Start-ups: Towards an Understanding of Their Role in Science

Aidan O'Brien

Master of Science in Data Science
The University of Bath
September 2021

Start-ups: Towards an Understanding of Their Role in Science

Submitted by: Aidan O'Brien

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Bachelor of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

BioNTech, a science-based start-up founded in 2008, laid the foundations for the first approved COVID-19 vaccine with their innovative mRNA technology. Evidently, the scientific research of start-ups can provide large amounts of social welfare. Analyses of bibliometric data have previously suggested that large companies play a vital role in the progression of science. To investigate the role of start-ups in science, this dissertation created a novel data set of over one million paper-institution pairs spanning 2000-2009 in the United Kingdom. Start-ups and other institutions were categorised following the use of lexical similarity techniques which allowed four main data sources to be combined: Scopus, the Global Research Identifier Database, Companies House, and Financial Analysis Made Easy. Statistical analyses were conducted to compare the research focus and quality of start-up papers to other institutions. Using word2vec, abstracts were then represented in a two-dimensional semantic space to understand the motivations of start-ups for publishing scientific research. This project presents the methodology used for creating the data set and the results from statistical testing. Furthermore, the methodology for creating abstract embeddings, as well as the results of visual exploration are also reported.

This project found that the research focus of start-ups was different to other institutions. Using citation and journal metrics, it was also found that the quality of start-up papers was worse than or equal to other institutions, depending on the research area. Seismology and structural chemistry were examples of areas of semantic space with a distinct lack of start-up papers. Two explanations were proposed for such absences: a lack of monetisation of the research and protection of intellectual property. The results of this project showed that start-ups did contribute to scientific knowledge, but the presence and quality of their research depended on the research topic. Furthermore, these findings will help future research to understand the role start-ups play in the transition from scientific knowledge to technological advancements and social welfare.

Contents

List of Figures

List of Tables

Acknowledgements

Thanks are due to the following people: My supervisor, Dr Tom Fincham Haines for his continued support and encouragement throughout this project. Dr Stefano Baruffaldi, Dr Virgilio Failla, and Dr Rossella Salandra from the School of Management for their extensive expertise in the entrepreneurial and start-up literature. Dr Maik Schneider from the Department of Economics for his inspirational lectures on growth theory and the economics of innovation and entrepreneurship that sparked my initial interest in this area of research.

I would also like to thank the University of Bath for their financial support throughout my undergraduate and postgraduate studies. With the support of the Global Leaders Scholarship, I have been able to enjoy the academically challenging experience that comes with university life.

Finally, I thank my friends and family who have always supported me throughout my academic studies. Special thanks go to my partner, Grace. She has provided endless support to this project with her extensive knowledge of the biomedical literature and her ability to make me forget about annoying bugs in the evening.

Chapter 1

Introduction

1.1 Background

Historically, small businesses have been overlooked in their importance as a means for economic growth, due to the dominance of large corporations during the industrial revolution (Ackermann, 2012). However, fast-forward to the 21st century where $136 billion was raised in venture capital funds in 2019 alone, and it is clear that small businesses play a vital role in the modern economy (Oecd, 2021). Start-ups are driving this surge in investment interest. This interest has delivered great value to the world economy in terms of jobs. According to Business Dynamics Statistics, start-ups contributed 2.9 million new jobs to the US economy per year between 1980 and 2010 (Decker et al., 2014). Defined as a company, partnership, or temporary organisation designed to search for a repeatable and scalable business model, start-ups also provide modern economies an avenue for continuous experimentation at the technology frontier (Blank, 2010; Ackermann, 2012).

Recent advances in mRNA technology from young companies, such as BioNTech, provide ample evidence for the vast amount of social value that science-based start-ups can bring to society. Founded in 2008 with a seed investment of $150 million, BioNTech started working on mRNA-based therapeutics that laid the foundations for the first approved COVID-19 vaccine (Ledford, Cyranoski and Van Noorden, 2020; SEC, 2020). With research published as early as 2012, BioNTech has been contributing to the scientific foundations of mRNA therapeutics and engineered cell therapies (Castle et al., 2012).

Despite the potential value that science-based start-ups can bring to society, the tail of Theranos offers an important warning of the dangers associated with start-ups and their technologies. Valued at $9 billion at its height, Theranos caught the attention of investors with its ground-breaking blood testing technology. It has been argued in the highly popular book *Bad Blood: Secrets and Lies at a Silicon Valley Startup*, that the company was based on fraudulent technology (Carreyrou, 2018). Following the unveiling of the scandal, Theranos was liquidated and concerns about the lack of peer-reviewed research for science-based start-ups were raised (Cristea, Cahan and Ioannidis, 2019).

The two contrasting case-studies of BioNTech and Theranos highlight the importance of scientific rigour in the application of start-up technology. More specifically, BioNTech was publishing research in peer-reviewed journals early on, whereas Theranos did not publish any research in the literature regarding their blood-sampling technology (Ioannidis, 2015).

1.2 Motivation

This project was motivated by three primary reasons. Firstly, the case study of BioNTech suggests that start-ups provide a valuable contribution to the progression of science. However, a thorough search of the literature revealed no research investigating the extent to which this statement is true. Previous work has looked at the characteristics of industry and academia research (Aghion, Dewatripont and Stein, 2008), yet little is understood about the research of start-ups.

Secondly, if start-ups do provide notable contributions to science, little is known about how their contributions differ from other institutions, such as universities and large companies. Furthermore, it is not understood how the research of start-ups translates into applied use cases. The advent of online bibliometric and patent data have allowed researchers to begin to understand how scientific knowledge is translated into technological advancements (Azoulay, Zivin and Sampat, 2012; Ahmadpoor and Jones, 2017; Marx and Fuegi, 2020). Recent research has even looked at the role that large companies play in this transition (Arora, Belenzon and Sheer, 2021), but there is much to be learnt about the role of start-ups in this process. This project will set the foundations for developing this understanding and will have policy implications for governments wanting to efficiently allocate resources to drive scientific and technological progress.

Thirdly, this project was motivated by the absence of research publications as a predictor of start-up survival in the literature. Although this literature has attracted a lot of attention due to the financial gains associated with it (Gartner, Starr and Bhat, 1999; Hyytinen, Pajarinen and Rouvinen, 2015), it is yet to formally incorporate start-up publications as a predictor. The benefit of identifying successful science-based start-ups early on is that large amounts of welfare will not be lost, like in the case of Theranos. Similar to Conti, Thursby and Thursby (2013), this project will also provide a basis for investigating investor signalling effects. However, the products of this project will allow scientific publications to be researched as a form of signalling as opposed to patents, like in Conti, Thursby and Thursby (2013).

1.3 Project Objectives

Following the motivations for this project and the literature review presented in Section 2, the project objectives can be formally stated as follows:

Objective One

Identify and collect data from relevant sources on scientific research papers and start-ups.

Objective Two

Create a novel data set linking scientific research papers to different types of institutions, including start-ups.

Objective Three

Explore the novel data set to evaluate the contribution of start-up publications to the progression of science.

Objective Four

Use natural language processing (NLP) techniques to provide further insights on the research differences of start-up companies.

1.4 Outline

The rest of this document is split into six main chapters. First, Chapter 2 provides a comprehensive review of the data sources available for creating a data set, as well as covering text similarity techniques. Chapter 3 then outlines the methodology used to achieve the first and second objectives of the project: create a data set linking research papers to institutional types. This process utilised four main data sources - Scopus, the Global Research Identifier Database (GRID), Companies House, and Financial Analysis Made Easy (FAME). Next, Chapter 4 details the experimentation procedure used for different methods of creating abstract embeddings. This chapter also proposes two different techniques for assessing the ability of the abstract embeddings to capture semantic meaning.

The creation of the data set was a significant contribution of this project which warranted extensive exploration and discussion. Therefore, Chapter 5 is dedicated to reporting a descriptive and statistical analysis of the data set. Statistical testing was based on the hypotheses outlined in Section 5.1, which were derived from the third objective. Discussion of the results in relation to relevant parts of the literature is also provided throughout this chapter.

Chapter 6 first reports the results of extensive experimentation carried out to produce abstract embeddings. Quantitative and qualitative evaluation validated the ability of the embeddings to capture the semantic meaning of the abstracts. After using a dimensionality reduction technique, the abstract embeddings were then visually explored to meet objective four. Namely, further insights into start-up research were generated using NLP techniques.

Chapter 7 summarises the results and achievements of this project. This chapter also reflects on the limitations of the methods used, as well as proposing alternative methods. Finally, some opportunities for future work are also discussed.

Chapter 2

Literature Review

2.1 Data

Given objectives one and two, it was clear that data relating to research publications and start-up companies were required for this project. This section is dedicated to exploring the different data sources that were available for use in this project.

2.1.1 Research Papers

Recently, there have been efforts from various platforms to challenge the dominance of Web of Science (WoS), Scopus, and Google Scholar for multidisciplinary bibliographic data. Such platforms include Microsoft Academic (MAG), Dimensions, and Crossref. As the number of data sources has increased, it is important to keep in mind their respective strengths and weaknesses. Fortunately, researchers have provided numerous attempts to compare these bibliographic data sources. It should be noted that these analyses take different approaches to their comparisons, as well as comparing different combinations of data sources.

For coverage, it has been found that MAG and Google Scholar are significantly better than Crossref, Dimensions, Scopus and WoS (Visser, van Eck and Waltman, 2021; Harzing, 2019). However, Google Scholar is only accessible through a search engine, meaning that it is not suitable for providing large scale data. In contrast, the previously mentioned sources, as well as PubMed, are readily available through an API or dump download. Comparing the coverage in different disciplines, Visser, van Eck and Waltman (2021) also found that MAG performed better than WoS, Dimensions and Crossref when compared to Scopus across multiple disciplines such as physical sciences, health sciences, life sciences, and social sciences and humanities. The authors noted that the use of Scopus as a baseline for comparison did not indicate that Scopus is considered the most comprehensive data source.

Cristea, Cahan and Ioannidis (2019) provided an example of constructing a data set showing biomedical unicorns (private companies valued over $1 billion) and the number of research papers associated with them. The authors used a combination of PubMed and Scopus for their research. More specifically, they used a list of pre-determined healthcare unicorns to search PubMed for affiliated institutions. Resulting publications were then cross-referenced with Scopus to find citation data that is absent from PubMed. An important limitation of PubMed that the authors failed to acknowledge relates to incomplete affiliation data. Namely,

PubMed only reported the affiliation of the first author prior to 2014. This means that a large amount of data that could link start-ups to papers is missing from the data source. In the context of this project, the approach taken by Cristea, Cahan and Ioannidis (2019) was deemed unsuitable as all start-up firms publishing papers should be identified. Despite the highly structured nature of PubMed, it also lacks the coverage of other data sources available due to its focus on biomedical research (Shultz, 2007). Therefore, creating a data set from PubMed would limit the project to biomedical firms and miss out on many start-ups.

This literature review found no further research linking start-up companies to research papers. As the approach taken by Cristea, Cahan and Ioannidis (2019) was deemed unsuitable for this project, only the more comprehensive data sources - MAG, Crossref, Dimensions, Scopus and WoS - were considered. Notably, MAG is no longer available due to the recent suspension of academic services by Microsoft. From the remaining data sources, Scopus and WoS were freely available through the University of Bath Library Services. Scopus had the advantage over WoS by being the largest abstract database of peer-reviewed literature (Scopus, 2021). In this project, abstracts were used to gain additional insights into the research of start-ups. Therefore, Scopus had the advantage over the other data sources discussed in this section.

2.1.2 Start-ups

There are numerous data sources available that have been used in the Economic and Management literature to study the dynamics of individual firms. For example, Compustat has been popular among academics as it provides a large database of financial and market information on active and inactive companies around the world (S&P, 2021). However, one important limitation of Compustat is that it only contains data on public companies. Given the focus on start-up companies for this project, this database was therefore deemed unsuitable. Fortunately, private company databases, such as CB Insights, CrunchBase, and FAME, have emerged as market research tools (CBInsights, 2021; CrunchBase, 2021; Dijk, 2021).

Cristea, Cahan and Ioannidis (2019) used CB Insights to identify unicorns in the healthcare industry. However, CB Insights has not been used widely in the literature due to a far more exhaustive database: CrunchBase. The use of CrunchBase led to more than 90 academic publications up to 2017, mostly concentrated in the Management literature (Dalle, den Besten and Menon, 2017). For example, Tata et al. (2017) created a data set of Twitter accounts of San Francisco based start-up founders through CrunchBase data. Although CrunchBase is more comprehensive than CB Insights for start-up data and is readily available through an API, it does cost to obtain use.

The FAME database contains extensive financial and other data for registered companies in the United Kingdom. Importantly, the University of Bath Library Services has access to this database. This gave it the advantage over CrunchBase and CB Insights. FAME has been used in previous research to look at the performance of high-tech small and medium-sized enterprises in the United Kingdom, as it contains multiple metrics for firm size (Crick and Spence, 2005). This data was vital for this project in determining whether a company was a start-up. However, an important limitation of FAME is that it does not have an API to work with, meaning that data is harder to access.

Fortunately, Companies House holds a record for all registered companies in the United Kingdom that is accessible through a free API (CompaniesHouse, 2021). As the official registry for incorporated firms in the United Kingdom, Companies House holds data such as incorporation

date, previous company names, and registered company address. Consequently, Companies House provided a rich set of firm data to match against data from research papers. Additionally, the registered company number found on Companies House can be used to search the FAME database. This expanded the available data that could be used to create the data set in this project.

2.2 Text Similarity

Determining the extent to which texts are similar was needed in this project for two reasons. Firstly, string data from research papers were matched against data on start-ups and other institutions. Secondly, the abstracts of scientific articles were compared. Fortunately, text similarity has been an active research area and techniques have been developed in areas such as information retrieval (Singhal et al., 2001), text classification (Sun and Lim, 2001), topic modelling (Blei, Ng and Jordan, 2003), and text summarisation (Barzilay, McKeown and Elhadad, 1999).

There are two main types of text similarity methods which have traditionally been applied: lexical and semantic. Given that the two approaches have specific use cases, advantages, and disadvantages, this section aims to understand the foundations of each approach. In addition to traditional techniques, this section also looks at recent developments in deep learning that have further developed semantic similarity techniques.

2.2.1 Lexical Similarity

At their core, lexical similarity measurements are finding an approximation for the match of a set of strings. As such, lexical similarity can often be thought of as string-based similarity metrics (Gomaa, Fahmy et al., 2013). There are two different groups of methods of lexical similarity: character-based and term-based. Character-based algorithms represent passages of text as sequences of characters. Smith-Waterman and Needleman-Wunsch, two character-based algorithms, have been particularly useful in bioinformatics due to the structured sequence of DNA (Sung, 2009). Other character-based measurements include Levenshtein and Jaro (Levenshtein et al., 1966; Jaro, 1989). Levenshtein uses insertion, deletion, and substitution of characters, whereas Jaro uses the order and number of common characters in two strings to calculate their similarity. Consequently, algorithms like these have been used to match records in databases (Winkler, 1994; Porter, Winkler et al., 1997). This made character-based lexical techniques good candidates for constructing the data set in this project.

In contrast to character-based metrics, term-based metrics represent passages of text as collections of words or vectors. Techniques such as Jaccard similarity work by using the word collection approach, calculating similarity as the intersection divided by the union of the two sets of words (Jaccard, 1912). Other techniques, such as Euclidean distance, typically rely on calculating the distance between vector representations. It is worth noting that Manhattan distance, cosine similarity, and Pearson correlation are just some examples that work similarly to Euclidean distance. Moreover, these techniques can often form the basis for comparing more sophisticated semantic techniques that are discussed later in Section 2.2.2. There is evidence suggesting that Euclidean distance may perform worse on text document clustering than other term-based techniques, therefore highlighting an important issue with lexical techniques (Huang et al., 2008). Namely, it is not clear which techniques will be most effective in a given

problem.

Although lexical similarity methods have achieved some success, an important limitation is that they are mostly unable to understand when two texts are similar in meaning but use different words. For example, consider the following phrases: *I own a dog* and *I have an animal* (Mihalcea et al., 2006). For a human, it is easy to understand that these phrases are very similar. However, lexical similarity methods often fail to capture this similarity. Despite this vital limitation, the interpretability and ease of implementation of lexical similarity methods have meant that they have been adopted by Arts, Hou and Gomez (2021) in the business literature. Before critically analysing this adoption, an understanding of semantic similarity methods must first be developed.

2.2.2 Semantic Similarity

Semantic similarity methods attempt to go beyond the idea of lexical similarity. Namely, "semantic relatedness refers to human judgments of the degree to which a given pair of concepts is related" (Pedersen et al., 2007). The previously mentioned phrases serve as an example of semantic similarity. The area of research concerned with helping computers *understand* the meaning of human language has developed some traditional techniques. These traditional techniques can be broken down into corpus and knowledge-based methods (Gomaa, Fahmy et al., 2013). However, recent advances in deep learning have led to a host of new techniques being developed that have also proven capable of capturing semantic meaning.

Latent semantic analysis (LSA), the most traditional corpus-based technique, starts by constructing a sparse occurrence matrix of terms within each document (Landauer and Dumais, 1997). It is worth noting that other corpus-based techniques, including many variants of LSA itself, start with constructing a similar matrix. In LSA, matrix construction is followed by singular value decomposition and a subsequent comparison of the resulting matrices. It has been shown that there are cases where LSA, as well as other corpus-based methods, perform worse than lexical similarity methods that use cosine similarity on term frequency-inverse document frequency representations (TF-IDF), another popular lexical method (Mihalcea et al., 2006). However, the same paper found that the corpus-based methods, on average, performed better than lexical methods. This highlights the importance of testing multiple techniques for any given application.

In contrast to corpus-based techniques, knowledge-based similarity methods estimate the degree of similarity between words using semantic networks. These semantic networks are graphical representations of knowledge. The most popular semantic network in NLP is WordNet, which groups nouns, verbs, adjectives, and adverbs into distinct concepts (Miller, 1995). In this sense, it can be thought of as a form of thesaurus with the addition of labelling the relationship between words in proximity to one another. For example, WordNet represents the relationship between *Barack Obama* and *president* as one being an instance of the other. An intuitive class of techniques naturally arises from such structured graphical representations. Specifically, the distance between words (vertices) in the semantic network can be computed. Leacock, Chodorow and Miller (1998) and Wu and Palmer (1994) were among the first researchers to propose methods relying on path distance. An alternative approach to using path distances is to take an information theoretic view. This approach takes advantage of the hierarchical structure of a network by calculating the information shared by two concepts as the information content of the concepts that come below them in the network (Resnik, 1995).

The previously discussed knowledge-based methods are domain-independent, relying on pre-determined semantic networks such as WordNet. This presents an issue when confronted with domains that contain a highly specialised vocabulary not covered by WordNet. Unfortunately, scientific articles relate to domains that suffer from this issue. As each scientific discipline discovers new phenomena and proposes theories, more highly specialised words are used to communicate effectively to the rest of the researchers in the field. SNOMED-CT® (2021) provides a domain specific network for the field of biomedicine, but there is a lack of similar works for other scientific disciplines. Interestingly, Pedersen et al. (2007) showed that corpus-based methods outperformed a variety of knowledge-based methods that utilised SNOMED-CT® (2021). This illustrates that domain dependent networks are not a necessity for capturing semantic meaning in difficult domains.

Although the research of Pedersen et al. (2007) provided interesting insights for the similarity of biomedical concepts, there were three important points that were considered in the context of this project. Firstly, the research of Pedersen et al. (2007) only focused on the similarity of two concepts, not two passages of text. Therefore, any conclusions reached cannot be generalised to the similarity of research article abstracts, as this project aimed to do. Secondly, findings in the field of biomedicine cannot be generalised to all scientific disciplines. Thirdly, it raises a major limitation of semantic similarity methods in general. Namely, there is no ground truth label of how semantically similar two texts are. In Pedersen et al. (2007), physicians and medical coders were used to rate the relatedness of pairs of medical concepts. With a correlation of only 0.51 between the ratings of the two groups, it is clear that the concept of semantic similarity is difficult to assess. Consequently, lexical similarity methods have an interpretability advantage over semantic similarity methods.

Embeddings

So far, this section has focused on the traditional methods used in lexical and semantic similarity text analysis. However, research in recent years has focused on learning word vectors, or word embeddings, for semantic analysis. One of the earliest advances in this domain was through the work of Bengio et al. (2003). In their paper, the authors proposed learning the meaning of words through neural networks, as opposed to co-occurrence matrices (as in LSA). The word vector representations are utilised to learn joint probability functions of word sequences in the training text. Consequently, it is possible to generalise to sentences which are unseen. This is because similar words are represented closely in vector space and therefore, combinations of word vectors are represented closely in sentence space.

Following the seminal work of Bengio et al. (2003), there have been further advances in word vector representations through global and local context approaches. GloVe (global vectors) by Stanford University illustrates the progress of global approaches which grew from methods like LSA (Pennington, Socher and Manning, 2014). More specifically, GloVe utilises a log-bilinear regression model which is trained only on the non-zero elements of the co-occurrence matrix of a corpus. Alternatively, approaches such as word2vec (Mikolov et al., 2013), ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), and the GPT series (Radford et al., 2018, 2019; Brown et al., 2020) learn word embeddings by trying to predict words within their local context. The latter three methods also take advantage of the advances in deep learning architectures to create large neural networks that produce state-of-the-art performance. One important limitation of such approaches, and indeed deep learning in general, is that they can be seen as a black box of operations. This further exacerbates the previously mentioned interpretability

issue associated with semantic similarity methods. Notably, there is an emerging body of research that seeks to understand how these black boxes come to understand natural language (Jawahar, Sagot and Seddah, 2019; Clark et al., 2019; Conneau et al., 2018b). Despite these efforts, interpretability is still an issue that is more prominent in semantic methods than in lexical ones.

As this project aimed to compare the abstracts of research articles published by start-ups and other institutional types, it was important to compare the semantic similarity of multiple passages of text with one other. Although BERT-based models currently dominate the leading edge on various NLP tasks, such as question answering, classification, and named entity recognition, it is important to acknowledge their limitations when comparing multiple texts (SuperGLUE Benchmark, 2021). More specifically, there are significant computational costs associated with BERT that limit its ability to make pairwise comparisons and cluster similar texts (Reimers and Gurevych, 2019).

One solution to this computational issue is to create additional embeddings whereby sentences, paragraphs or documents are represented in vector space, similar to word embeddings. A simple method would involve summing or averaging the relevant word embeddings that make up a given text. Later dubbed doc2vec, Le and Mikolov (2014) proposed an alternative method to this simple approach that involves training a paragraph vector which is unique to each passage of text, thereby capturing the context or topic of that text. Following the success of BERT-based models, researchers have also attempted to create sentence embeddings from BERT. This task proves difficult as BERT, unlike doc2vec, does not explicitly compute sentence embeddings. However, it is possible to pass a single sentence as input and derive a fixed sized vector by averaging outputs. In a similar manner to Le and Mikolov (2014), Reimers and Gurevych (2019) provided evidence that this somewhat naive approach produces poor embeddings and subsequently proposed an alternative. Their approach, Sentence-BERT, is among the many proposed methods for producing sentence embeddings (Kiros et al., 2015; Conneau et al., 2018a; Bowman et al., 2016).

One commonality between the recent advances in language models is that training is carried out on large general corpora, and then fine-tuning is performed on domain specific corpora. Notable exceptions for this project are Sci-BERT and SPECTER by AllenAI (Beltagy, Lo and Cohan, 2019; Cohan et al., 2020). Sci-BERT, based on BERT, is pre-trained on 1.14 million scientific papers from various fields. SPECTER is then designed to produce document level representations of scientific texts. Interestingly, SPECTER is initialised using Sci-BERT but is then trained using an objective function based on the citation graph of the papers. The authors claimed that this means the embeddings capture inter-document relatedness more than language models alone. As SPECTER is a recent contribution to the literature, it is important to be sceptical of its abilities. Although evaluation was conducted in the original paper, the methodology used was a new evaluation framework that the authors also proposed in the same paper.

This review of the literature has revealed an important theme underlying NLP research. Namely, moving from lexical to semantic, and then applying deep learning approaches, increasingly sacrifices the interpretability of comparisons. However, this loss of interpretability is exchanged for an increase in the performance in various NLP tasks. With this trade-off in mind, the recent application of NLP by Arts, Hou and Gomez (2021) to the business literature will now be critically discussed.

2.3 Applications

One recent noteworthy paper in the business literature that attempted to use text similarity techniques was by Arts, Hou and Gomez (2021), which built on the prior work of Arts, Cassiman and Gomez (2018). The reason why this paper is particularly interesting is that it was weighted towards lexical over semantic similarity techniques. The authors of the paper proposed the use of cosine similarity and N-grams to estimate the impact of technologies that have been patented by the United States Patent and Trademark Office. They argued that these similarity measurements reflect the impact of a technology by comparing its difference to prior work and similarity to future work. For N-grams, this means that an impactful patent introduced new sequences of words (uni-grams, bi-grams and tri-grams) that had not previously occurred in patents, yet were frequently used in proceeding patents. Cosine similarity between a given focal patent and all patents filed five years before and after the focal patent were calculated and averaged to produce a metric with the same logic as the N-grams approach. These methods successfully identified influential patents such as Google's patent for delivering, targeting, and measuring advertising over networks (US5948061) as highly impactful. However, there are important limitations that must be considered in the context of this project.

Firstly, there were a number of problems with the methods used in the paper. For example, the authors calculated an impact measure as the average forward cosine similarity divided by the average backward cosine similarity. Although it is reasonable to expect backward cosine to reflect the novelty of a patent, forward cosine should not be used to calculate this impact measurement. This is because all future patents that are similar to an impactful patent will be represented close to each other in vector space. Furthermore, cosine similarity may be confounded by similar writing styles. For instance, if a patent attorney wrote their first patent, their writing style may differ from prior patents. Moreover, if the patent attorney is successful and takes on more clients, their writing style will become more common in the pool of patents. In this example, the cosine similarity measurement would mistake a new writing style for an impactful patent. In addition to the limitations of cosine similarity, N-grams also suffers from technical problems. Namely, it does not capture the semantic meaning of words. This means that a patent could be building on the work of a very similar patent but use a different phraseology such that it is identified as novel.

One obvious solution to the issue of missing semantic meaning is to use one of the semantic similarity techniques discussed in Section 2.2.2. However, a key part of their approach was to separate the novelty of a technology (difference from prior work) from its impact (similarity to future work). Therefore, they argued that learning word and concept relationships (semantic similarity) from all patents would lose the separability aspect they relied on. This could be solved by training a model to learn semantic relationships between words on patents up to a certain patent. Consequently, the time separability aspect would be intact and not affect the analysis. Unfortunately, such an approach becomes computationally expensive as the number of documents increases. Therefore, there is once again a trade-off. Namely, the ability to *understand* natural language comes at the expense of interpretability and usability.

Secondly, there are problems that stem from the restriction of the data set that are more poignant in this project than in Arts, Hou and Gomez (2021). More specifically, the authors noted that it is possible that the real technological breakthrough of a patent came from a patent that was rejected or granted in another jurisdiction, or came from a scientific publication. Therefore, such lexical techniques may overestimate the impact of any given patent. In the context of this project, this limitation is even more pronounced due to the incomplete nature

of publication data sets. Unlike the data set used in Arts, Hou and Gomez (2021), the data set created in this project does not claim to cover all scientific knowledge. Specifically, it only spans ten years and is localised to the United Kingdom. Furthermore, the work of Arts, Hou and Gomez (2021) was fundamentally looking at something different to this project. The authors wanted to identify technological impact through NLP because citation data for patents are less formalised than in scientific papers. In this project, the impact of a paper can be identified through citation and journal metrics.

2.4 Summary

This chapter presented an overview of the data sources available for constructing the data set. It was noted that the open-access nature, as well as memberships though the University of Bath Library Services, led certain data sources to be favoured over others. Although these online data sources have spurred research looking at the relationship between academia and industry, the role of start-up companies in this space has been left unexplored. As a result, this project aimed to take advantage of the Scopus, Companies House, GRID, and FAME databases to fill this space.

This chapter also reported the findings of a comprehensive review of the literature relating to text similarity. Despite a lot of research being conducted in this topic, few applications have been attempted in the business and economic literature. Lexical techniques, such as the Jaro and Levenshtein algorithms, were identified as useful for matching records to create data sets. Other lexical, as well as semantic, techniques were identified as potential methods for comparing longer texts. Throughout the literature review, it was made clear that the best technique for a given use case is not predictable ex-ante. Therefore, experimentation using various methods was required in this project.

Chapter 3

Method: Data Set

Relating to the first and second objectives outlined in Section 1.3, the creation of the data set was an integral part of this project. This is because it was the first data set to link research papers to start-up companies, and it formed the basis for subsequent analyses. This chapter outlines the approached taken to combine four main data sources using lexical string matching. The additional metrics used to improve the analysis are also stated in Section 3.3.

3.1 Research Papers

The foundation of the data set was made up of research papers published by affiliations located in the United Kingdom between 2000 and 2009.[1] This data was collected from Scopus and resulted in an initial data set of 828,235 papers. Although the data were structured, tidying and transforming the data into a more useful format left 818,666 papers. This was primarily due to affiliations having large amounts of data missing that would prevent further analysis. Once the initial data were tidied, a unique set of affiliations in the United Kingdom that published research papers was established. This provided the basis for more data collection to categorise the affiliations by institutional type.

3.2 Institution Categorisation

To measure the research contribution of startup companies, there first needed to be a valid comparison against other institutional types, such as universities, government agencies and more established companies. Therefore, a procedure for categorising institutional types was developed. Additionally, categorising the affiliations as being a company allowed for a more robust identification of startups later on. Figure 3.1 presents a flow chart for the categorisation process, which used the following data sources:

1. Scopus: An online database for bibliometric data.
2. Global Research Identifier Database (GRID): A global database of 101,637 notable research institutions, providing established date, location and institutional type data.[2]

[1] The following query was used to collect data from the advanced search functionality on Scopus: 'DOCTYPE (ar) AND AFFIL (united AND kingdom) AND PUBYEAR = <year>'.

[2] Available for download at: `https://www.grid.ac/downloads`.

3. Companies House API: An API for retrieving a wide range of data on incorporated companies within the United Kingdom.
4. Postcodes: A database mapping UK postcodes to latitude and longitude coordinates.[3]

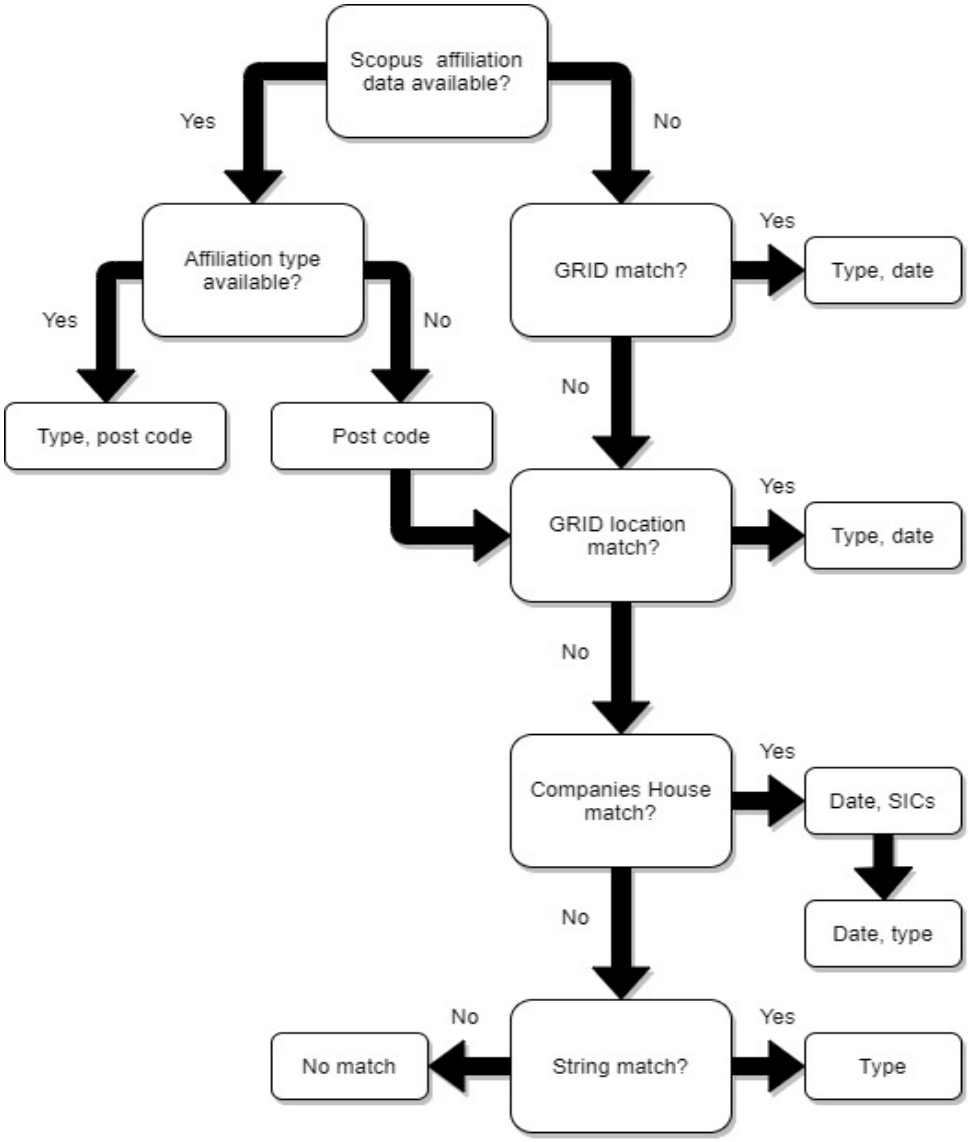

Figure 3.1: Flow chart of the institutional type decision-making process during data set creation. The smaller boxes represent a scenario where data was collected on an affiliation. Acronyms: Global Research Identifier Database, GRID; Standard Industrial Classification of Economic Activities, SIC.

Scopus and GRID

The decision-making process illustrated in Figure 3.1 was put in place to combat the fact that only a small percentage of Scopus affiliations are curated by Scopus. More specifically, Scopus only provided institutional type data for 2,928 out of the 48,060 affiliations in the data set. Therefore, the GRID was used to increase the number of affiliations which the institutional type was known. Matching the Scopus affiliation to the GRID entry was done through the

[3]Available for download at: `https://www.freemaptools.com/download-uk-postcode-lat-lng.htm`.

Python library RapidFuzz. RapidFuzz uses the Levenshtein algorithm discussed in Section 2.2.1 to score the similarity between a string and a list of candidate strings. A cut-off score of 91 out of 100 was chosen through visual inspection of the resulting matches to ensure the matches were sensible.

To increase the number of GRID matches, postcodes taken from the Scopus Affiliation Retrieval API were used to calculate the geodesic distance from GRID candidate matches. Geodesic distance, implemented by the GeoPy library, finds the shortest distance on the surface of an ellipsoidal model of the earth using latitude and longitude coordinates. To convert postcode data from Scopus to latitude and longitude coordinates, the postcode database was used. A cut-off score of 90 and distance of 2 km were again chosen through visual inspection to ensure the matches were sensible. After using the GRID, the number of affiliations matched increased by 3,286 to 6,214.

Companies House API

The Companies House API was used to search for fuzzy string matches for the remaining affiliations. Despite dealing with rate limits, the API offered the benefit of searching by current and previous company names. This search functionality was particularly beneficial because the data set created spanned from 2000-2009. Therefore, it was likely that some companies changed their names after publishing their research paper. Fuzzy matching was carried out on the current and previous names of the search results, and a cut-off score of 88 was chosen through visual inspection. The data collected from Companies House included: registered company number, incorporation date, and standard industrial classification of economic activities (SIC) codes. Registered company numbers and incorporation dates were used for identifying startups, whereas SIC codes were used to further categorise institutional types. Namely, companies with SIC codes relating to public administration and defence, education, and human health and social work activities were categorised as government, education, and health respectively. With the rest of the matches from Companies House categorised as companies, Companies House provided an additional 16,421 matches.

Standardisation and Heuristics

Standardisation of institutional types was carried out on the categories provided by Scopus and GRID. An additional 4,531 affiliations were categorised through simple string heuristics. For example, if the affiliation name contained "school", this was categorised as education.

3.2.1 Start-up Definition

Once a set of affiliations had been established, the companies were then used to create startup candidates based on their incorporation or established date relative to the publication year. Incorporation, established and publication dates were provided by Companies House, GRID and Scopus respectively. This approach followed from the literature using the year of appearance on the official registry of a country to define a startup (Klapper, Laeven and Rajan, 2006; Conti and Valentini, 2018). The number of years chosen was ten due to the range of industries captured within the data set. Although this may include companies in the growth stage, previous research looking at start-ups has found that findings were robust across a range of one to ten years for defining a start-up (Conti and Valentini, 2018).

The approach used in this project differed from the literature by also incorporating an indication of firm size. Although Companies House does provide financial data relating to companies, it is often recorded in PDF format and therefore not easily accessible. Fortunately, the data source FAME provides a batch search functionality whereby registered company numbers can be used to search for a wide range of data relating to company size. Consequently, the top 97th percentile in terms of 2009 revenue were rejected as startup candidates. This prevented subsidiaries of large corporate companies and newly merged firms being falsely identified as startups. For example, GlaxoSmithKline Plc and Oxoid Limited were correctly rejected as startup candidates despite their incorporation dates falling within the permissible window for startups.

To ensure the methodology used in this project correctly categorised start-up companies, two random samples were taken from the start-up (n=50) and non-start-up (n=50) affiliations. The number of false positives and false negatives were determined by manually investigating the website, LinkedIn profiles, and other online data for each affiliation. The results of these random samples are reported in Section 5.2.

3.3 Additional Metrics

To gain a deeper understanding of the contribution of startups to the scientific literature, additional data regarding research area and paper quality were also collected. The research area of a paper was determined at the journal level. Namely, the unique journal identifier of a paper was mapped to its corresponding All Science Journal Classification (ASJC) codes using data provided by Scopus. However, any given journal can have multiple ASJC codes associated with it. Therefore, the research area of each journal was disambiguated based on the research area of other journals where the same authors of the given journal had been publishing. Furthermore, journals without a corresponding ASJC code were assigned a research area based on the research area of similarly named journals.

To judge the quality of research articles, citation counts and journal quality metrics were utilised. Citation counts were mined from Scopus ten years after the publication date of each paper.[4] This allowed for a more valid analysis when comparing citation data as older research articles had more time to accumulate citations. Journal quality data provided alternative metrics for evaluating the quality of research articles, therefore acting as a form of robustness check on the findings. Journal quality metrics from Scopus included CiteScore and CiteScore percentile for a given research area. CiteScore can be seen as an alternative metric to the popular Journal Impact Factor, but it is based on the Scopus database. As CiteScore is a new open metric, there has been debate on its use over Journal Impact Factor (da Silva and Memon, 2017; Fernandez-Llimos, 2018). CiteScore was used in this project as it naturally aligned with the data set. CiteScore percentile calculates the CiteScore percentile of a journal within a given research area.

[4]The following query was used to collect citation data from the advanced search functionality on Scopus: 'REF(<eid>) AND PUBYEAR BEF <pub_year> + <num_years> + 1' where num_years was chosen as ten.

Chapter 4

Method: Semantic Similarity

In this chapter, the methods used for generating abstract embeddings are outlined. The challenge of assessing the quality of abstract embeddings is also discussed. A quantitative and qualitative evaluation method are proposed to assess the ability of the abstract embeddings to capture semantic meaning. Finally, the use of dimensionality reduction techniques as a means for visualising abstract embeddings is outlined.

4.1 Word2vec

To create abstract embeddings, the continuous bag of words (CBOW) version of the word2vec model proposed by Mikolov et al. (2013) was used. The model was implemented using the gensim Python library (Řehůřek and Sojka, 2010). The CBOW version was used instead of the skip-gram version of word2vec as the model trains much faster, therefore allowing for more experimentation with alternative models for abstract embedding generation. Unlike other algorithms, such as doc2vec, that explicitly create document vectors, word2vec requires averaging the word embeddings that make up a given document. Figure 4.1 displays the neural network architecture for how these word embeddings are created.

The word2vec model uses a window that slides over the sentences in a corpus to learn word vector representations. Figure 4.1 illustrates the case when window size equals five, meaning that the current word is being predicted using a context of four words: two preceding and two proceeding words. The context words are one-hot encoded and then transformed into a continuous vector space through a projection layer. The projection layer is simply a set of weights that does not involve an activation function and is shared across all inputs. The reason why this version of word2vec is named CBOW stems from the fact that the projections are aggregated by averaging. This means that the word order is removed from the representation of context words, hence the context is now a bag of words which is also represented continuously.

Training was carried out using an initial learning rate of 0.025 with linear decay that approaches zero at the end of the last training epoch. This choice was taken to mirror the hyperparameters of the original word2vec paper that produced good quality word embeddings (Mikolov et al., 2013). In instances where hyperparameters were not reported in the original paper, default values were used from the gensim library. The exceptions to this were the dimensionality of the word embeddings and the number of epochs. The dimensionality of word vectors was tuned using the quantitative evaluation method described in Section 4.3.1. During early

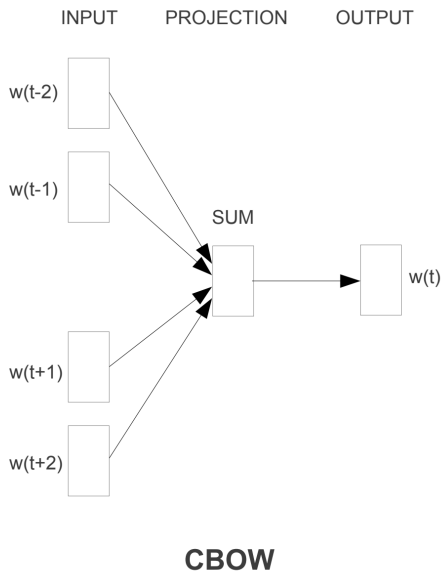

Figure 4.1: CBOW word2vec model architecture. Adapted from Mikolov et al. (2013). Acronym: Continuous Bag of Words, CBOW.

experimentation, it was noted that increasing the number of epochs to ten increased the quality of the abstract embeddings without compromising on computation speed.

4.2 Alternative Models

Although word2vec was found to be the best model for capturing the semantic meaning of abstracts, its performance was ranked against multiple other techniques. This followed the evidence found in Section 2.2.2 that suggested experimentation for each use case was vital as the best model is hard to predict ex-ante. Table 4.1 provides an outline of the alternative techniques used, as well as information on their training data and sources.

Technique	Document Specific	Trained	Pre-trained	Source
TF-IDF		✓		Scikit-learn
GloVe			Wikipedia 2014 + Gigaword 5	gensim
word2vec		✓	Google News	gensim
doc2vec	✓	✓		gensim
SPECTER	✓		Semantic Scholar	AllenAI

Table 4.1: Techniques used to generate abstract embeddings. Acronyms and contractions: Term Frequency-Inverse Document Frequency, TF-IDF; Global Vectors, GloVe.

TF-IDF, a lexical technique, was used to establish a baseline for abstract embedding quality. TF-IDF was chosen as a baseline, as the literature review in Section 2.2.1 found that lexical techniques struggle to understand semantic meaning. Next, two word embedding techniques were used: word2vec and GloVe. Notably, there was a pre-trained and trained version of word2vec. These techniques required averaging word embeddings to generate abstract level embeddings. Finally, two document level embedding techniques were used: doc2vec and SPECTER. Doc2vec was trained using the hyperparameters from Lau and Baldwin (2016) as these were shown to produce good embeddings. As mentioned is Section 2.2.2, SPECTER is

pre-trained on 1.14 million research papers and the corresponding citation graph. The title and abstract were fed into SPECTER to generate embeddings.

4.3 Embedding Evaluation

As with any machine learning task, it is important to ensure that the outcome of the task is of high quality. Given that the generation of abstract embeddings is an unsupervised learning task, traditional metrics such as mean squared error and cross entropy loss were not available to assess the quality of the embeddings. Therefore, to perform an assessment, a quantitative and qualitative method were used.

4.3.1 Quantitative Evaluation

Quantitative evaluation was carried out by utilising the research area of a paper to create 10,000 triplets of research papers. As noted in Section 3.3, any given paper can have multiple research areas associated with it. Therefore, to increase the validity of the triplet evaluation, each triplet consisted of two papers that had a common research area, as well as a paper that did not have any research areas in common with the other two. Next, the similarity between the abstract embeddings of the three research papers was calculated using cosine similarity, as shown in equation 4.1. Cosine similarity was used as it was identified in Section 2.2.1 as being a common way to compare sophisticated vector representations. The idea behind this evaluation method was that if the abstract embeddings were semantically meaningful, papers in the same research area would be more similar, as reflected in the cosine similarity measure.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{4.1}$$

4.3.2 Qualitative Evaluation

Qualitative evaluation of the best abstract embeddings was conducted by visual inspection of the resulting vectors. More specifically, t-distributed stochastic neighbour embedding (t-SNE) was used as a form of dimensionality reduction such that the abstract embeddings could be visualised on a Cartesian plane. Visual inspection involved highlighting two research areas with overlapping but also very distinct research topics. Economics and psychology were chosen a priori as they often share research topics. For example, the emerging field of behavioural economics aims to bring insights from psychology into the field of economics. Despite this overlap, each subject has distinct topics which the other subject does not involve itself, such as clinical psychology and economic history. Consequently, quality abstract embeddings should be able to display some level of overlap, as well as separation, between the two subjects. A hyperparameter sweep was performed for the perplexity of the t-SNE algorithm to find the value that best visualised the research areas.

Chapter 5

Results and Discussion: Data Set

5.1 Introduction

Section 3 outlined the methodology used to create the data set that achieved objectives one and two. The third objective was to evaluate the contribution of start-up research to science. The purpose of this section comes directly from objective three, as well as some further exploration of the data set. More specifically, the primary motivation for this section is to test the following null hypotheses:

H_1 There is no difference in the research area focus between papers by start-ups and papers by other types of institutions.

H_2 There is no difference in the quality of research papers between start-ups and other types of institutions.

The rest of this chapter is laid out in four parts. Firstly, Section 5.2 presents and discusses some summary statistics relating to the data set. Secondly, Section 5.3 represents the distribution of start-up papers in the United Kingdom geographically. Sections 5.2 and 5.3 can be seen as a form of exploratory data analysis of the data set. They are reported in this section of the project because a large part of the output and novelty of this project came from creating the data set. Therefore, they warrant significant discussion that naturally fits into the results and discussion section. Thirdly, Section 5.4 compares the research areas of start-up papers to other institutional types, as well as other companies. Finally, Section 5.5 compares the quality of research papers to other institutional types and other companies. The last two sections relate back to the two null hypotheses presented above. All statistical testing was carried out using Stata Statistical Software v16.[1] It should be noted that although the data set permits further analysis in multiple directions, such analyses have been omitted to stay within the provided time frame and scope of the project.

5.2 Summary Statistics

The primary challenge of creating the data set was to categorise affiliations from the Scopus database into different institutional types. The method used for creating the data set, as outlined in Section 3, resulted in a final data set of 1,044,261 affiliation-paper pairs. Start-up

[1] The .do files used for the statistical analysis are available to view on the GitHub repository for this project.

categorisation was validated. Two random samples of start-ups (n=50) and non-start-ups (n=50) identified ten false positives and two false negatives. This validated the approach used to categorise affiliations as start-ups.

A summary of the number of affiliations and subsequent papers by matching technique is provided in Appendix A.1. Summary statistics for each of these institutional categories are presented in Table 5.1.

	Number of papers	**Mean (s.d.) citations**	**Most common research areas**
Archive	6,322	28.60 (82.40)	AGRI > EART > MEDI
Company	45,162	32.52 (106.10)	MEDI > ENGI > CHEM
Education	704,207	37.52 (102.70)	MEDI > PHYS > BIOC
Government	19,092	39.07 (95.30)	MEDI > EART > AGRI
Healthcare	161,433	39.19 (112.90)	MEDI > BIOC > NURS
Non-profit	13,390	68.60 (280.80)	MEDI > BIOC > AGRI
Other	7,590	38.89 (71.60)	AGRI > MEDI > ENVI
Research Institute	76,621	54.94 (141.80)	MEDI > PHYS > BIOC
Start-up	10,444	34.34 (154.80)	MEDI > ENGI > BIOC

Table 5.1: Summary statistics of institutional types. For research area acronyms, see Appendix A.2.

From Table 5.1, it is clear that affiliations categorised as education produced the most research papers between 2000 and 2009. Affiliations categorised as start-ups published the third-fewest number of research papers during the same time frame, with only 'Archive' and 'Other' producing fewer papers. Non-profit and archive affiliations were the most and least cited institutional types, respectively. For seven of the nine institutional types, medicine was the most common research area. Interestingly, companies and start-ups had the same top two most common research areas (medicine and engineering) but differed with their third: chemistry versus biochemistry respectively.

Start-up papers received an average of 34.34 citations ten years after publication and a standard deviation of 154.80 citations. A large standard deviation suggested that start-up papers received a large range of citations. This is visually demonstrated by the histogram in Figure 5.1(a). With the mean (34.34) falling to the right of the median (12.00), Figure 5.1(a) shows that citation data for start-ups and other institutional types were positively skewed. Very high skewness values of 44.06 and 45.03 more formally confirmed the skewness of citation data for start-ups and other institutional types, respectively. The rest of Figure 5.1 is discussed in Section 5.5.

The results in Table 5.1 and Figures 5.1(a) and 5.1(b) summarise three interesting results that have not yet been shown in the literature. Firstly, it is clear that start-ups do contribute to the scientific literature in some form. This is reinforced by the fact that Frontier Science Limited, Field Genetics Limited, and Oxford Nanopore Technologies were all start-ups that published research papers receiving over 2,000 citations ten years after publication. Start-ups might be publishing scientific research for a number of reasons. Examples include: to send a signal to good candidates that the firm is a good place to work; to send signals to investors that the firm is a good investment; to form part of the "prior work" for future patents; to share knowledge to scientists about existing patents the firm already possesses.

Secondly, the most common research areas of start-up papers were medicine, engineering and

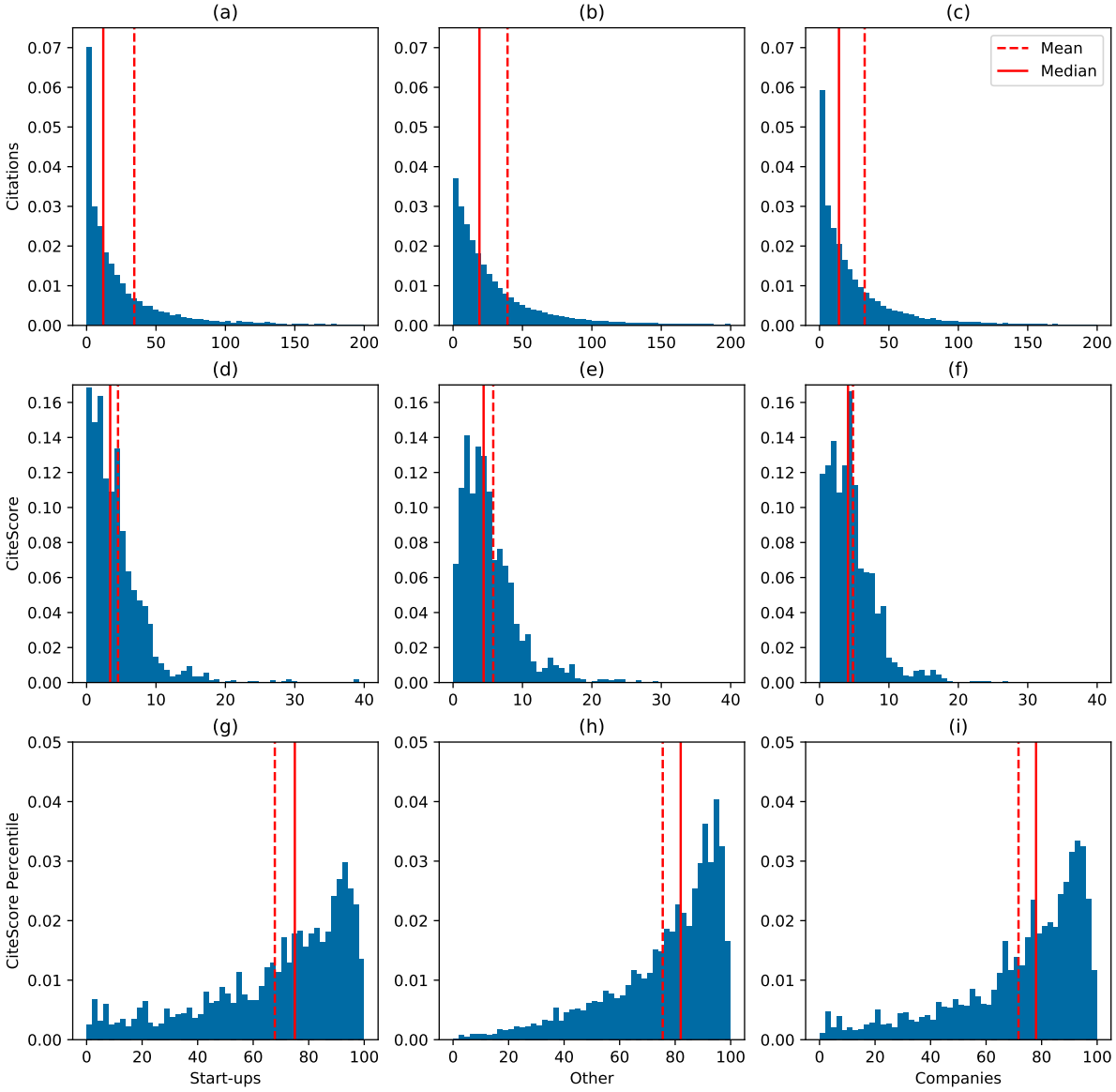

Figure 5.1: Histograms of three measures of paper quality for start-ups, other institutional types and companies.

biochemistry. The reason for the popularity of medicine and biochemistry can be explained by the size of the pharmaceutical and biotechnology markets. These markets were valued at $405.52 billion and $752.88 billion in 2020 and are expected to grow at 11.34% and 15.83% up to 2028, respectively (GVR, 2021a,b). This clearly indicates that there is money to be made in these markets. Given the highly scientific nature of the markets, start-ups stand to benefit from conducting scientific research in these areas. Engineering is also a research area that is very applied, meaning that start-ups can conduct research and expect to produce a product.

Thirdly, the distribution of citation data for start-ups and other institutional types were very similar, as seen in Figures 5.1(a) and 5.1(b). Although the distributions appeared to be similarly shaped, it is worth noting that start-ups had more papers closer to zero citations than other institutional types. This provides a preliminary indication that the quality of start-up papers were not as good as other institutional types. However, the quality of research papers is more

formally investigated in Section 5.5.

5.3 Geographic Distribution

Postcode data collected from Scopus and Companies House were utilised to visualise the distribution of start-up papers in the United Kingdom. The distribution was created using Tableau v2021.1.2 and is displayed in Figure 5.2.[2] Customised geocoding data were used to create postal areas for the United Kingdom (Ordnance Survey, 2021).

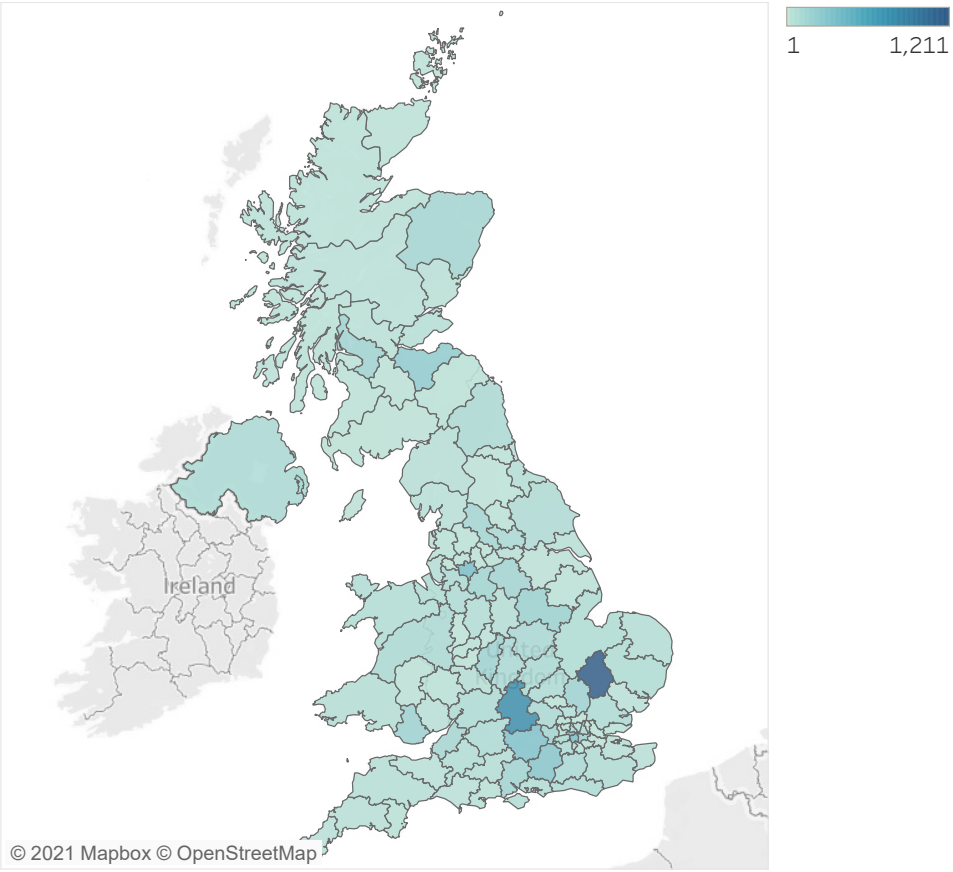

Figure 5.2: Geographic distribution of start-up papers in the United Kingdom between 2000 and 2009. Darker shades of blue correspond to more start-up papers being published in the postal area.

From Figure 5.2, it is immediately clear that Cambridgeshire and Oxfordshire were the two postal areas with the most start-up papers between 2000 and 2009 (1,211 and 797 respectively). However, the eight postal areas that make up London corresponded to 1,597 papers in total. Manchester and Edinburgh were the only two postal areas not in the South East that made it in to the top ten. Interestingly, this means that the highest concentration of start-up papers were centred around the golden triangle of universities (Oxford, Cambridge, and London) (Mullins, 2005). Furthermore, adding Manchester and Edinburgh universities to the golden

[2]The Tableau Public workbook allows interested readers to filter by research area to see the effect on the geographical distribution of start-up papers. This can be found on the GitHub repository for this project, which is linked in Section 7.3.

triangle completes the list of the top eight universities in the United Kingdom, according to the QS World University Rankings (QS, 2021).

The clustering of start-up papers around the elite universities of the United Kingdom suggests that there is some relationship between start-ups and universities. This result relates to two ideas from the economic literature: knowledge transfer and localisation. More specifically, it has been argued that the innovation performance of firms is more efficient when they are able to use knowledge generated from external sources. Furthermore, this acquisition or transfer of knowledge is more effective when universities are in proximity to firms (Smith, 2007). Therefore, the transfer of knowledge from universities to nearby start-ups explains why start-ups were clustered around the elite universities in this data set.

5.4 Research Area

An important aspect to understand about the research of start-up companies is what areas they are focusing their efforts on. Comparing the research focus of start-ups to other institutions will aid in the efficient allocation of resources to progress science and technology. Using the research area of each journal, as determined by the disambiguation technique described in Section 3.3, the research areas of start-up papers were compared to other institutional types. To perform a more granular comparison, companies that were not classified as start-ups were also compared to start-ups. Figure 5.3 visually illustrates the difference in research areas between start-ups, other institutional types, and companies.

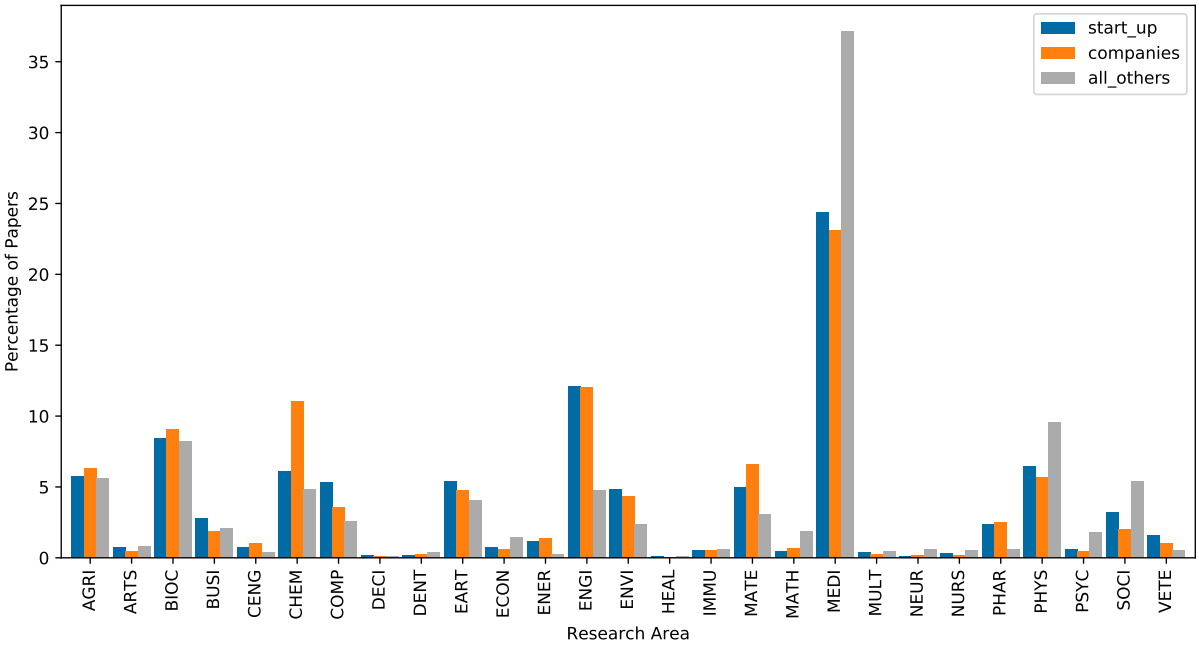

Figure 5.3: Research area as a percentage of total papers for start-ups, companies, and other institutional types. For research area acronyms, see Appendix A.2.

There are three clear results that appear from Figure 5.3. Firstly, the most common research area for start-ups, other institutional types, and companies was medicine (24.40%, 37.12%, and 23.10% respectively). Biochemistry was also a common research area for the three groups. Secondly, there were commonalities between start-ups and companies when compared to other institutional types. For example, medicine, physics, and the social sciences and humanities

represented a higher percentage of papers for other institutional types than start-ups and companies. However, a higher percentage of start-up and company papers were published in areas such as computer science, engineering, and material sciences than other institutional types. Thirdly, there were differences between start-ups and companies. Notably, chemistry papers made up a smaller percentage of start-up papers (6.07%) than company papers (11.04%), whereas computer science papers made up a larger percentage of start-up papers (5.33%) than company papers (3.54%).

The difference in focus on computer science research was found to be a consequence of the Microsoft Research Lab in Cambridge being categorised as a start-up. Debating the validity of this institution as a start-up is beyond the scope of this project. However, it does highlight the empirical methods for defining a start-up as an important area for future work. One explanation for start-ups producing a smaller percentage of chemistry papers is that there were fewer start-ups involved in chemistry in the first place. This could be because it is harder to create a product than in other research areas, or that larger companies have a competitive advantage over start-ups. An alternative explanation is that start-ups are involved in chemistry research, but they do not publish papers about their work. The semantic similarity analysis conducted in Section 6.3 provides evidence to support the latter explanation.

5.4.1 Statistical Analysis

Chi-squared tests were carried out to determine whether the research focus of start-ups differed from that of other institutional types and companies. The null hypothesis for the first test was that there was no association between being a start-up or another institutional type and the research area of papers. The null hypothesis of the second test only differed such that the research area of start-ups were only compared to companies. Both null hypotheses were rejected, indicating there was a difference in the research focus between start-ups and all other institutions, as well as between start-ups and companies (P=0.00 for both).

Figure 5.3 and the results of the chi-squared tests both provided evidence to reject null hypothesis one. This implies that the research focus of start-ups was significantly different to other institutional types. One potential explanation for this can be found in the economic literature for innovation. More specifically, it has been argued that there are two types of research: basic and applied. Basic research has the purpose of advancing our understanding of natural or other phenomena, whereas applied research finds a practical solution to a specific application (NSF, 1953). Therefore, the monetary incentives of a start-up naturally align with applied research, whereby a practical solution is produced and money can be made. This is evidenced by the fact that the three biggest research areas for start-ups were medicine, engineering, and biochemistry. All three of these areas have the potential to produce some form of product that can be commercialised. In contrast, other institutional types are more aligned with basic research due to the incentive structures. This is reflected in the fact that areas that are less related to commercial applications, such as mathematics, physics, and the social sciences and humanities, made up a larger percentage of the papers by other institutions than by start-ups.

5.5 Paper Quality

As described in Section 3.3, the quality of a research paper was assessed using two distinct methods: the number of citations ten years after publication and the quality of the journal the paper was published in. For each method, start-ups were compared to all other institutional types, and then separately to other companies. Looking into the quality of research conducted provides an indication of how much start-ups contribute towards scientific progress. Citation and journal quality data are presented in Figure 5.1.

Figure 5.1 shows that citations and CiteScore were positively skewed, whilst CiteScore percentile was negatively skewed. Skewness and kurtosis tests for normality confirmed that all distributions in Figure 5.1 were not normally distributed (P=0.00 for all). The distributions belonging to start-ups (Figures 5.1(a), 5.1(d), and 5.1(g)) were more concentrated towards the lower bound than the corresponding distributions of other institutional types (Figures 5.1(b), 5.1(e), and 5.1(h)). Since both paper quality metrics are defined such that larger values mean better quality, this further suggests that the quality of start-up papers was worse than other institutional types. The distributions of citations and CiteScore percentile for start-ups and companies were visually very similar. Figures 5.1(d) and 5.1(f) showed signs that start-up papers might be lower quality than companies, but were inconclusive. To more formally analyse the distributions of paper quality metrics, a statistical analysis was carried out.

5.5.1 Statistical Analysis

The results of statistical testing for citation and journal quality data are reported in Table 5.2. Further statistical testing at the research area level was also carried out. Due to the large quantity of tests conducted, the full results are reported in Appendix A.3. Mann-Whitney U tests were conducted for all statistical testing following the skewness and kurtosis tests for normality showing a violation of the normality assumption required for t-tests.

	Start-ups	Others		Companies	
		Mean	P-value	Mean	P-value
Citations	34.34	39.24	0.00	32.52	0.00
CiteScore	4.52	5.79	0.00	4.82	0.00
CiteScore Percentile	67.80	75.53	0.00	71.67	0.00

Table 5.2: Results of Mann-Whitney U tests performed on citations ten years after publication, 2011 CiteScore, and 2011 CiteScore percentile. P-values correspond to the statistical comparisons made against start-ups.

With a P-value of 0.00, Table 5.2 shows that start-up papers received significantly fewer citations than papers by other institutional types. However, results displayed in Appendix A.3 indicate that start-up papers in the area of medicine and physics obtained significantly more citations than papers by other institutional types (P=0.00 for both). The majority of the other research areas saw more citations by other institutional types than start-ups, as can be seen in Appendix A.3. Compared to other companies, start-up papers received significantly more citations (P=0.00). Research areas where start-ups received significantly more citations than companies included medicine, agriculture, and engineering, among others. Notably, of the research areas where start-ups received more citations than companies, there were only three areas where the difference was more than an average of five citations: computer science, decision sciences, and physics.

P-values of 0.00 for both CiteScore and CiteScore percentile indicate that start-up papers were published in journals of worse quality than papers by other institutional types. Diving deeper found that there were no research areas where CiteScore was significantly larger for start-ups than other institutional types. Dentistry was the only research area where CiteScore percentile was significantly larger for start-ups than other institutional types (P=0.07). Compared to other companies, the CiteScore and CiteScore percentile of start-up papers were significantly lower. There were five and three research areas where start-ups received significantly higher CiteScores and CiteScore percentiles than other companies. In contrast, there were nine and eleven research areas where start-ups received significantly lower CiteScores and CiteScore percentiles than other companies.

The results of the statistical testing of citation and journal quality data imply that, on average, start-ups produced lower quality papers than other institutional types. As grouped averages can hide nuances in the data set, it was broken down into research areas. However, this showed that start-up papers were not significantly better than the papers of other institutional types. Namely, physics and medicine, the two areas which reported significantly more citations for start-ups than other institutional types, did not correspond to significantly better quality journals for start-ups. This meant that there were no research areas where citations and journal quality metrics were better for start-ups than other institutional types. In contrast, there were 14 research areas where citations and CiteScores were significantly higher for other institutional types than start-ups. This indicates that the quality of start-up papers was the same or worse than papers by other institutional types, depending on the research area. Therefore, at a high level null hypothesis two may be rejected, but this is not always the case at a more granular level.

One possible reason why start-up papers were lower quality than papers by other institutional types relates to the incentive structures within the different institutions. More specifically, the careers of researchers at universities can depend on the quantity and quality of published research (Fanelli, 2010). This competitive culture has been popularised under the phrase "publish or perish" (Rond and Miller, 2005). This means that university researchers have strong incentives to produce research papers in high-impact factor journals that are also highly-cited. In contrast, the careers of researchers at start-ups depend more on the survival of the start-up than on bibliometric statistics. This difference in incentives means that some university researchers are likely to devote more time and effort into getting their research published in higher quality journals. It is worth noting that this "publish or perish" culture can also result in low quality research from academic institutions. This aspect of the culture is discussed later in Section 6.3.1.

Another interesting result was that, on average, start-up papers received more citations than companies but were published in lower quality journals. One possible explanation for this result stems from the publication process. Namely, it can cost up to $3,900 to publish a paper in a high-impact factor journal (Solomon and Björk, 2012). This means that it can be a financial burden for start-ups to publish in high quality journals. Furthermore, high-impact factor journals require higher standards of writing, meaning that there may be multiple revisions before publication, costing additional time and money for start-ups. Given that start-ups often have limited financial resources, this could be one reason why start-ups published in lower quality journals.

Chapter 6

Results and Discussion: Semantic Similarity

6.1 Introduction

This chapter details the results of the generation, evaluation, and visual exploration of abstract embeddings. Abstract embeddings were generated to achieve objective four outlined in Section 1.3. Namely, the aim of generating abstract embeddings was to provide further insights on the research differences of start-up companies. The rest of this chapter is laid out in three distinct phases. Firstly, the results of quantitative evaluation of multiple abstract embedding techniques are presented. Secondly, the results of dimensionality reduction and qualitative evaluation of the best embeddings are reported. Finally, the visual exploration of the semantic meaning of start-up papers relative to other institutional types is outlined and discussed.

6.2 Embedding Evaluation

6.2.1 Quantitative Evaluation

Quantitative evaluation, in the form of triplets of research papers, was conducted on the six abstract embeddings techniques outlined in Section 4.1. A hyperparameter sweep was performed on the dimensions of the embeddings for four of the six techniques used. The results of the evaluation are presented in Figure 6.1.

From Figure 6.1, it is clear that word2vec and TF-IDF produced the best and worst abstract embeddings, respectively. More specifically, word2vec with embedding dimensions of 500 and 1,000 achieved 75.81%, the highest accuracy out of all the techniques and hyperparameter combinations used. Interestingly, the two document-specific methods, doc2vec and SPECTER, performed worse than word2vec, a word-specific technique. Another interesting result was that the pre-trained word2vec model performed only 1.21 percentage points worse than the word2vec model trained specifically on this data set.

There are two possible reasons that can explain why word2vec outperformed doc2vec and SPECTER. Firstly, it is possible that word2vec was simply a superior model for capturing the semantic meaning in this specific data set, given the hyperparameter values tested. Note that this does not mean that word2vec would necessarily be a superior model after more rigorous

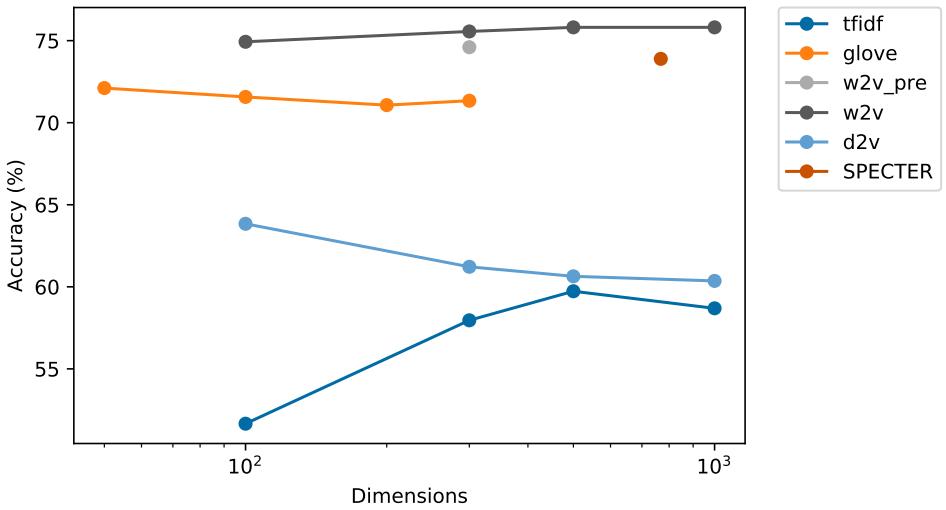

Figure 6.1: Triplet evaluation results of the different abstract embedding techniques. Acronyms and contractions: Term Frequency-Inverse Document Frequency, TF-IDF; Global Vectors, GloVe; pre-trained word2vec, w2v_pre; word2vec, w2v; doc2vec, d2v.

hyperparameter tuning. The fact that word2vec outperformed doc2vec coincides with the findings of Lau and Baldwin (2016). Specifically, they showed that the performance of doc2vec, relative to word2vec, increases as the length of documents increases. Therefore, the short abstracts used in this project can explain why word2vec outperformed doc2vec. The same logic can also explain why word2vec outperformed SPECTER.

Secondly, it is possible that the triplet evaluation technique used did not assess semantic similarity well enough. This explanation was identified early on as a threat to the validity of the quantitative evaluation. Consequently, triplets were created using all research areas associated with the paper, as opposed to the one research area determined by the disambiguation technique. This meant that there was more confidence that the unrelated paper was in fact unrelated, therefore increasing the ability of the evaluation method to assess semantic similarity. Although measures were taken to reduce the possibility of this explanation confounding the results, the nature of the abstract embeddings makes it hard to rule out entirely. Due to this, a qualitative evaluation method was also used.

6.2.2 Qualitative Evaluation

After word2vec was identified as the model most capable of capturing semantic meaning, dimensionality reduction and a qualitative evaluation method was carried out. Principal component analysis (PCA) and t-SNE were used to reduce the dimensions of the abstract embeddings. Qualitative evaluation, in the form of visual inspection of economics and psychology papers, was also carried out. The results of this experimentation and visualisation are displayed in Figure 6.2.

Figure 6.2 shows that the abstract embedding representations of psychology and economics papers were clustered together for all dimensionality reduction methods. For t-SNE with perplexity equal to ten and PCA, papers were sparsely distributed in the two-dimensional space. As the perplexity of the t-SNE algorithm increased, the papers were represented more densely around the origin. Interestingly, Figure 6.2 also shows the varying degrees of separation and

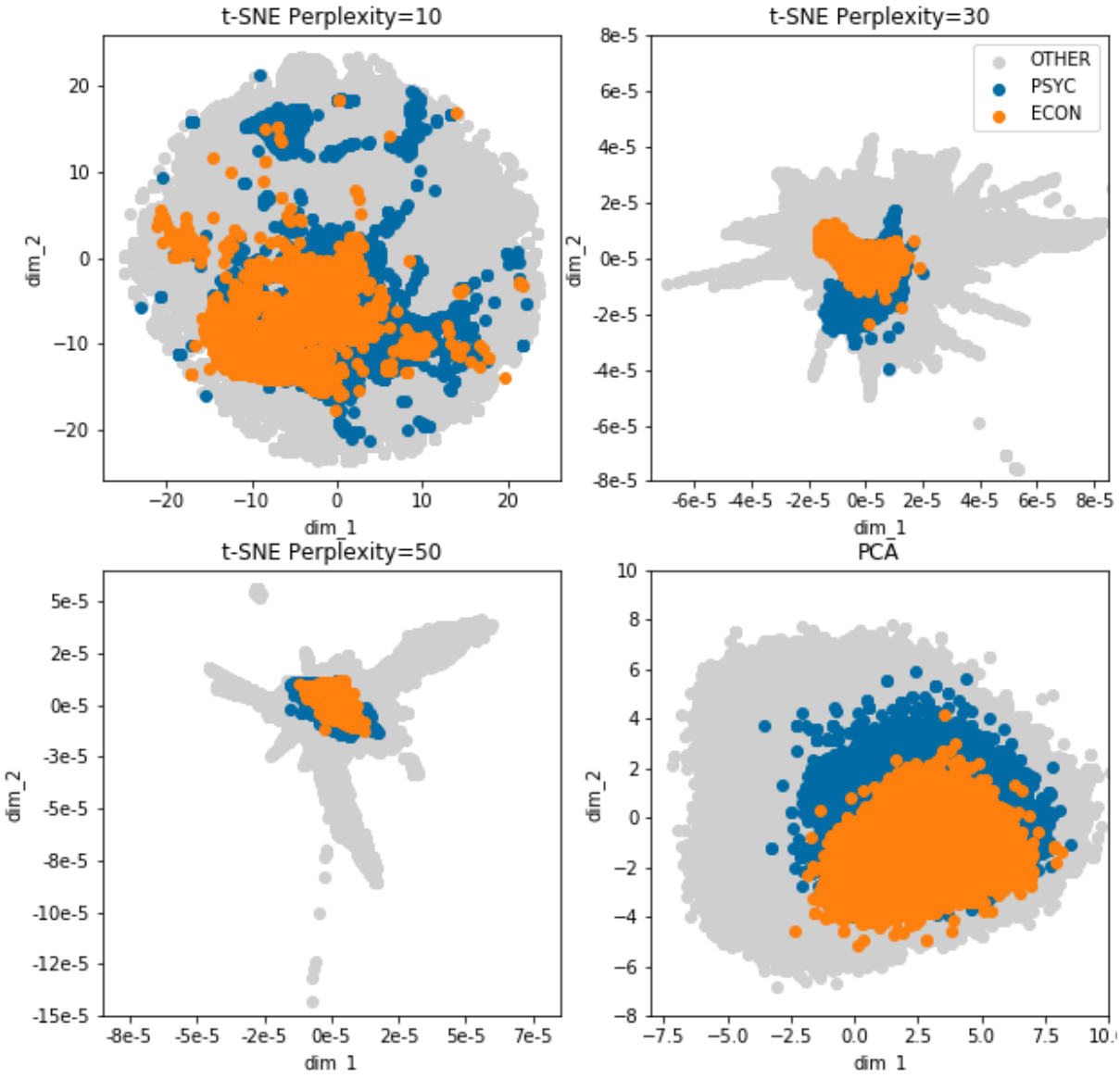

Figure 6.2: Dimensionality reduction experimentation. Abstract embeddings were generated from the word2vec model with 500 dimensional vectors. Acronyms and contractions: t-distributed Stochastic Neighbour Embedding, t-SNE; Principal Component Analysis, PCA; dimension, dim; psychology, psyc; economics, econ.

overlap between psychology and economics papers that the dimensionality reduction techniques were able to achieve. For example, the t-SNE algorithm with perplexity equal to 10 and 50 exhibited a high degree of overlap but little separation. The t-SNE algorithm with perplexity equal to 30 showed separate areas of only psychology and only economics papers, as well as areas of overlap. Although the PCA visualisation was able to show distinct regions for psychology papers, this was not the case for economics. Therefore, the t-SNE algorithm with perplexity equal to 30 was chosen as the best method for visualising the abstract embeddings.

The results from Figure 6.2 suggest that the abstract embeddings generated by word2vec were able to capture the semantic meaning of papers. This is because all four methods represented the papers from psychology and economics, two related fields, closely in two-dimensional space. The fact that increasing perplexity resulted in more densely distributed data stems

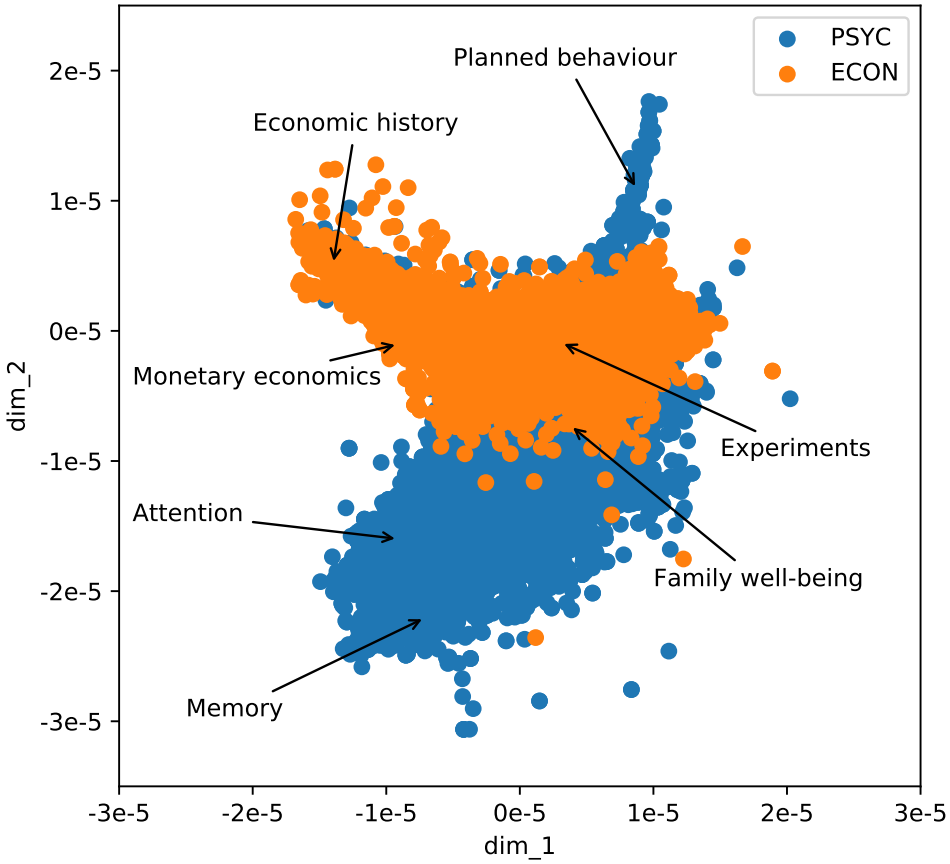

Figure 6.3: Abstract embedding representations of psychology and economic papers. Visualisation was produced using the t-SNE algorithm with perplexity equal to 30. Annotations indicate regions where the majority of papers corresponded to one research topic. Acronyms: psychology, psyc; economics, econ; dimensions, dim.

from perplexity being roughly interpreted as the number of neighbours close to each point (Van der Maaten and Hinton, 2008). Therefore, a perplexity of 30 was able to densely cluster similar papers and avoid clustering all papers, unlike perplexity values of 10 and 50 respectively. To further examine the validity of the two-dimensional representation of papers, Figure 6.3 displays a close up of psychology and economics papers, along with labels identifying subfields for each research area.

Figure 6.3 illustrates how word2vec and t-SNE were able to successfully represent the overlapping and separate nature of psychology and economics papers. Firstly, psychology had subfields relating to memory, attention and planned behaviour that were easily distinguished from economics papers. Secondly, economics had subfields relating to monetary economics and economic history that were easily distinguished from psychology papers. It is worth noting that there were some psychology papers in proximity to the economic history papers. However, they were not within the tightly packed cluster of economic history papers. Thirdly, there were regions where the research topic was similar for psychology and economics. For example, there was a research cluster relating to child and family well-being located near the lower boundary of economics papers, as well as an experiments-based research cluster near the centre of the economics papers.

The results presented in Sections 6.2.1 and 6.2.2 provide evidence that the abstract embeddings

generated by word2vec captured the semantic meaning of the research papers. Therefore, the abstract embeddings produced were deemed suitable for analysing the semantic similarity of papers published by start-ups and other institutional types.

6.3 Semantic Similarity

This section starts off by highlighting the areas in semantic space where start-ups did not publish, relative to other institutional types. Sections 6.3.1 and 6.3.2 then dive deeper into the data by looking at two of the most common research areas for start-ups to provide two reasons to explain the absence of start-up papers: a lack of monetisation and protecting intellectual property. The embeddings created by word2vec, which were subsequently reduced to two dimensions using t-SNE, are displayed in Figure 6.4. The figure distinguishes between papers written by start-ups and other institutional types to aid comparison. The annotations in Figure 6.4 indicate regions of the graph where start-up papers were absent and the majority of papers corresponded to one or two research areas.

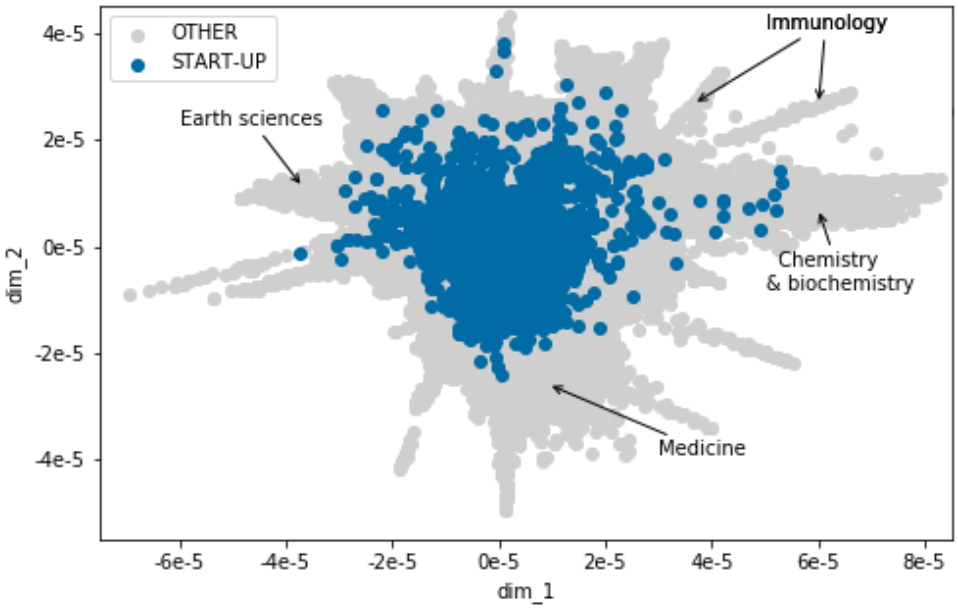

Figure 6.4: Abstract embeddings of papers by start-ups and other institutional types represented in a two-dimensional space. Annotations indicate regions where start-up papers were absent and the majority of papers corresponded to one or two research areas. Embeddings were generated using word2vec and t-SNE. Acronyms and contractions: dimensions, dim; t-distributed Stochastic Neighbour Embedding, t-SNE.

Figure 6.4 shows that start-up papers were more concentrated around the origin, whilst other institutional types also spread out into off-shoots. There were four distinct areas of semantic space where start-up papers were absent, which also corresponded to papers from one or two research areas. Firstly, the two immunology areas in the northeastern region of the graph were related to the discovery of new bacteria and fungi in plants, soils, and animals. Secondly, there was a large area of papers in chemistry and biochemistry where start-ups did not publish. Thirdly, there was a large area of medicine papers in the southern region of the graph. The research topics of these papers related to pregnancy and osteoporosis. Finally, there was an

area of papers in the earth sciences in the western region of the graph relating to seismology and plate tectonics, where there was an absence of start-up papers.

6.3.1 Monetisation

The first reason for the absence of start-up papers in a semantic area is due to the lack of monetisation of the underlying research. For example, Figure 6.4 showed an absence of start-up papers in an area relating to seismology and tectonic plates. It is plausible that research in these fields does not easily translate to a product. Therefore, start-ups tend not to do research in this semantic area. In this example, the research topic has a direct effect on the abstract embeddings and the monetisation of the research. However, the lack of monetisation can also have a direct effect on the abstract embeddings through phraseological differences.

As previously mentioned, the markets for biotechnology and pharmaceuticals are very large. Therefore, the lack of monetisation should not be a problem for immunology papers. However, Figure 6.4 shows two branches of immunology papers where there was an absence of start-up papers. These branches consisted of papers characterising new bacteria and fungi that did not correspond to a practical application. If the bacteria or fungi discovered had corresponded to a practical application, the phraseology of the abstract would have been very different in order to emphasise the application. Consequently, the resulting embeddings would have also been represented differently in semantic space. In other words, the lack of monetisation had a direct effect on the representation of the abstract. An example paper in these branches was entitled: "Streptococcus marimammalium sp. nov., isolated from seals" (Lawson et al., 2005). This paper characterised a previously unknown bacteria that was found in seals, but had no direct link to a practical application. In contrast, immunology papers not found in the two branches tended to have a practical application.[1]

It could be argued that these immunology papers fall victim to the "publish or perish" culture mentioned earlier. More specifically, academic immunologists may need to publish a paper regardless of whether they have found an application for their findings. This idea is reinforced by the fact that these immunology papers received fewer citations (16.98) than the average immunology paper (36.23). They were also published in lower quality journals (CiteScore: 3.69 and 5.90 respectively). In contrast, start-up researchers do not have the same career-driven incentives to publish papers on findings that do not translate to an application or monetisation. Therefore, despite the potential monetisation of research in immunology, there were no start-up papers in these two off-shoots because they do not directly translate to practical applications. An alternative interpretation of these results is that academic immunologists publish more basic research than start-up immunologists, and that basic research receives less attention than applied research. This is due to basic research having unknown practical applications.

Further evidence for the monetisation theory comes from visual inspection of research papers in engineering. Figure 6.5 shows the abstract embeddings of start-ups, companies, and other institutional types for engineering. Three areas where there was an absence of start-up papers have been labelled with the research topics that were prominent in those areas.

[1] An example start-up paper that was not in the two branches was "Analysis of equid herpesvirus 1 strain variation reveals a point mutation of the DNA polymerase strongly associated with neuropathogenic versus nonneuropathogenic disease outbreaks" (Nugent et al., 2006). An example company paper that was not in the two branches was "In vivo characterization of Lactobacillus johnsonii FI9785 for use as a defined competitive exclusion agent against bacterial pathogens in poultry" (La Ragione et al., 2004). Note how both these papers have a practical application and are not characterising a novel organism without a clear purpose.

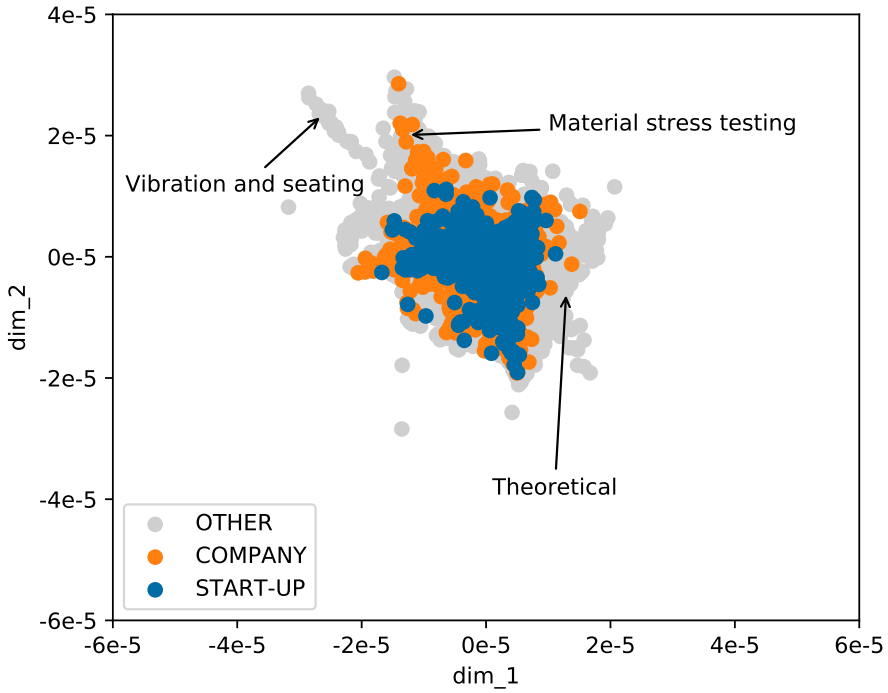

Figure 6.5: Abstract embeddings of engineering papers by start-ups and other institutional types. Embeddings were generated using word2vec and t-SNE. Annotations indicate regions where there was an absence of start-up papers and the majority of papers corresponded to one research topic. Acronym: dimensions, dim; t-distributed Stochastic Neighbour Embedding, t-SNE.

Notably, Figure 6.5 shows a distinct lack of start-up papers on the right-hand side of the semantic space. These papers were primarily focused on theoretical or mathematically heavy research. An example paper in this semantic space was entitled: "Factorization on a Riemann surface in scattering theory" (Antipov and Silvestrov, 2002). From the title and the abstract, it is not clear how this research relates to a practical application. The fact that there is also a distinct lack of company papers in this semantic space further suggests that a lack of monetisation or practical application can be one reason for the absence of start-up papers.

6.3.2 Intellectual Property

The second reason for the absence of start-up publications is the protection of intellectual property. Namely, start-ups might be working on a research topic, but they choose to not publish their findings in academic journals as they do not have a form of intellectual property rights (e.g. a patent) yet. To investigate this claim, Figure 6.6 displays biochemistry papers from start-ups, companies, and other institutional types. Again, the figure has been annotated to indicate to research topics where start-ups did not publish.

The largest area of semantic space in Figure 6.6 where there was an absence of start-up papers relates to structural chemistry. This area corresponds to the whole branch on the right-hand side of the semantic space. The figure shows that there were company papers in this area, but they were few in numbers. The semantic space relating to databases and research tools also

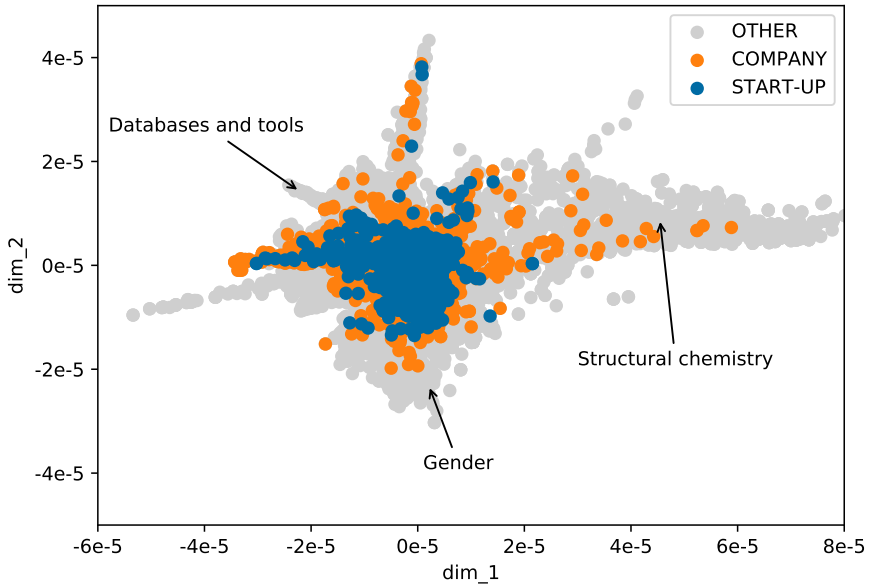

Figure 6.6: Abstract embeddings of biochemistry papers by start-ups and other institutional types. Embeddings were generated using word2vec and t-SNE. Annotations indicate regions where there was an absence of start-up papers and the majority of papers corresponded to one research topic. Acronyms and contractions: dimensions, dim; t-distributed Stochastic Neighbour Embedding, t-SNE.

lacked the presence of start-up and company papers. The semantic space annotated "Gender" corresponded to papers researching sex and mating.

The lack of start-up papers in the semantic space of databases and tools, and structural chemistry, can be explained by the need to protect intellectual property. Successful research in both of these areas could yield a competitive advantage for a start-up. Therefore, the start-up would not want to publish their research before ensuring they have intellectual property rights for their discovery. For example, Transgenomic Limited filed multiple patents before 2003 for their research on liquid chromatography and mutation analysis (JUSTIA, 2021). Then, in 2003 they published some of their research in the academic literature (Bayat et al., 2003). Notably, they did not detail the structure of their compounds in the academic literature before ensuring their research successfully related to an application, and that they had a patent for it. It is also worth noting that the research that was published after the patent did not fall into the right-hand side of Figure 6.6 due to the phraseology of the research emphasising the application.

The fact that start-ups have strong disincentives to publish research without intellectual property rights can be explained by the characteristics of the intellectual property market. Namely, it has been previously argued that numerous low quality patents have been granted due to the insufficient resources and capabilities of patent examiners to successfully evaluate candidate patents against prior art (Sampat, 2010).[2] This means that if a start-up were to publish a paper detailing the structural and biological properties of a compound, it is

[2]Prior art refers to the information from patents, research papers, and other sources that is available to the public before the origin date of a patent.

possible that a larger company could be granted a patent for this same compound. Given that start-ups often have limited resources, it is likely that the larger company would be able to financially overpower the start-up and settle in court for a relatively small amount (Graham and Sichelman, 2008). Consequently, start-ups have strong disincentives to publish structural chemistry research without having intellectual property rights.

Chapter 7

Conclusion

The advent of online bibliometric data has started to allow social scientist to understand the role that large companies play in the progression of science. However, it is unclear what role start-up companies play in this process. This project was the first attempt to shed light on this research topic.

This was done by creating a novel data set linking research papers to start-up companies in the United Kingdom. The data set spanned from 2000 to 2009 and also categorised associated affiliations into distinct institutional types to aid the comparison. The creation of the baseline data set utilised lexical similarity techniques for string matching and took advantage of four main data sources: Scopus, the Global Research Identifier Database, Companies House, and the Financial Analysis Made Easy database.

The baseline data set was then expanded to find that the quality of start-up papers was worse than other institutional types. The quality of papers was assessed using citation and journal metrics. Using additional data from Scopus, the research area of each paper was used to investigate this finding at a more granular level. The results showed that the quality of start-up papers was worse than or equal to the quality of papers by other institutional types, depending on the research area. Notably, there were no research areas where start-up papers were of higher quality than papers by other institutional types.

Using geographic data from Scopus and Companies House, a geographic distribution of start-up papers was visualised. This visualisation was beyond the initial scope of the project. However, it was included in the report as it provided some interesting insights. Namely, the high concentration of start-up papers around the golden triangle of universities - Cambridge, Oxford, and London - indicated that there were localised knowledge spillovers from universities to start-ups. Furthermore, the additional data collected from Scopus allowed this visualisation to be filtered by research area. Although this has been omitted from the report, it is available to view on Tableau Public. The link for this in provided in Section 7.3.

Finally, abstracts were represented in a semantic space to highlight semantic differences between papers by start-ups and other institutional types. With the highest accuracy in the quantitative evaluation task, word2vec was chosen as the model that produced the best abstract embeddings. After t-SNE was used to reduce the dimensionality of the embeddings, qualitative evaluation further confirmed the ability of this model to capture semantic meaning. With the semantic embeddings represented in a Cartesian plane, two ideas were proposed to explain the absence of start-up papers in areas of semantic space: a lack of monetisation of

the research and the protection of intellectual property.

7.1 Limitations

This section reflects on the limitations of the methods used during this project. The two main areas of limitations were affiliation categorisation and similarity comparison. Using the knowledge gained from the project, recommendations for alternative methodologies have also been provided where appropriate.

7.1.1 Affiliation Categorisation

The first step taken to identify start-up companies involved using lexical similarity techniques to match data from Scopus to Companies House. One issue with this approach was that some data provided by Scopus were uninformative and led to incorrect categorisations. Consequently, there were more false positives than false negatives for start-ups in the data set. However, there are two main ways that the methods used in this project can be altered to combat these false positives. Firstly, more filters can be applied to Scopus affiliations to ensure good data quality. Secondly, the final heuristic string matching used in the institution categorisation process, as shown in Figure 3.1, could be carried out before Companies House matching. This would result in fewer false positives for companies. Given that start-ups were defined as a subset of the companies, this would also reduce the number of false positives for start-ups.

There were also a number of design choices that were made due to the restricted time frame. However, if time had permitted, there were certain techniques that could have been used to increase the quality of the data set. Firstly, disambiguation of Scopus affiliations could have been carried out. This could be done through simple lexical string matching or a form active learning. This would decrease the number of duplicate affiliations that came from the Scopus database. As a result, the number of companies and start-ups publishing research would be a valid area for analysis. The data set in its current form does not permit this kind of analysis due to duplicate affiliations.

Secondly, Companies House and FAME provide rich sources of data to produce more accurate definitions of start-ups. This project followed the literature by using the date of appearance on the official registry, but also utilised metrics for company size. Companies House and FAME provide data on company founders, company ownership, and SIC codes. If time had permitted, this project would have utilised these data to produce more accurate definitions of start-ups.

7.1.2 Similarity Comparison

The similarity comparison conducted in Section 6.3 has two potential limitations: poor quality embeddings and limited scope. Poor quality embeddings imply that semantic meaning is not captured within the vectors. In this project, poor quality embeddings would have resulted in invalid conclusions being drawn from the visual inspection of semantic space. Although two forms of evaluation were carried out to ensure quality embeddings were generated, the lack of ground truth labels mean that it is hard to assess quality. Therefore, it is still possible that the embeddings were not capturing semantic meaning. To increase confidence that the embeddings generated were of good quality, further evaluation could have been conducted. For example, Cohan et al. (2020) proposed an evaluation framework for scientific document embeddings which includes citation prediction, classification and recommendation.

It should also be noted that the analysis conducted in this project was based only on the abstracts of papers. This means that a lot of information from within the full text was not captured by the embeddings. Unfortunately, obtaining full access to research papers dramatically reduces the size of the data set.

Again, if time had permitted, there is a host of other natural language processing techniques that could have been used to evaluate start-up papers. For example, persuasion and emotion detection are techniques used in the literature that could provide interesting insights (Young et al. (2011); Shivhare and Khethawat (2012)). These techniques could be used to understand more about the differences in writing style between start-ups and other types of institutions.

7.2 Future Work

As this project explored a novel research topic, there are multiple directions in which future work can be directed. To aid this area of research, potential directions have been split into short-term and medium-term.

7.2.1 Short-Term

Future work that can be carried out immediately involves conducting more analyses with the current data set or with simple extensions. One short-term direction is to analyse how the publishing performance of start-ups and companies changes throughout the life cycle of the firm. This can provide an indication of how much firms rely on research at different stages. Importantly, this can guide governments to provide funding at the optimal time for firms in their life cycle. To conduct this type of analysis, disambiguation of Scopus affiliations would be required. Another simple extension to the data set could be to look at how papers are received outside the scientific literature. This can be done using the PlumX Metrics API available through Scopus. This API provides a variety of data, such as policy citations, social media usage and article mentions, to indicate the impact of scientific research. Using a variety of metrics will provide a more holistic view of the impact of start-up research relative to other institutions.

As mentioned in Section 7.1.2, there are a variety of other linguistic analysis tools that can shed light on the difference between start-up and other research papers. Future work can use the current data set to understand the linguistic properties, such as persuasion, emotion, and bias, of the research that start-ups carry out.

7.2.2 Medium-Term

Future work that can be carried out in the medium-term either involve the creation of a new data set or substantial additions to the current one. This project provided a first report on start-up companies and their contribution to scientific research in the United Kingdom. Future work can look at creating a similar data set for other countries. However, a significant barrier to this type of research is the availability of a national registry, like Companies House in the United Kingdom. For example, the United States deals with the incorporation of companies at a state level. Therefore, the focus for this direction of future work will be data consolidation.

An important area of future work that requires substantial extension of the current data set relates to the prediction of start-up success. More specifically, additional data from Companies

House and FAME can be used to create a panel data set of start-up companies in the United Kingdom. Then, the role of publishing scientific papers can be assessed in the success of start-up companies. It is worth noting that affiliation disambiguation will also have to be carried out as it is common for companies to change names over time. This could be a significant contribution to the literature predicting start-up success and result in efficiency gains for the venture capitalism market.

A third direction for future work involves diving deeper into how the different types of start-ups contribute to the scientific literature. For example, data on company founders from Companies House and FAME can help identify university spin-offs. Large company spin-offs and charitable start-ups can also be identified using these data.

Finally, the most important direction for future work relates to a more detailed investigation of the transition from scientific knowledge to technological progress. Similar to Ahmadpoor and Jones (2017), this will involve creating a data set linking scientific papers with patents. The contribution in this direction will come from understanding the role that start-ups play in this dual frontier. However, access to patent data often comes with a significant fee. Therefore, funding would be required in order to pursue this direction of future work.

7.3 Source Code

The code for this project, including the statistical testing, is available on GitHub at the following URL: `https://github.com/aidan-o-brien/Dissertation`. Links for the final data set and Tableau Public workbook are also available on the README file of the GitHub repository.

Bibliography

Ackermann, S., 2012. *Are small firms important? their role and impact*. Springer Science & Business Media.

Aghion, P., Dewatripont, M. and Stein, J.C., 2008. Academic freedom, private-sector focus, and the process of innovation. *The rand journal of economics* [Online], 39(3), pp.617–635. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-2171.2008.00031.x`.

Ahmadpoor, M. and Jones, B.F., 2017. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351), pp.583–587.

Antipov, Y. and Silvestrov, V., 2002. Factorization on a riemann surface in scattering theory. *Quarterly journal of mechanics and applied mathematics*, 55(4), pp.607–654.

Arora, A., Belenzon, S. and Sheer, L., 2021. Knowledge spillovers and corporate investment in scientific research. *American economic review*, 111(3), pp.871–98.

Arts, S., Cassiman, B. and Gomez, J.C., 2018. Text matching to measure patent similarity. *Strategic management journal* [Online], 39(1), pp.62–84. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2699`.

Arts, S., Hou, J. and Gomez, J.C., 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research policy*, 50(2), p.104144.

Azoulay, P., Zivin, J.S.G. and Sampat, B.N., 2012. *2. the diffusion of scientific knowledge across time and space*. University of Chicago Press.

Barzilay, R., McKeown, K. and Elhadad, M., 1999. Information fusion in the context of multi-document summarization. *Proceedings of the 37th annual meeting of the association for computational linguistics*. pp.550–557.

Bayat, A., Walter, J., Lamb, H., Ferguson, M. and Ollier, W., 2003. Rapid denaturing high-performance liquid chromatography (dhplc) for mutation scanning of the transforming growth factor β3 gene using a novel proof-reading polymerase. *European journal of immunogenetics*, 30(5), pp.335–340.

Beltagy, I., Lo, K. and Cohan, A., 2019. Scibert: A pretrained language model for scientific text. *arxiv preprint arxiv:1903.10676*.

Bengio, Y., Ducharme, R., Vincent, P. and Janvin, C., 2003. A neural probabilistic language model. *The journal of machine learning research*, 3, pp.1137–1155.

Blank, S., 2010. What's a startup? first principles. Available from: `https://steveblank.com/2010/01/25/whats-a-startup-first-principles/` [Accessed 2021-05-02].

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *the journal of machine learning research*, 3, pp.993–1022.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R. and Bengio, S., 2016. Generating sentences from a continuous space. `1511.06349`.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al., 2020. Language models are few-shot learners. *arxiv preprint arxiv:2005.14165*.

Carreyrou, J., 2018. *Bad blood: Secrets and lies at a silicon valley startup*. USA: Penguin Random House.

Castle, J.C., Kreiter, S., Diekmann, J., Löwer, M., Roemer, N. Van de, Graaf, J. de, Selmi, A., Diken, M., Boegel, S., Paret, C. et al., 2012. Exploiting the mutanome for tumor vaccination. *Cancer research*, 72(5), pp.1081–1091.

CBInsights, 2021. Tech insights, vc database competitive intelligence platform. Available from: `https://www.cbinsights.com/` [Accessed 2021-05-02].

Clark, K., Khandelwal, U., Levy, O. and Manning, C.D., 2019. What does bert look at? an analysis of bert's attention. *arxiv preprint arxiv:1906.04341*.

Cohan, A., Feldman, S., Beltagy, I., Downey, D. and Weld, D.S., 2020. Specter: Document-level representation learning using citation-informed transformers. `2004.07180`.

CompaniesHouse, 2021. Companies house api oveview. Available from: `https://developer.company-information.service.gov.uk/` [Accessed 2021-05-02].

Conneau, A., Kiela, D., Schwenk, H., Barrault, L. and Bordes, A., 2018a. Supervised learning of universal sentence representations from natural language inference data. `1705.02364`.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L. and Baroni, M., 2018b. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. `1805.01070`.

Conti, A., Thursby, J. and Thursby, M., 2013. Patents as signals for startup financing. *The journal of industrial economics*, 61(3), pp.592–622.

Conti, R. and Valentini, G., 2018. Super partes? assessing the effect of judicial independence on entry. *Management science*, 64(8), pp.3517–3535.

Crick, D. and Spence, M., 2005. The internationalisation of 'high performing'uk high-tech smes: a study of planned and unplanned strategies. *International business review*, 14(2), pp.167–185.

Cristea, I.A., Cahan, E.M. and Ioannidis, J.P., 2019. Stealth research: lack of peer-reviewed evidence from healthcare unicorns. *Eur j clin invest*, 49(4), p.e13072.

CrunchBase, 2021. Discover innovative companies and the people behind them. Available from: `https://www.crunchbase.com/` [Accessed 2021-05-02].

Dalle, J.M., Besten, M. den and Menon, C., 2017. *Using Crunchbase for economic and managerial research* [Online]. (Oecd science, technology and industry working papers 2017/08). OECD Publishing. Available from: https://doi.org/10.1787/6c418d60-en.

Decker, R., Haltiwanger, J., Jarmin, R. and Miranda, J., 2014. The role of entrepreneurship in us job creation and economic dynamism. *The journal of economic perspectives* [Online], 28. Available from: https://doi.org/10.1257/jep.28.3.3.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arxiv preprint arxiv:1810.04805*.

Dijk, B.V., 2021. Financial analysis made easy. Available from: https://www.bvdinfo.com/en-gb/our-products/data/national/fame [Accessed 2021-08-30].

Fanelli, D., 2010. Do pressures to publish increase scientists' bias? an empirical support from us states data. *Plos one*, 5(4), p.e10271.

Fernandez-Llimos, F., 2018. Differences and similarities between journal impact factor and citescore. *Pharmacy practice (granada)*, 16(2).

Gartner, W., Starr, J. and Bhat, S., 1999. Predicting new venture survival: An analysis of "anatomy of a start-up." cases from inc. magazine. *Journal of business venturing*, 14(2), pp.215–232.

Gomaa, W.H., Fahmy, A.A. et al., 2013. A survey of text similarity approaches. *International journal of computer applications*, 68(13), pp.13–18.

Graham, S.J. and Sichelman, T., 2008. Why do start-ups patent. *Berkeley tech. lj*, 23, p.1063.

GVR, 2021a. Biotechnology market size report, 2021-2028. Available from: https://www.grandviewresearch.com/industry-analysis/biotechnology-market [Accessed 2021-08-15].

GVR, 2021b. Pharmaceutical manufacturing market size report, 2021-2028. Available from: https://www.grandviewresearch.com/industry-analysis/pharmaceutical-manufacturing-market [Accessed 2021-08-15].

Harzing, A.W., 2019. Two new kids on the block: How do crossref and dimensions compare with google scholar, microsoft academic, scopus and the web of science? *Scientometrics* [Online], 120. Available from: https://doi.org/10.1007/s11192-019-03114-y.

Huang, A. et al., 2008. Similarity measures for text document clustering. *Proceedings of the sixth new zealand computer science research student conference (nzcsrsc2008), christchurch, new zealand*. vol. 4, pp.9–56.

Hyytinen, A., Pajarinen, M. and Rouvinen, P., 2015. Does innovativeness reduce startup survival rates? *Journal of business venturing*, 30(4), pp.564–581.

Ioannidis, J.P., 2015. Stealth research: is biomedical innovation happening outside the peer-reviewed literature? *Jama*, 313(7), pp.663–664.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2), pp.37–50.

Jaro, M.A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the american statistical association*, 84(406), pp.414–420.

Jawahar, G., Sagot, B. and Seddah, D., 2019. What does bert learn about the structure of language? *Acl 2019-57th annual meeting of the association for computational linguistics*.

JUSTIA, 2021. Patents assigned to transgenomic, inc. [Online]. Available from: `https://patents.justia.com/assignee/transgenomic-inc?page=4` [Accessed 2021-08-28].

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R. and Fidler, S., 2015. Skip-thought vectors. `1506.06726`.

Klapper, L., Laeven, L. and Rajan, R., 2006. Entry regulation as a barrier to entrepreneurship. *Journal of financial economics*, 82(3), pp.591–629.

La Ragione, R., Narbad, A., Gasson, M. and Woodward, M.J., 2004. In vivo characterization of lactobacillus johnsonii fi9785 for use as a defined competitive exclusion agent against bacterial pathogens in poultry. *Letters in applied microbiology*, 38(3), pp.197–205.

Landauer, T.K. and Dumais, S.T., 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), p.211.

Lau, J.H. and Baldwin, T., 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arxiv preprint arxiv:1607.05368*.

Lawson, P.A., Foster, G., Falsen, E. and Collins, M.D., 2005. Streptococcus marimammalium sp. nov., isolated from seals. *International journal of systematic and evolutionary microbiology*, 55(1), pp.271–274.

Le, Q.V. and Mikolov, T., 2014. Distributed representations of sentences and documents. `1405.4053`.

Leacock, C., Chodorow, M. and Miller, G.A., 1998. Using corpus statistics and wordnet relations for sense identification. *Computational linguistics*, 24(1), pp.147–165.

Ledford, H., Cyranoski, D. and Van Noorden, R., 2020. The uk has approved a covid vaccine-here's what scientists now want to know. *Nature*, 588(7837), pp.205–206.

Levenshtein, V.I. et al., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*. Soviet Union, vol. 10, pp.707–710.

Maaten, L. Van der and Hinton, G., 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Marx, M. and Fuegi, A., 2020. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic management journal* [Online], 41(9), pp.1572–1594. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.3145`, Available from: `https://doi.org/https://doi.org/10.1002/smj.3145`.

Mihalcea, R., Corley, C., Strapparava, C. et al., 2006. Corpus-based and knowledge-based measures of text semantic similarity. *Aaai*. vol. 6, pp.775–780.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arxiv preprint arxiv:1301.3781*.

Miller, G.A., 1995. Wordnet: a lexical database for english. *Communications of the acm*, 38(11), pp.39–41.

Mullins, J., 2005. England's golden triangle. *New scientist*, (2496).

NSF, 1953. The third annual report of the national science foundation [Online]. Available from: https://www.nsf.gov/pubs/1953/annualreports/start.htm [Accessed 2021-08-18].

Nugent, J., Birch-Machin, I., Smith, K., Mumford, J., Swann, Z., Newton, J., Bowden, R., Allen, G. and Davis-Poynter, N., 2006. Analysis of equid herpesvirus 1 strain variation reveals a point mutation of the dna polymerase strongly associated with neuropathogenic versus nonneuropathogenic disease outbreaks. *Journal of virology*, 80(8), pp.4047–4060.

Oecd, 2021. Available from: https://stats.oecd.org/Index.aspx?DataSetCode=VC_INVEST [Accessed 2021-05-02].

Ordnance survey, 2021. Available from: https://osdatahub.os.uk/downloads/open.

Pedersen, T., Pakhomov, S.V., Patwardhan, S. and Chute, C.G., 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3), pp.288–299.

Pennington, J., Socher, R. and Manning, C.D., 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*. pp.1532–1543.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations. *arxiv preprint arxiv:1802.05365*.

Porter, E.H., Winkler, W.E. et al., 1997. Approximate string comparison and its effect on an advanced record linkage system. *Advanced record linkage system. us bureau of the census, research report*. Citeseer.

QS, 2021. Qs world university rankings 2022 [Online]. Available from: https://www.topuniversities.com/university-rankings/world-university-rankings/2022 [Accessed 2021-08-17].

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *Openai blog*, 1(8), p.9.

Řehůřek, R. and Sojka, P., 2010. EnglishSoftware Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp.45–50. http://is.muni.cz/publication/884893/en.

Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 1908.10084.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. *arxiv preprint cmp-lg/9511007*.

Rond, M.D. and Miller, A.N., 2005. Publish or perish: Bane or boon of academic life? *Journal of management inquiry* [Online], 14(4), pp.321–329. https://doi.org/10.1177/1056492605276850, Available from: https://doi.org/10.1177/1056492605276850.

Sampat, B., 2010. When do applicants search for prior art? *The journal of law and*

economics [Online], 53(2), pp.399–416. https://doi.org/10.1086/651959, Available from: https://doi.org/10.1086/651959.

Scopus, 2021. What is scopus preview? [Online]. Available from: https://service.elsevier.com/app/answers/detail/a_id/15534/supporthub/scopus/#tips [Accessed 2021-08-28].

SEC, 2020. Biontech se [Online]. Available from: https://www.sec.gov/Archives/edgar/data/1776985/000119312519241112/d635330df1.htm [Accessed 2021-05-02].

Shivhare, S.N. and Khethawat, S., 2012. Emotion detection from text. 1205.4944.

Shultz, M., 2007. Comparing test searches in pubmed and google scholar. *Journal of the medical library association : Jmla* [Online], 95, pp.442–5. Available from: https://doi.org/10.3163/1536-5050.95.4.442.

Silva, J.A.T. da and Memon, A.R., 2017. Citescore: A cite for sore eyes, or a valuable, transparent metric? *Scientometrics*, 111(1), pp.553–556.

Singhal, A. et al., 2001. Modern information retrieval: A brief overview. *Ieee data eng. bull.*, 24(4), pp.35–43.

Smith, H.L., 2007. Universities, innovation, and territorial development: a review of the evidence. *Environment and planning c: Government and policy*, 25(1), pp.98–114.

Solomon, D.J. and Björk, B.C., 2012. A study of open access journals using article processing charges. *Journal of the american society for information science and technology* [Online], 63(8), pp.1485–1495. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22673.

S&P, 2021. Sp global market intelligence [Online]. Available from: https://www.spglobal.com/marketintelligence/en/?product=compustat-research-insight [Accessed 2021-05-02].

Sun, A. and Lim, E.P., 2001. Hierarchical text classification and evaluation. *Proceedings 2001 ieee international conference on data mining*. IEEE, pp.521–528.

Sung, W.K., 2009. *Algorithms in bioinformatics: A practical introduction*. CRC Press.

Superglue benchmark, 2021. Available from: https://super.gluebenchmark.com/leaderboard/ [Accessed 2021-05-02].

Tata, A., Martinez, D.L., Garcia, D., Oesch, A. and Brusoni, S., 2017. The psycholinguistics of entrepreneurship. *Journal of business venturing insights* [Online], 7, pp.38–44. Available from: https://doi.org/https://doi.org/10.1016/j.jbvi.2017.02.001.

SNOMED-CT®, 2021. Why snomed ct? [Online]. Available from: https://www.snomed.org/snomed-ct/why-snomed-ct [Accessed 2021-05-02].

Visser, M., Eck, N.J. van and Waltman, L., 2021. Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic. 2005.10732.

Winkler, W.E., 1994. Advanced methods for record linkage.

Wu, Z. and Palmer, M., 1994. Verb semantics and lexical selection. *arxiv preprint cmp-lg/9406033*.

Young, J., Martell, C., Anand, P., Ortiz, P., Gilbert IV, H.T. et al., 2011. A microtext corpus for persuasion detection in dialog. *Workshops at the twenty-fifth aaai conference on artificial intelligence*.

Appendix A

A.1 Institution Categorisation Summary

Technique	No. Affiliations	No. Paper-Affiliation Pairs
Scopus	1401	160896
GRID exact	2195	707134
GRID fuzzy	1844	138828
GRID location	294	712
Companies House	11787	18008
Heuristics	5037	8239
FAME	5116	10444

Table A.1: Institution categorisation summary.

A.2 Research Area Acronyms

Acronym	Research Area
AGRI	Agriculture
ARTS	Arts
BIOC	Biochemistry
BUSI	Business
CENG	Chemical Engineering
CHEM	Chemistry
COMP	Computer Science
DECI	Decision Science
DENT	Dentistry
EART	Earth Sciences
ECON	Economics
ENER	Energy
ENGI	Engineering
ENVI	Environmental Sciences
HEAL	Health
IMMU	Immunology
MATE	Material Sciences
MATH	Mathematics
MEDI	Medicine
MULT	Multidisciplinary
NEUR	Neurology
NURS	Nursing
PHAR	Pharmacy
PHYS	Physics
PSYC	Psychology
SOCI	Social Sciences and Humanities
VETE	Veterinary

Table A.2: Definitions of research area acronyms.

A.3 Statistical Testing Results

	Start-ups			Others			Companies			Start-ups Versus Others			Start-ups Versus Companies		
	Cites	CS	CS percentile	Cites	CS	CS percentile	Cites	CS	CS percentile	Cites	CS	CS percentile	Cites	CS	CS percentile
AGRI	28.13	3.10	69.33	31.34	4.41	78.07	23.22	3.42	75.04	0.00	0.00	0.00	0.00	0.00	0.00
All	34.34	4.52	67.80	39.24	5.79	75.53	32.52	4.82	71.67	0.00	0.00	0.00	0.00	0.00	0.00
ARTS	8.64	1.30	71.40	11.12	1.02	68.63	7.68	1.00	69.21	0.72	0.11	0.22	0.06	0.08	0.36
BIOC	48.91	7.72	74.62	52.72	8.52	78.87	50.61	6.90	74.98	0.02	0.00	0.00	0.56	0.00	0.02
BUSI	14.67	1.80	59.70	30.22	2.48	69.38	15.47	1.77	57.07	0.00	0.00	0.00	0.24	0.50	0.20
CENG	18.94	3.08	67.20	20.97	3.41	75.29	18.95	3.21	71.14	0.00	0.09	0.14	0.11	0.54	0.80
CHEM	30.26	5.50	74.18	32.53	6.45	78.42	31.59	5.92	77.81	0.01	0.00	0.00	0.01	0.00	0.00
COMP	41.01	3.78	62.15	27.33	3.50	63.60	24.33	3.24	57.95	0.10	0.45	0.72	0.00	0.01	0.01
DECI	23.82	1.92	59.46	15.86	2.16	60.16	9.33	2.14	61.43	0.14	0.34	0.35	0.03	0.14	0.20
DENT	22.93	3.34	81.43	19.55	2.94	75.17	22.93	4.31	87.34	0.19	0.25	0.07	0.68	0.12	0.80
EART	19.39	2.65	62.77	36.26	4.37	79.90	18.10	2.93	68.32	0.00	0.00	0.00	0.43	0.00	0.00
ECON	13.85	1.75	57.78	24.12	2.09	63.39	16.97	1.67	57.11	0.00	0.01	0.04	0.81	0.69	0.72
ENER	8.12	2.49	57.76	24.91	3.64	74.91	7.67	1.59	49.06	0.00	0.00	0.00	0.21	0.01	0.01
ENGI	13.05	1.96	57.13	20.63	2.73	73.51	12.80	2.06	61.46	0.00	0.00	0.00	0.00	0.01	0.00
ENVI	22.34	3.47	66.67	36.77	4.55	78.55	27.08	3.98	72.25	0.00	0.00	0.00	0.01	0.00	0.00
HEAL	5.63	1.40	47.86	10.45	1.49	55.96	4.00	1.33	51.83	0.18	0.70	0.56	0.56	0.82	0.94
IMMU	42.12	5.91	75.35	36.18	5.90	73.43	37.04	4.90	71.42	0.96	0.99	0.64	0.89	0.03	0.04
MATE	18.06	3.20	62.64	28.94	4.83	80.17	22.51	3.89	71.90	0.00	0.00	0.00	0.00	0.00	0.00
MATH	10.63	1.63	53.07	15.28	1.96	61.48	14.43	2.06	60.64	0.38	0.32	0.01	0.91	0.09	0.05
MEDI	56.10	6.84	73.06	49.78	7.45	75.62	54.65	7.24	77.17	0.00	0.00	0.00	0.00	0.00	0.00
MULT	90.57	15.18	93.68	107.75	15.41	93.63	129.73	15.75	93.95	0.20	0.73	0.73	0.04	0.55	0.45
NEUR	140.14	5.06	62.50	51.34	7.00	78.75	46.02	6.03	71.41	0.46	0.09	0.03	0.37	0.48	0.27
NURS	15.18	1.90	67.97	16.45	2.08	70.98	13.20	1.79	63.47	0.58	0.39	0.33	0.90	0.60	0.63
PHAR	21.36	3.54	61.26	25.87	4.45	72.03	27.60	4.71	73.72	0.00	0.00	0.00	0.00	0.00	0.00
PHYS	47.41	4.44	71.29	35.81	6.18	80.87	28.44	4.22	70.19	0.00	0.00	0.00	0.05	0.02	0.19
PSYC	37.19	2.49	60.04	36.83	3.57	67.80	24.94	2.81	61.08	0.33	0.00	0.01	0.34	0.31	0.75
SOCI	13.24	1.52	62.85	21.66	1.73	68.59	13.44	1.43	61.01	0.00	0.00	0.00	0.17	0.13	0.20
VETE	19.90	1.91	70.20	19.51	2.20	74.42	18.20	1.94	69.63	0.50	0.00	0.00	0.62	0.90	0.78

Table A.3: Mann-Whitney U statistical testing results for research area.

A.4 Ethics Documentation

This form must be attached to the dissertation as an appendix.

Department of Computer Science

12-Point Ethics Checklist for UG and MSc Projects

Student Aidan O'Brien

Academic Year or Project Title 2020/1

Supervisor Tom Fincham-Haines

Does your project involve people for the collection of data other than you and your supervisor(s)? ~~YES~~ / NO

If the answer to the previous question is YES, you need to answer the following questions, otherwise you can ignore them.

This document describes the 12 issues that need to be considered carefully before students or staff involve other people ('participants' or 'volunteers') for the collection of information as part of their project or research. Replace the text beneath each question with a statement of how you address the issue in your project.

1. *Will you prepare a Participant Information Sheet for volunteers?* YES / NO
   This means telling someone enough in advance so that they can understand what is involved and why – it is what makes informed consent informed.

2. *Will the participants be informed that they could withdraw at any time?* YES / NO
   All participants have the right to withdraw at any time during the investigation, and to withdraw their data up to the point at which it is anonymised. They should be told this in the briefing script.

3. *Will there be any intentional deception of the participants?* YES / NO
   Withholding information or misleading participants is unacceptable if participants are likely to object or show unease when debriefed.

4. *Will participants be de-briefed?* YES / NO
   The investigator must provide the participants with sufficient information in the debriefing to enable them to understand the nature

of the investigation. This phase might wait until after the study is completed where this is necessary to protect the integrity of the study.

5. *Will participants voluntarily give informed consent?* YES / NO

Participants MUST consent before taking part in the study, informed by the briefing sheet. Participants should give their consent explicitly and in a form that is persistent –e.g. signing a form or sending an email. Signed consent forms should be kept by the supervisor after the study is complete. If your data collection is entirely anonymous and does not include collection of personal data you do not need to collect a signature. Instead, you should include a checkbox, which must be checked by the participant to indicate that informed consent has been given.

6. *Will the participants be exposed to any risks greater than those encountered in their normal work life (e.g., through the use of non-standard equipment)?* YES / NO

Investigators have a responsibility to protect participants from physical and mental harm during the investigation. The risk of harm must be no greater than in ordinary life.

7. *Will you be offering any incentive to the participants?* YES / NO

The payment of participants must not be used to induce them to risk harm beyond that which they risk without payment in their normal lifestyle.

8. *Will you be in a position of authority or influence over any of your participants?* YES / NO

A position of authority or influence over any participant must not be allowed to pressurise participants to take part in, or remain in, any experiment.

9. *Will any of your participants be under the age of 16?* YES / NO

Parental consent is required for participants under the age of 16.

10. *Will any of your participants have an impairment that will limit Their understanding or communication?* YES / NO

Additional consent is required for participants with impairments.

11. *Will the participants be informed of your contact details?* YES / NO

All participants must be able to contact the investigator after the investigation. They should be given the details of the Supervisor as part of the debriefing.

12. *Will you have a data management plan for all recorded data?* YES / NO

Personal data is anything which could be used to identify a person, or which can be related to an identifiable person. All personal data (hard copy and/or soft copy) should be anonymized (with the exception of consent forms) and stored securely on university servers (not the cloud).