic Relationships (TSR) recommender algorithm was developed partly during an ICASE placement with KnowNow Information. A partner company, Launch International LTD provided the data and labelling for the Isle of Wight Supply Chain Dataset on which the algorithm was demonstrated. Launch International LTD have incorporated this approach into their technology platform "Find Engine". A testimonial from the company's Chief Technical Officer is included below.

> "The algorithm has been used as the core service in Launch to provide recommendations for businesses to connect to other businesses based on their profiles. The algorithm is trained based on sampling data from a list of Isle of Wight businesses. In the database, we have experts to label the supply chain relationships between any two businesses from the list. Then the TSR algorithm analyses the profile of each business and learns the supplier-provider relationships from the labelled data. Then we have applied the algorithm in a large dataset based on a couple of LEP's business catalogues and automatically recommend which businesses can potentially collaborate on supply chain. The software has been used by Solent LEP, Oxford Innovation and some business relationship management departments in universities."
>
> - **Dr Yunjia Li, CTO Launch International LTD**

5.3.2 Impact in Parliament

The results from Chapter 4 for the COVID-19 Expert Concerns dataset, including the alternate categorisation scheme produced through automated clustering and the figures and visualisations from that chapter were presented to UK Parliament and the Lords COVID-19 Committee as a report similar to those generated by the Text Insights Pipeline. This evidence was used by POST, UK Parliament, and select committee staff to produce the Areas of Research Interest (ARIs) for COVID-19 (Parliamentary Office of Science and Technology, 2020a).

5.3.3 Collaboration with James Lind Alliance and Wessex Institute

The Text Insights Pipeline (TIP) produced as part of the work described in Chapter 4 has already received interest from external organisations including the Wessex Institute and the James Lind Alliance (JLA). The JLA have in particular expressed interest in using the tool for their Priority

Setting Partnerships, similar to how the results for the Expert Concerns dataset were used by POST. Demonstrations of the tool have been made to the Wessex Institute and JLA on their own datasets; the tool was given praise for its performance, particularly in its ability to produce many similar observations as human analysts, but in much less time.

At the time of writing, a joint project is currently underway with the Wessex Institute and JLA to formally evaluate the performance of the tool and its suitability for applications within these organisations. Funding is also currently being sought to develop this tool further with the aim of making it available as a research tool. It is expected this will be with the support of either the Wessex Institute, JLA, or both, and with the aim of working with other research organisations as the project progresses.

5.4 Future Work

5.4.1 Evaluating TSR on Other Datasets

The TSR model has been demonstrated to produce good results on the challenging IWSC dataset tasks, where other recommender algorithms would be unsuitable for the reasons discussed in Chapter 3.

In future research, TSR might be evaluated on more general datasets to allow direct comparison with other recommender systems. As TSR is designed specifically for difficult scenarios where there are very few training labels, it is likely that it will only outperform existing solutions in these cases, and that the existing solutions designed with the assumption of abundant training labels will perform better on such datasets.

However, while direct comparisons of algorithms would be biased in favour of those optimised for datasets of that nature, the relationship between the performance of each algorithm and the number of labels in the dataset is of interest. Therefore, standard benchmark datasets might be used for evaluation but with varying numbers of labels included in the training set, similar to the approach used by Cer et al., (2018) to determine the impact of transfer learning with different quantities of labelled data. By varying the number of labels available for training, the suitability of each algorithm can be assessed for different degrees of sparsity.

Testing on other datasets should prove the generality of the TSR model, and comparison to other models when varying the amount of training data should demonstrate their suitability for different scenarios. The expected result would be that when there are very few labels in the training set TSR will be the best performing algorithm, but when there are many labels TSR will underperform.

An exhaustive list of models and datasets that might be tested cannot be provided here as it would be subject to the compute resources available and future advances in the field. However, some models of interest include traditional Collaborative filtering, Neural collaborative filtering (He et al., 2017), and the averaging of embeddings technique typically used to generate recommendations from neural language models such as used for a baseline by Suglia et al. (2017). Some of these methods do not support cold-starts, and so may not be suitable for all testing scenarios.

5.4.2 Investigating Effects of Embedding Models

Both the TSR algorithm and some analysis steps in the Text Insights Pipeline are highly dependent on effective measures of semantic similarity between items. In this project, the Universal Sentence Encoder (Cer et al., 2018) model has been used to generate the content embeddings used to calculate similarity. The reasons for choosing this model for the initial investigation are given in section 3.8, but no formal investigation of alternative models has been conducted so far.

Both TSR and TIP make use of fixed-length content embeddings as input but do not dependent on any particular model, so long as all embeddings for a given set of data come from the same model. As a result, it is trivial to switch the upstream model for any other that outputs fixed-length semantic content embeddings. The TSR evaluation toolkit was built with the capability to specify the embedding model to be used so that this could be easily investigated in future, and the TIP web tool is expected to be given this functionality later in its ongoing development, giving analysts easy opportunity to explore the efficacy of different models.

For comparison of performance, models of interest include some of those discussed in section 2.3, and particularly the recent state-of-the-art models discussed in section 2.3.7.

5.4.2.1 Beyond English Language Text

An additional possibility beyond the scope of this thesis is the application of these approaches (TSR and TIP) to other languages. Embedding models have been produced for many different natural languages, including multi-language models (Conneau et al., 2020) and the algorithms employed by TSR and TIP are agnostic to the language used; this is purely dependent on the embedding model. For TIP, items within a dataset should all be in the same language due to the approach used for summarisation and category naming. TSR is completely language agnostic and multi-lingual datasets could be used with a multi-lingual embedding model. TSR could also be applied to non-text content embedding such as those produced by multi-modal embedding models (Sun et al., 2018; Sung et al., 2017), these could also be used with TIP but some text data would still be required for summarisation and category naming.

5.4.2.2 Fine Tuning

The benefits of fine-tuning upstream models on downstream task performance are well known (Cer et al., 2018; Conneau et al., 2017). These sources attribute better vocabulary coverage of domain-specific terminology to be a significant factor in the performance improvements.

In the case of investigating supply-chain, this is likely to be a significant factor as domain-specific terminology is common. As such, fine-tuning the upstream embedding model on company

descriptions or similar documents could potentially provide significant performance improvements on IWSC tasks.

For exploring survey data, a model fine-tuned on domain-specific text may enable better performance, particularly in domains that use highly specialised vocabulary (such as medicine and the sciences).

5.4.3 Choice of Algorithms for TIP

The approach presented in Chapter 4 for producing automated analysis and summarisation of text datasets makes use of various existing techniques for automated text summarisation and clustering. As previously discussed in concern to the chosen embedding model, these may be interchanged with other models and algorithms for these tasks.

In Chapter 4, extractive summarisation is used for extracting key sentences to produce the summaries for each category and for identifying keywords to form provisional category names. That chapter gives the rationale for choosing extractive techniques over abstractive ones, however, there is room for an investigation into their suitability. A detailed study of the generality and biases of pre-trained learning models when applied in this way could be of interest both for building on the approach presented here and also for consideration by the creators of the upstream models. Category naming may also benefit from abstractive summarisation as an alternative to the keyword/keyphrase extraction used.

Additionally, clustering methods besides K-means and K-medoids could be employed which may offer more tuneable parameters or perform better on datasets with particular distributions, such as where clusters are more clearly separable or non-convex. Some alternative approaches for topic modelling, such as LDA, could extend functionality to cover cases like the assignment of items to multiple categories.

More generally, other algorithms and models could be used in place of the ones used in the Text Insights Pipeline while still following the same methodology to further improve its performance and extend functionality to specific technical domains, other languages, other types of media, and a wide range of other use cases and applications.

List of References

Adomavicius, G., & Tuzhilin, A. (2001). Using data mining methods to build customer profiles. *Computer*, *34*(3), 74–82. https://doi.org/10.1109/2.901170

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 207–216. https://doi.org/10.1145/170035.170072

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, *07-09-Janu*, 1027–1035.

Banik, R. (2017). *TED Talks | Kaggle*. https://www.kaggle.com/rounakbanik/ted-talks

Barrios, F., López, F., Argerich, L., & Wachenchauzer, R. (2015). Variations of the Similarity Function of TextRank for Automated Summarization. *44th Argentine Conference on Informatics*. http://arxiv.org/abs/1602.03606

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, *17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

Bendakir, N., & Aïmeur, E. (2006). Using association rules for course recommendation. *AAAI Workshop - Technical Report*, *WS-06-05*, 31–40.

Bengio, Y., Ducharme, R., & Vincent, P. (2001). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, *3*, 1137–1155.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. https://doi.org/10.18653/v1/D15-1075

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

List of References

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1/7), 107–117. https://doi.org/10.1.1.109.4049

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-Decem*. https://commoncrawl.org/the-data/

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 104047. https://doi.org/10.1016/j.jml.2019.104047

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder for English. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174. https://doi.org/10.18653/v1/D18-2029

Chen, W. H., Cai, Y., Leung, H. F., & Li, Q. (2010). Generating ontologies with basic level concepts from folksonomies. *Procedia Computer Science*, *1*(1), 573–581. https://doi.org/10.1016/j.procs.2010.04.061

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 93–98. https://doi.org/10.18653/V1/N16-1012

Conneau, A., & Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC} 2018)*, *abs/1803.05449*. http://arxiv.org/abs/1803.05449

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, *abs/1705.02364*, 670–680. https://doi.org/10.18653/v1/D17-1070

Conneau, A., Rinott, R., Lample, G., Schwenk, H., Stoyanov, V., Williams, A., & Bowman, S. R. (2020). XNLI: Evaluating cross-lingual sentence representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2475–2485. https://doi.org/10.18653/V1/D18-1269

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Deldjoo, Y., Dacrema, M. F., Constantin, M. G., Eghbal-zadeh, H., Cereda, S., Schedl, M., Ionescu, B., & Cremonesi, P. (2019). Movie genome: alleviating new item cold start in movie recommendation. *User Modeling and User-Adapted Interaction*, *29*(2), 291–343. https://doi.org/10.1007/s11257-019-09221-y

Devlin, J., Chang, M.-W. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, *1*(abs/1810.04805), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dolan, W. B., & Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 9–16. https://research.microsoft.com/apps/pubs/default.aspx?id=101076

Dupplaw, D., Matthews, M., Richard, J., & Lewis, P. (2012). LivingKnowledge: a platform and testbed for fact and opinion extraction from multimodal data. *Eternal Systems*, *255*, 100–115. https://eprints.soton.ac.uk/350579/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gowda, T., & May, J. (2020). Finding the Optimal Vocabulary Size for Neural Machine Translation. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3955–3964. https://doi.org/10.18653/v1/2020.findings-emnlp.352

Grady, C., & Lease, M. (2010). Crowdsourcing Document Relevance Assessment with Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, *June*, 172–179. http://dl.acm.org/citation.cfm?id=1866696.1866723

Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets. *ACM Transactions on Interactive*

Intelligent Systems, *5*(4), 1–19. https://doi.org/10.1145/2827872

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural Collaborative Filtering. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 173–182. https://doi.org/10.1145/3038912.3052569

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work - CSCW '00*, 241–250. https://doi.org/10.1145/358916.358995

Hill, F., Cho, K., & Korhonen, A. (2016). Learning Distributed Representations of Sentences from Unlabelled Data. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1367–1377. https://doi.org/10.18653/v1/N16-1162

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. https://doi.org/10.1037/h0071325

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). BibSonomy: A Social Bookmark and Publication Sharing System. *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, 87–102. http://www.kde.cs.uni-kassel.de/hotho/pub/2006/iccs_tools_ws_final.pdf

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*, 168. https://doi.org/10.1145/1014052.1014073

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM '13*, 2333–2338. https://doi.org/10.1145/2505515.2505665

Ibrahim, N., Chaibi, A. H., & Ghézala, H. Ben. (2017). Scientometric re-ranking approach to improve search results. *Procedia Computer Science*, *112*, 447–456. https://doi.org/10.1016/j.procs.2017.08.020

IWChamber. (2018). https://www.iwchamber.co.uk

IWTechnology. (2018). http://iwtechnology.co.uk/

Jackson Kristi, & Bazeley Pat. (2019). *Qualitative Data Analysis with NVivo* (3rd ed.). Sage

Publications.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. https://doi.org/10.1145/582415.582418

Joffe, H., & Yardley, L. (2003). Content and thematic analysis. In *Research Methods for Clinical and Health Psychology* (pp. 56–68). SAGE Publications. https://www.researchgate.net/publication/313157845_Content_and_thematic_analysis

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-Aware Neural Language Models. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2741–2749. https://doi.org/2

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-Thought Vectors. *Advances in Neural Information Processing Systems*, *786*, 3294–3302. http://arxiv.org/abs/1506.06726

Kong, W., Li, R., Luo, J., Zhang, A., Chang, Y., & Allan, J. (2015). Predicting Search Intent Based on Pre-Search Context. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, 503–512. https://doi.org/10.1145/2766462.2767757

Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1401890.1401944

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, *32*, 29–30. https://doi.org/10.1145/2740908.2742760

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. https://doi.org/10.3115/v1/P14-2050

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and Understanding Neural Models in NLP. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the*

Conference, 681–691. https://doi.org/10.18653/V1/N16-1082

Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, *September*, 1722–1732. https://doi.org/10.18653/v1/d15-1200

Li, X., & Roth, D. (2002). Learning Question Classifiers. *COLING '02 Proceedings of the 19th International Conference on Computational Linguistics*, *1*, 1–7. https://doi.org/10.3115/1072228.1072378

Liebling, D., Bennett, P. N., & White, R. (2012). Anticipatory Search: Using Context to Initiate Search. *The 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1035–1036. https://doi.org/10.1145/2348283.2348456

Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., & Luis, T. (2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, *September*, 1520–1530. https://doi.org/10.18653/v1/D15-1176

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., & Allen, P. G. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://github.com/pytorch/fairseq

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lops, P., De Gemmis, M., Semeraro, G., Musto, C., & Narducci, F. (2013). Content-based and collaborative techniques for tag recommendation: An empirical evaluation. *Journal of Intelllgent Information Systems*, *40*(1), 41–61. https://doi.org/10.1007/s10844-012-0215-6

Maaten, L. Van Der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Maedche, A., & Staab, S. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, *16*(2), 72–79. https://doi.org/10.1109/5254.920602

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. *Lrec*, *May*, 216–223. https://doi.org/10.1.1.486.5834

Marine Southeast. (2018). http://www.marinesoutheast.co.uk/

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, *22*(1), 54–88. https://doi.org/10.1145/963770.963773

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions -*, *4*, 20-es. https://doi.org/10.3115/1219044.1219064

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. https://www.aclweb.org/anthology/W04-3252

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop Papers*. https://doi.org/10.1162/153244303322533223

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*, 1–9. https://doi.org/10.1162/jmlr.2003.3.4-5.951

Musto, C., Semeraro, G., de Gemmis, M., & Lops, P. (2016). Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 729–734. https://doi.org/10.1007/978-3-319-30671-1_60

Musto, C., Semeraro, G., de Gemmis, M., Lops, P., Ferro, N., Crestani, F., Moens, M. F., Mothe, J., Silvestri, F., Di Nunzio, G. M., Hauff, C., & Silvello, G. (2016). Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9626*, 729–734. https://doi.org/10.1007/978-3-319-30671-1_60

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

Nakhaeizadeh, G., Hipp, J., & Güntzer, U. (2000). Algorithms for association rule mining - a general survey and comparison. *ACM Sigkdd Explorations Newsletter*, *2*(1), 58–64.

OWL 2 Web Ontology Language Document Overview (Second Edition). (2012).

https://www.w3.org/TR/owl2-overview/

Pang, B., & Lee, L. (2005). Seeing stars. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, *1*, 115–124. https://doi.org/10.3115/1219840.1219855

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 271-es. https://doi.org/10.3115/1218955.1218990

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Parliamentary Office of Science and Technology. (2020a). *COVID-19 Areas of Research Interest - POST*. https://post.parliament.uk/covid-19-areas-of-research-interest/

Parliamentary Office of Science and Technology. (2020b). *COVID-19 outbreak: What are experts concerned about? - POST*. https://post.parliament.uk/covid-19-outbreak-what-are-experts-concerned-about/

Parliamentary Office of Science and Technology. (2020c). *Expert acknowledgements - POST*. https://post.parliament.uk/expert-aknowledgements/

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). Sage publications.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., … Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://scikit-learn.sourceforge.net.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Porter, M. F. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Ralph, D. (2021). *Text Insights Pipeline Demo*. https://davidralph.github.io/TIP-Demo/

Ralph, D., Li, Y., Wills, G., & Green, N. G. (2019, July 30). *Transitive Semantic Relationships (TSR)*. https://doi.org/10.5281/ZENODO.3355448

RDF 1.1 Concepts and Abstract Syntax. (2014). https://www.w3.org/TR/rdf11-concepts/

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Reimers, N., & Gurevych, I. (2020). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence A Modern Approach*. Prentice Hall.

Sculley, D. (2010). Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 1177. https://doi.org/10.1145/1772690.1772862

See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1073–1083. https://doi.org/10.18653/v1/P17-1099

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., & Riedl, J. (2006). Tagging, communities, vocabulary, evolution. *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work - CSCW '06*, 181. https://doi.org/10.1145/1180875.1180904

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1715–1725. https://doi.org/10.18653/v1/p16-1162

Seo, S., Kim, C., Kim, H., Mo, K., & Kang, P. (2020). Comparative Study of Deep Learning-Based Sentiment Classification. *IEEE Access*, *8*, 6861–6875. https://doi.org/10.1109/ACCESS.2019.2963426

Shalaby, W., Alaila, B. E., Korayem, M., Pournajaf, L., Aljadda, K., Quinn, S., & Zadrozny, W. (2018). Help me find a job: A graph-based approach for job recommendation at scale. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 1544–1553. https://doi.org/10.1109/BigData.2017.8258088

Snow, R., Connor, B. O., Jurafsky, D., Ng, A. Y., Labs, D., & St, C. (2008). Cheap and Fast — But is it Good ? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *EMNLP '08*, 254–263. http://dl.acm.org/citation.cfm?id=1613715.1613751

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://www.aclweb.org/anthology/D13-1170

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Suglia, A., Greco, C., Musto, C., De Gemmis, M., Lops, P., & Semeraro, G. (2017). A deep architecture for content-based recommendations exploiting recurrent neural networks. *[UMAP2017]Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 202–211. https://doi.org/10.1145/3079628.3079684

Sun, M., Li, F., & Zhang, J. (2018). A Multi-Modality Deep Network for Cold-Start Recommendation. *Big Data and Cognitive Computing*, *2*(1), 7. https://doi.org/10.3390/bdcc2010007

Sung, J., Lenz, I., & Saxena, A. (2017). Deep multimodal embedding: Manipulating novel objects with point-clouds, language and trajectories. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2794–2801. https://doi.org/10.1109/ICRA.2017.7989325

Takase, S., Suzuki, J., Okazaki, N., Hirao, T., & Nagata, M. (2016). Neural Headline Generation on Abstract Meaning Representation. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1054–1059. https://doi.org/10.18653/V1/D16-1112

TED Conferences LLC. (n.d.). *TED: Ideas worth spreading*. Retrieved July 5, 2021, from https://www.ted.com/

Vuurens, J. B. P., Larson, M., & de Vries, A. P. (2016). Exploring Deep Space: Learning Personalized Ranking in a Semantic Space. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems - DLRS 2016*, 23–28. https://doi.org/10.1145/2988450.2988457

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C. L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text

classification. *Neurocomputing*, *174*, 806–814. https://doi.org/10.1016/j.neucom.2015.09.096

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210. https://doi.org/10.1007/s10579-005-7880-9

Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. http://arxiv.org/abs/1609.08144

Xie, K., Wang, C., & Wang, P. (2021). A Domain-Independent Ontology Learning Method Based on Transfer Learning. *Electronics*, *10*(16), 1911. https://doi.org/10.3390/electronics10161911

Xu, Z., Chen, C., Lukasiewicz, T., Miao, Y., & Meng, X. (2016). Tag-Aware Personalized Recommendation Using a Deep-Semantic Similarity Model with Negative Sampling. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, *24-28-Octo*, 1921–1924. https://doi.org/10.1145/2983323.2983874

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 1445–1456. https://doi.org/10.1145/2488388.2488514

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*. https://github.com/zihangdai/xlnet

Yin, D., Nobata, C., Langlois, J.-M., Chang, Y., Hu, Y., Tang, J., Daly, T., Zhou, M., Ouyang, H., Chen, J., Kang, C., & Deng, H. (2016). Ranking Relevance in Yahoo Search. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 323–332. https://doi.org/10.1145/2939672.2939677

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding Neural Networks Through Deep Visualization. *Deep Learning Workshop, 31st International Conference on Machine Learning*. http://arxiv.org/abs/1506.06579

Yuan, J., Shalaby, W., Korayem, M., Lin, D., Aljadda, K., & Luo, J. (2016). Solving cold-start problem in large-scale recommendation engines: A deep learning approach. *Proceedings - 2016 IEEE*

International Conference on Big Data, Big Data 2016, 1901–1910. https://doi.org/10.1109/BigData.2016.7840810

Zhang, F., Yuan, N. J., Lian, D., Xie, X., & Ma, W.-Y. (2016). Collaborative Knowledge Base Embedding for Recommender Systems. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 353–362. https://doi.org/10.1145/2939672.2939673

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys*, *52*(1), 1–38. https://doi.org/10.1145/3285029

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 19–27. https://doi.org/10.1109/ICCV.2015.11

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *ICLR 2017 5th International Conference on Learning Representations*, 1–12. http://arxiv.org/abs/1702.04595