UNIVERSITY OF
LINCOLN

Use of Word Embeddings in a Literature-Based Discovery System

Toby Stephen Reed
REE13487582

School of Computer Science

College of Science

University of Lincoln

Submitted in partial satisfaction of the requirements for the
Degree of Master of Science by Research
in Computer Science

Supervisor Dr. Vassilis Cutsuridis

July 2020

Acknowledgements

I would firstly like to thank my supervisor Dr Vassilis Cutsuridis for not only his invaluable expertise and direction throughout this project but also his motivation and friendship.

I would also like to thank my colleagues and friends Arthur, Ian, Aiden, Fabio and Bashar for the many trips to the shops, teas, coffees and (sometimes relevant) conversations we shared throughout the last year. My colleagues at Wren Kitchens also get the warmest of thankyous whilst they have only been a part of this journey towards the end they have encouraged and motivated as the finish line has approached.

There is also a special thank you to all my other colleagues from the School of Computer Science including but not limited to Matthew Beddows, Sean and Arthur Jones for their friendship and willingness to help in whatever way they can.

Thank you to all my other friends, family and everyone else who had a part to play in this journey. Whilst I do not have the space to thank everyone individually your input has not been forgotten and I truly could not have done any of this or got where I have without you.

Finally, all the work I have put in over this period of research is dedicated to one man, Pete, my grandfather. Thankyou, for everything you have done for me.

Thankyou.

Abstract

Since Don R. Swanson's first works in the field of Literature-Based Discovery (LBD) in the 1980s, there has been a keen interest in the process's abilities to retrieve new relationships from already published articles. In addition to this, the explosion of biomedical literature added to the public domain daily makes these automated systems more vital as time goes on with a researchers ability to keep up to date with their specialism let alone any potentially related fields. Furthermore, this emergence of LBD and the explosion of published knowledge has come at a time where the pharmaceutical industry is beginning to understand the importance of repurposing existing compounds as a method of reducing costs whilst still managing new and old conditions. This thesis proposes a system that utilises the Word2Vec group of models to implement an LBD system. These tasks are undertaken utilising seven different corpora comprised of biomedical articles related to varying levels from Raynaud Disease to Hematological journals published on the MEDLINE database and retrieved through PUBMED. This data was then fed through a specially developed pre-processing pipeline to normalise the data and then passed through a Word2Vec model and using the cosine similarity metric the most semantically similar phrases to any phrases containing the word "Raynaud". Finally, these phrases are filtered based upon their UMLS semantic type and compared to the terms found by both Weeber and Swanson to evaluate the usefulness of this method. These experiments found that when using these corpora the majority of links, an average of 88% of B-Terms for Open Discovery and an average of 81% for Closed Discovery, can still be formed. However there is still a large degree of manual curation necessary due to the imprecision of the process. This thesis shows that the development and imple-

mentation of such a system with improvements to its precision can be of use to the research community.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The average cost of drug development in the United States has increased from $403 million in 2000 (DiMasi, Hansen and Grabowski, 2003) to $648 million in 2017 and with costs ranging between $157.3 million to $1.950 billion (Prasad and Mailankody, 2017). There is widespread demand, both in the pharmaceutical industry and also within many healthcare facilities for more work to be done in the field of reusing compounds in diseases as a method to escape some of the most expensive and time-consuming processes in drug discovery (See Fig. 1.1).

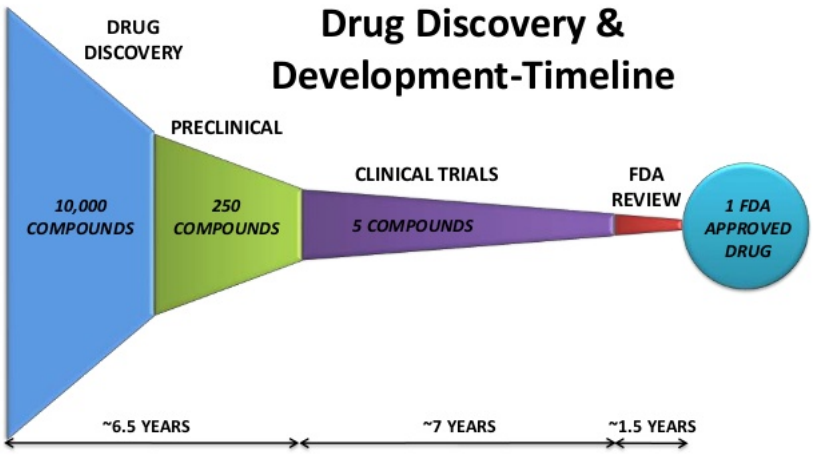

Figure 1.1: Drug discovery and drug development timeline until Food and Drug Administration (FDA) approval

The highly expensive field of drug discovery has driven many companies and researchers to try and focus their efforts on the much cheaper idea of repurposing (also known as repositioning) existing drugs to similar diseases. Whilst many people have focused on the cost of the development of new drugs as a barrier for much of the fields research. It has also been shown in research that another large contributing factor to many companies being more responsive to the idea of repurposing previously approved medication for new uses is that many government agencies have much more strict approval rates. This new found strictness is related to the fact that in previous years in response to the fact that some 450 previously approved drugs were later found to be dangerous, with the main cause being liver damage (Dialani, 2019).

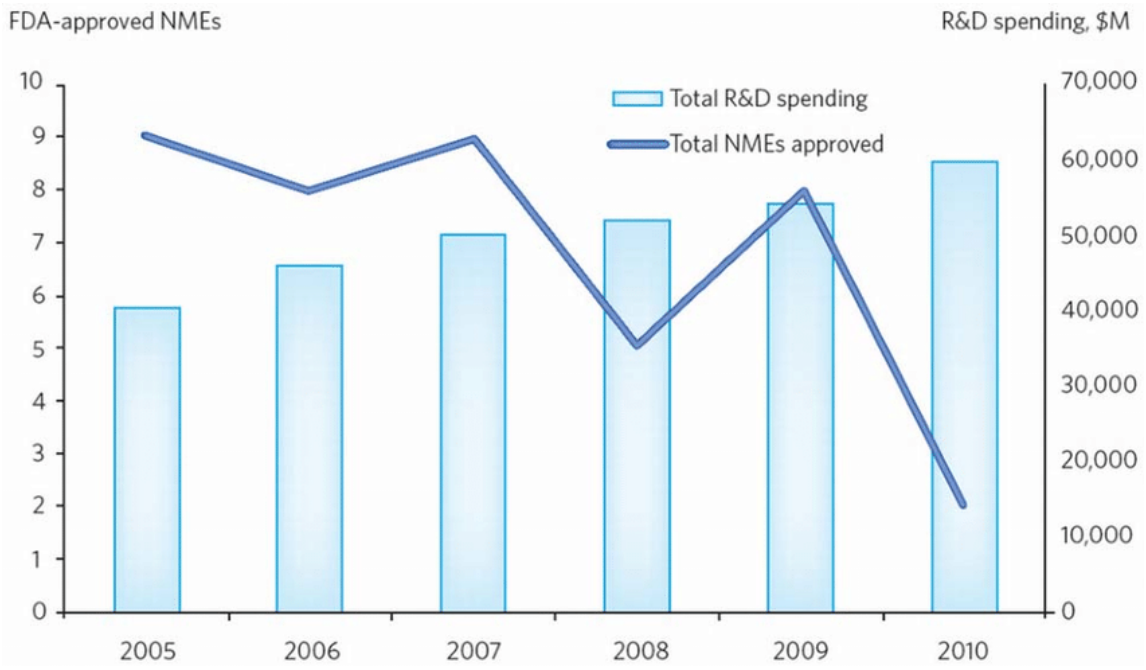

Figure 1.2: FDA number of drug approvals with respect to their R&D costs

The reason R&D cost, as seen in Figure 1.2, is still rising when the number of approved drugs is decreasing is because the R&D process is inefficient (Ayyadurai, 2014). When this argument is looked at alongside the previous fact research output is at a rapid level (See Fig. 1.1), it makes a large amount of sense why companies are wondering as to whether starting their own research on brand new undiscovered compounds is both financially plausible, but whether it also has a decent chance of being successful.

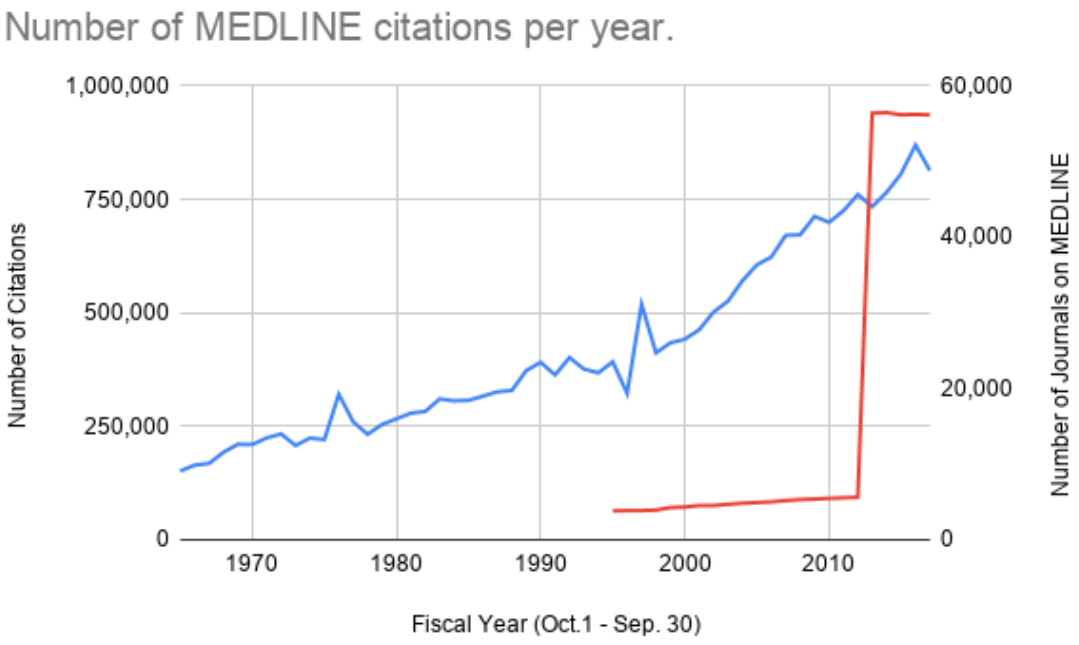

Figure 1.3: Number of MEDLINE citations per year

With a successful track record in the field, after most famously Sildenafil (Viagra) being repurposed from cardiovascular diseases to erectile dysfunction and Minoxidil from being an anti-hypertensive to being repurposed into Rogaine (Azvolinsky, 2017). The use of drug repositioning has been shown to have the potential to be beneficial not only to the healthcare facilities and pharmaceutical companies previously mentioned, but also to the everyday consumer due to the fact that if the process to finding and developing cures becomes cheaper, then the actual to consumer cost of treatment will likely decrease. There is also an argument not only for drug repurposing, but also for the effect that many dietary/vitamin supplements may have on certain conditions. For example, in 1986, Don Swanson in his article entitled "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge" (Don. R. Swanson, 1986a) found for the first time via the manual curation of published literature evidence the consumption of Fish Oil supplements may ease the symptoms of Raynaud's Syndrome, a medical condition which reduces blood flow to a persons extremities. Swanson's work was so groundbreaking that it birthed the technique currently known as Literature-Based Discovery (LBD). LBD is a form of Knowledge Extraction that uses academic literature that are currently seen as "Noninteracting Literatures" to

potentially uncover previously unreported links such as those between Raynaud's and Fish Oil (Don. R. Swanson, 1986a). In the years after Swanson reported his hypothesis lab experiments were able to concur and thus prove the link between Fish Oil and Raynauds Disease (Digiacomo, Kremer and Shah, 1989).

Since Swanson's work in 1986 many researchers have found that LBD has as the potential to not only potentially speed up the hypothesis generation phase of drug discovery, but of also its ability to link together previously existing but currently un-linked, islands of scientific literature. This is largely due to the rate of growth seen in many scientific fields, for example, the biomedical database MEDLINE has shown a growth from 16,113,221 in 2007 to 24,335,332 in 2017, a total increase of 51.03% and an average year on year growth of 4.2% (See Fig. 1.3 and Table 1.1). As this graph makes abundantly clear, it would be impossible for any scientific researcher to keep up-to-date with not only their own area of research, much less so any neighbouring and currently un-linked research areas.

Early LBD experiments utilised co-occurrence lexical statistics to extract concept relationships in text. Whilst these studies have had a moderate level of success it has become clear that they struggled to explain the relationships themselves. Later experiments attempted to overcome this issue through the utilisation of text-mining techniques, for example mapping text to concepts with the use of the UMLS (Unified Medical Language System) (Weeber et al., 2001) and Term Frequency - Inverse Document Frequency (TF-IDF) for term weighting (P. Srinivasan, 2004).

Fiscal Year (Oct. 1-Sep. 30)	Number of Journals Indexed in Index Medicus	Number of Journals in MEDLINE	Number of Citations in MEDLINE	Total Citations	Percentage Change on Previous Year
2017	5,150	5,617	813,598	24,335,332	3.42%
2016	5,136	5,623	869,666	23,531,948	5.09%
2015	5,123	5,618	806,326	22,391,870	3.75%
2014	5,118	5,647	765,850	21,582,742	4.29%
2013	5,067	5,640	734,052	20,695,240	3.61%
2012	5,025	5,633	760,903	19,974,272	4.28%
2011	4,946	5,559	724,831	19,155,303	4.44%
2010	4,866	5,484	699,420	18,340,055	3.96%
2009	4,759	5,394	712,675	17,641,559	4.46%
2008	4,660	5,319	671,904	16,888,640	4.81%
2007	4,520	5,194	670,943	16,113,221	N/A

Table 1.1: MEDLINE Indexing Statistics from 2007-2017 as found on the NIH website *Detailed Indexing Statistics: 1965-2017* n.d.

To give the LBD process a higher degree of autonomy the aim of this thesis is

to replicate the Raynaud-Fish Oil discovery (Don. R. Swanson, 1986b; Weeber et al., 2001) with the usage of recently developed deep learning methods of word embedding generation to prove that use of Word Embeddings is a suitable method for the task. This will involve the crafting of a variety of corpora, the normalisation of any potentially significant text, the creation of word embeddings from the content of this corpora and the filtering of words found to be potential matches to target words by their semantic types defined by the UMLS and its associated biomedical thesauri. This project also aims to investigate the optimisation of these word embeddings to potentially envisage any potential patterns in the optimal parameters that may potentially be applied to other diseases and corpora.

1.2 Description of Chapters

I have provided an overview of the motivation behind the approach I am going to follow in this thesis. The remaining of this thesis is organised in the following five chapters:

- Chapter 1 introduces the field of LBD. It also provides a justification as to why the research community have taken a keen interest in the field in recent years with a brief discussion on drug discovery and drug repositioning.
- Chapter 2 provides a detailed introduction to the field of traditional LBD, its progress throughout the years whilst highlighting the problems that these systems encountered and how future LBD systems tried to solve these problems. Particular focus was paid not only to Swanson's original work but also on the experiments that were subsequently conducted to validate the "Raynaud-Fish Oil" hypothesis.
- Chapter 3 provides a detailed explanation of the implementation of an LBD system that aims to replicate Swanson's hypothesis through the use of word embeddings generated by the Word2Vec group of models. It is in this section that my proposed system and its individual components are broken down, starting with the retrieval of the free text data from the PubMed/MEDLINE database, moving onto the pre-processing of the text data through the use of text normalisation methods. This section also contains a detailed methodology which explores all of the libraries and systems that are the backbone of this research and how they themselves have been created and the algorithms behind them.
- Chapter 4, presents the results of the experiments undertaken. It is here that it becomes clear as to whether the usage of Word Embeddings are utilisable as a variant of both discovery methods as defined throughout the history of LBD.
- Chapter 5 concludes this thesis as well as it identifies the potential areas for further research and how this research can help further the field of literature based discovery.

1.3 Value of This Thesis

In this thesis an innovative pipeline has been developed for the (re)discovery of the Raynaud-Fish Oil Hypothesis. This pipeline has been developed utilising some of the many publicised NLP techniques found to be effective in previous literature. These are:

- The use of Word2Vec to replicate the Raynaud - Fish Oil Discovery.
- The development of a pipeline from pre-processing PubMed data into a Word2Vec model and thus usable as a method of Literature-Based Discovery.
- The investigation into whether the specificity and generality of a corpus can improve/worsen the results of a Word2Vec literature based discovery system.

The utilisation of these methods will allow for more data to be employed than before. It will also give an insight as to whether these methods can be successfully used for other biomedical literature-based discovery methods. This work will also provide a fully-fledged set of tools to read data from the XML format given by PubMed into data ready for word embedding generations which will then be used to retrieve semantically similar terms.

Chapter 2

Related Work

2.1 Swanson's Initial Works in the Field

The first research works on LBD were published by Don R. Swanson in the late eighties. These publications were primarily aimed towards defining, utilising and verifying a new method of scientific discovery. The first of these three papers released by Swanson "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge." defined this process. It is within this paper that Swanson utilises co-occurrence between elements in the titles to test the strength of his "Raynaud-Fish Oil" hypothesis. To do this, Swanson took two groups of specifically chosen literature to match his starting (Raynaud) and ending (Fish Oil) literature, otherwise known as "closed discovery", this differs from Swanson's other discovery type "Open Discovery" (OD) because OD does not require both sets of literature, as it searches the initial (Raynaud) literature for any potential pathways. These groups, while chosen for their relatedness to specific attributes such as blood viscosity, platelet aggregability and vascular reactivity the literatures were completely isolated from one another. This paper explained the idea that due to dietary fish oils being hypothesised to help contain different symptoms of Raynaud's, such as Thrombosis in one set of literature, the theory that these oils could be potentially connected to Raynaud's disease is plausible. Whilst these papers have suggested a viable and working manual process to find these new connections in disjointed literature, it wasn't until 1991 for Swan-

son to propose a potentially automated process (Don R. Swanson, 1991). Following this, a system named ARROWSMITH was developed and integrated in 1997, details of which are discussed in section 2.1. This process still utilised the potential for linking elements in each title e.g. Migraine could be linked to magnesium due to serotonin. This level of thinking would understandably make the manual process a lot more tedious, whilst this 1991 paper focuses on the Magnesium-Migraine hypothesis its process is very interchangeable with the aforementioned Raynaud hypothesis as shown by the works published by Weeber in 2001. Swanson struggled in his paper to find the "intermediate" stage records simply from the terms that co-occur with migraine in the title due to the large number of them approx 120,000 (Don R. Swanson, 1991). However, Swanson did state that a certain categorisation/filtering strategy could be utilised to narrow these down.

2.2 Arrowsmith

As Swanson's ideas for finding undiscovered knowledge in already published literature were becoming more accepted in the research community, it became of additional value for an automated system that was able to utilise these methods to be developed and tested. One of the first tools designed for this task is known as "Arrowsmith", produced by Swanson and Smalheiser in 1997 as the subject of their paper "An interactive system for finding complementary literature's: a stimulus to scientific discovery". The Arrowsmith system works by taking a user's input, setting that as the "C" literature, a secondary set of literature which is complementary to the first is also provided, known as the "A" literature. This system then automatically retrieves any terms that co-occur with the C-Terms, these terms are then known as B-Terms. One feature that Arrowsmith incorporates, which improves upon the manual methods, is that A/B-Terms are ranked allowing users to guide their own decisions on the importance of a term. The technique utilised for this procedure is a multi-step method, which also includes the filtering of the list. The first of these steps is the removal of all unsuitable words, for example, those which are off-topic. However, the system also performs a search of the MEDLINE database to retrieve

how many different article titles each B-Term appears in, then calculating how many of these results are related to the current search, retaining only those which have a small probability of leading to a number of co-occurrences with migraines (Don R Swanson and Smalheiser, 1997). This process also utilised manual curation which is shown in step three; this step consists of a human examining the filtered list and judging themselves which entries are not-suitable. The rank of the remaining terms is calculated through a search of MEDLINE through these B-Terms, and the current word occurrences are found and used to form the base A-B relationships. Whilst the procedure for finding the B-Terms is slightly different than the one previously defined for finding the A-Terms it is due to the fact that the authors have attempted to prevent the possibility of retrieving already known links.

2.3 DAD-System

Published four years later than the Arrowsmith system in 2001 by Marc Weeber et al., the DAD-system attempted to use a mixture of concepts and statistical Natural Language Processing (NLP) techniques to utilise PubMed citations as a method of text-based discovery. In this research, the authors have attempted to codify the process as a method of assistance for biomedical research. However, they have made a conscious effort to make sure that the user is still at the centre of the process. This system is similar to the work of Gordon and Lindsey but very different to the Arrowsmith system explored in section 2.1 of this literature review. This similarity lies in the fact that the DAD system starts the discovery process with the C literature to find the B and A terms, but then in a second step the system examines both the A and C literature to test their hypothesis, whereas the Arrowsmith system utilises a one-step system to generate its hypotheses. Additionally this system expands on Swanson et al's work by resembling both open and closed discovery in their two procedures whereas ARROWSMITH only utilises the "closed discovery" method (Smalheiser and Don R Swanson, 1998). One other advancement that this piece of work has made over Arrowsmith is that it does not only treat words as words but also as concepts. The DAD system does this by only analysing those that are

found in the UMLS (Unified Medical Language System), this is done due to it only allowing words of interest, otherwise defined as those words with meaning in the biomedical domain to be analysed whilst also allowing even further specialisation and filtering through the semantic types allocated to these words which would limit the number of pathways found in a more automatic method than the one proposed by Swanson and Smalheiser. Another advancement that this paper has made is that it uses a more advanced and automatic method of n-gram generation and classification of meaningful n-grams from those which are meaningless. Whereas Swanson and Smalheiser utilise an extensive list of stop-words (Smalheiser and Don R Swanson, 1998), this piece of work has utilised the UMLS (Unified Medical Language System) to map the free-text into different biomedical concepts, which are then utilised within the discovery process. One important point that should be noted here is that whilst this method is slightly more automated, the authors make the note that these semantic filters are not a one-size fits all method and are dependent on the query being processed.

2.4 Lit-Linker

Published two years after the DAD-System, LitLinker is also a system that attempts to utilise NLP methods to be able to mine the currently released biomedical literature for new potential links (Pratt and Yetisgen-Yildiz, 2003). To do this they, like most others, attempted to build on Swanson's approach which differs to many of the approaches such as Arrowsmith through the use of intermediate literature to be able to limit the search space. Their approach is, however, similar to Swanson's Arrowsmith tool due to the fact they are utilising only the titles of articles. In contrast, the experiments by Weeber were made using titles and abstracts of the relevant data. The LitLinker paper restricts their concepts only to titles as it allows for a limit in the number of terms found, thus making it easier to compare and prune the list. Whilst the DAD-System solely utilises the UMLS knowledge base for its automated concept pruning whilst having it backed by manual intervention the Lit-Linker system found their were three types of erroneous connections:

1. Too general terms - Terms such as problem, test, therapeutic
2. Too closely related terms to the start term
3. Terms that do not make sense as connections

One method the authors utilised to remove these terms was to remove certain terms that appeared at the second level of the UMLS. However they found that this process only eliminated a small number of found terms. Due to this fact they then had to include another step that removed terms that were found too frequently, ten-thousand occurrences, in the titles. This step automates the pruning process even further than those provided in the Arrowsmith and DAD-Systems. After this point however, they then followed the same step as Weeber and utilised a semantic filter. Whilst this may reduce in a slightly smaller list of potential terms, it should be noted that the author's technique did eliminate some previously found terms in Swanson's Magnesium-Migraine hypothesis.

2.5 Other LBD Experiments

While our previous sections have focused purely on the usage of literature-based discovery techniques when used in the Raynaud-Fish Oil hypothesis. It should be noted that there have been other pieces of work in the field in relation not only to other diseases (Pyysalo et al., 2018; Meng et al., 2018) but as previously mentioned to entirely new areas such as Material Sciences with (Tshitoyan et al., 2019) and Neuroscience with brainSCANr which is the result of the works published by Brad and Jessica Voytek in 2013. These three papers all replicate known scientific hypotheses in their fields, such as the relationship between Progesterone and Aging (J. B. Voytek and B. Voytek, 2012) and Thshitoyan et al. verified previously made thermoelectric claims (Tshitoyan et al., 2019). As the previous statement shows, the majority of LBD experiments have been focused upon the fields of biomedicine and drug repositioning. This research trend has continued into more recent years as seen by the LION tool mentioned above and also in regards to research into whether Ketamine is a valid therapy for those with Alzheimer's disease (Smalheiser, 2019)

where the authors utilised the Arrowsmith tool previously discussed to find three new viable pathways as to how Alzheimer's disease can be affected through the usage of ketamine. These pathways were found to be:

- VGF Levels
- mTOR Regulation
- The process of Autophagy
- Inflammatory cytokines

As is shown in this section, the field of Literature Based Discovery is largely limited to biomedicine where it can be used to potentially kickstart research in areas that could be potentially struggling, this does raise the question as to how the systems could be utilised in fast-changing and developing pandemics such as the COVID-19 pandemic as a method of speeding up vaccine development.

2.6 Word Embeddings

As discussed in Section 2.4 many recent advancements in the field of biomedical LBD has been focused on the utilisation of word embeddings, however, this trend does not seem to have stopped at this small corner of Natural Language Processing. Wang et al. did a comparison of these word embeddings for the field in their 2018 paper. In this paper the authors' trained different types of skip-gram Word2Vec word embeddings on different sets of literature ranging from biomedical publications to news articles, for some of these they did utilise other collections such as the Google News set of Word Embeddings and also the GloVe model which allows for a comparison between different generation methods (Y. Wang et al., 2018). In this paper they found that the embeddings produced by Word2Vec using biomedical academic literature outperformed those produced by GloVe and Google News. These results were correlated by other publications such as Schnabel et al. who evaluated different unsupervised word embeddings and their uses in more general NLP tasks. It was in Schnabel's research that Word2Vec, albeit the continuous bag-of-words (CBOW) model, outperform many of the other word embedding methods utilised in

their experiments, for example, GloVe (Schnabel et al., 2015). This article did not directly compare a model employing a CBOW architecture with a model that uses the skip-gram method. However, S Henry, C Cuffy, and B T. McInnes published a paper in 2018 which experiments with both models which found that both Word2Vec models create better word embeddings than singular value decomposition and explicit co-occurrence vectors. Still, there was not any significant increase/decrease from a specific method (Henry, Cuffy and McInnes, 2018). However, there have been other works released which have found the skip-gram model to not only be preferable for the task of biomedical NLP (TH, Sahu and Anand, 2015). On the other hand, there have been been more in-depth releases that evaluate dimensionality but also when training on an extensive, 1.25 million article, corpora taken from PubMed (Chiu et al., 2016).

Chapter 3

Materials and Methods

3.1 Materials

3.1.1 XML Parsing

XML Parsing is the act of converting data formatted as XML to usable in-memory text data. This method has been employed in this project as it is by far the easiest and fastest way of retrieving a comprehensive record of all information stored about each found document. This is due to the fact that by design, an XML document is heavily structured which allows a parser to specify tags to search for and thus allows them to ignore all other data stored in the file.

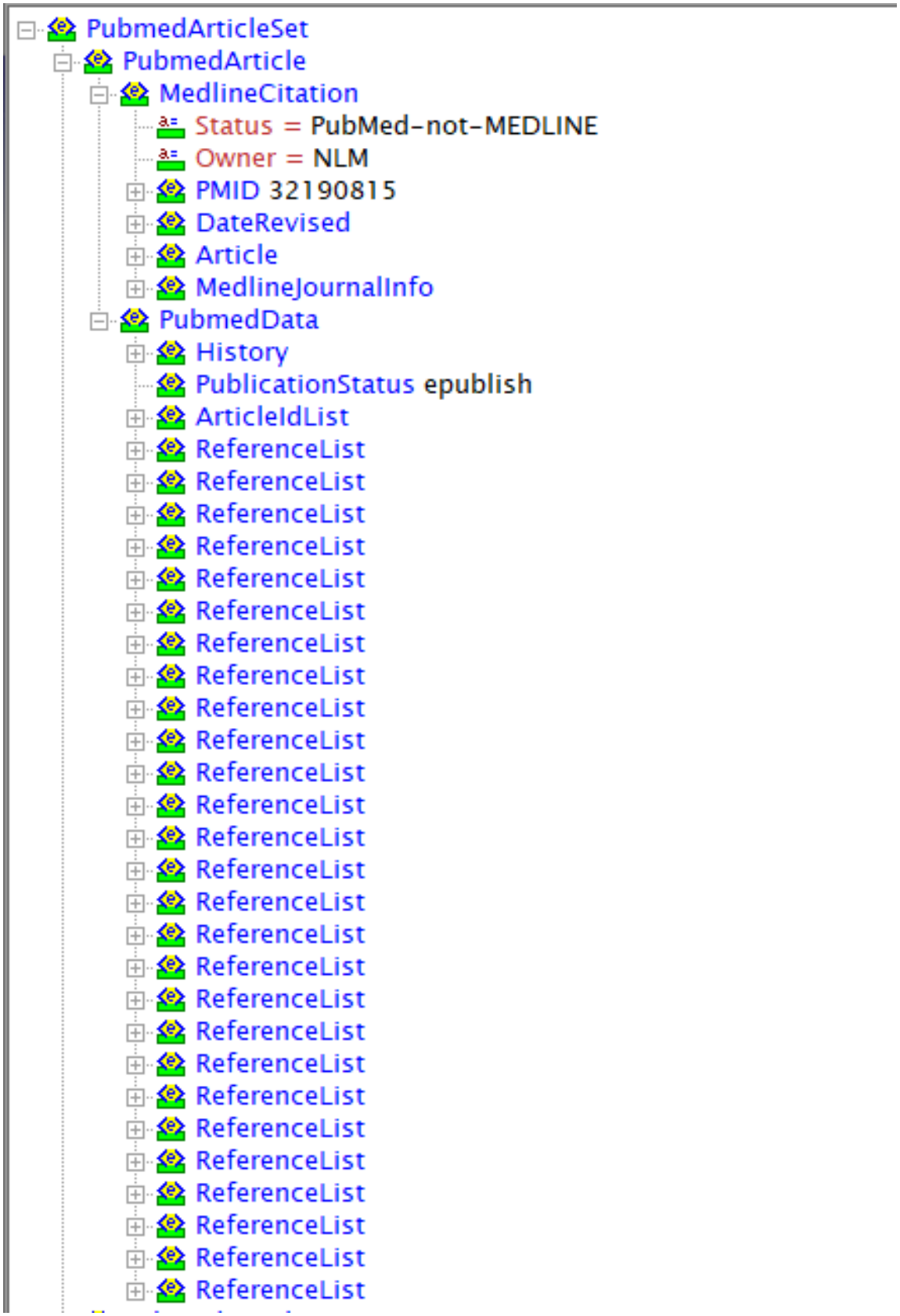

Figure 3.1: PubMed Article Set Base XML

This XML file, see 3.1, contains a lot of tags that are not necessarily required, however this file can be unzipped and when the MedlineCitation is opened it contains tags that are of interest such as the articles title, abstract and published language (See Fig. 3.2).

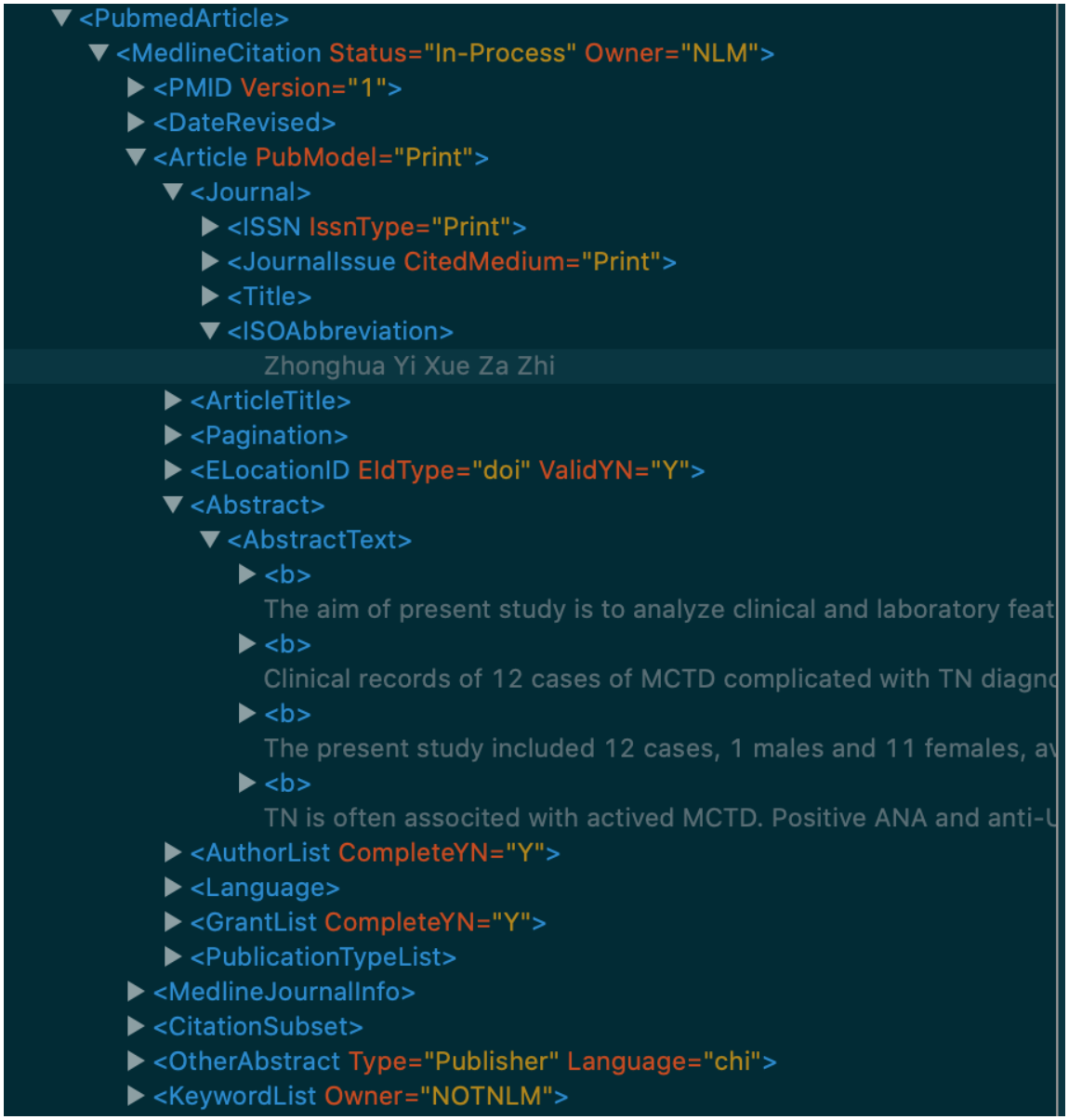

Figure 3.2: PubMed Article XML

3.1.2 Word2Vec

A group of models introduced by Mikolov et al. in 2013 (Mikolov, Chen et al., 2013; Mikolov, Sutskever et al., 2013), Word2Vec is known to be a method of generating Word Embeddings, these models are called continuous skip-gram was introduced in 2013 by Mikolov. Continuous skip-gram was introduced as a method that attempts to predict the current word from its context by inputting the current n-gram into a log-linear classifier and predicting the phrases within a specific range before and after the current word (Mikolov, Chen et al., 2013). A second method called the Continuous Bag of Words model which was also introduced as a method of generating word embeddings without taking into account the order of the words found however future words are taken into account. These methods have both been found to successfully find subtle semantic relationships when created using large dimensional word vectors from a large amount of data.

Continuous Bag Of Words (CBOW)

As mentioned above in Section 3.1.2, the simple definition of the CBOW model is a method of generating word embeddings without the syntactic order of those words being taken into account. This architecture is defined as being similar to a Feedforward Neural Network Language Model (NNLM) where the non-linear hidden layer has been removed and the projection layer is shared between all words (Mikolov, Chen et al., 2013) which is seen in its shallow nature, with the model comprising of only an input layer, a projection layer and the output layer. Due to the fact that this model does not take into account the location of the words in a sentence, it means that it is also able to take into account words also used in the future allowing it to more accurately classify the "middle" word.

Continuous Skipgram

Unlike the CBOW model defined above, the Skipgram model attempts to classify a word based upon other words defined in the same sentence. This done through the model taking the current word as an input to a log-linear classifier which through

the use of a continuous projection layer, maps each word index to this vector space which is then fed into the output layer which holds the probability that the next word is each word in the vocabulary. Mikolov et al. in 2013 found that increasing the "range", otherwise known as the context window will improve the quality of the resultant word vectors. However, this does increase the training complexity of this model as shown below:

$$Q = C \times (D + D \times \log_2(V)) \tag{3.1}$$

Where C is equal to the maximum distance between words, D being the dimensionality and V being the size of the vocabulary (Mikolov, Chen et al., 2013).

It should be noted that due to the skip-gram model utilising more information than the CBOW model, it is predominately the slower of the two but does a better job for infrequent words. There has however been a large amount of work into improving the performance of the skip-gram model with some authors focusing on an extension of the skip-gram model inclusive of negative sampling. When this was initially defined by Mikolov needed to go through the training data a minimum of two times whereas in 2017 an extension of this model was provided to allow for incremental model (Kaji and Kobayashi, 2017). Negative Sampling works through a sigmoid function and by getting a smoothed unigram probability distribution and only keeps those words it finds occurring enough times.

3.1.3 MetaMap

Developed to map free-text to biomedical concepts found in the UMLS, MetaMap uses a group of lexical analysis techniques to find the best matched mapped terms to terms in a given phrase (A. R. Aronson and F. M. Lang, 2010). The tool starts by creating phrases of all the text after parsing the text into predominately simple noun phrases which allows for easier limitation of further processing, this is done through the use of a minimal commitment parser for SPECIALIST which is a large English language lexicon of biomedical terms (McCray, S. Srinivasan and Browne, 1994).

This parser utilises the Xerox part of speech tagger to assign generic syntactic tags to those words that do not have a unique tag in the SPECIALIST lexicon. Once this is done, the knowledge of the SPECIALIST lexicon is supplemented with a database of synonyms and a generator is used to find all variants including any acronyms, abbreviations, variants, and synonyms. These are stored alongside their POS-Tag in the order the variant was created. MetaMap then generates and retrieves a subset of the metathesaurus of all phrases that contain any of the variants found. To improve performance MetaMap includes options to ignore those terms of only one or two characters. Once this is done the tool then evaluates every candidate using a weighted average of the centrality (involvement of the head), the variation (the average of the inverse distance scores), the coverage (how well it matches the term), and the cohesiveness of the term (how many pieces this term has) (Alan R Aronson, 2001). The closer this score is to one thousand the better a match the term is. Each mapped term found by MetaMap is assigned to one of 134 different semantic categories, which are discussed below.

3.1.4 UMLS

The Unified Medical Language System (UMLS) is the most comprehensive collection of biomedical vocabularies which was released in 1986 by the National Library of Medicine (Humphreys et al., 1998). The UMLS has experienced such growth that in 2004 the UMLS consisted of over two million different names for over nine-hundred thousand unique concepts (Bodenreider, 2004). However, the latest release, 2019AA, consists of 14.6 million concept names for over 3.85 million concepts held in 210 different sources(Health, 2019). Due to the massive number of available concepts in the UMLS, this project utilises the tool alongside the mapping tool MetaMap that is discussed in Section 3.2.3 to map phrases from the free text found to its closest possible entry in the UMLS.

Semantic Types

A method of assigning biomedical concepts to different semantic categories stored in the UMLS Semantic Network makes it easier to distinguish between two concepts (Bodenreider, 2004). A semantic filter is a method of reducing the amount of data returned by only including those found to be in the same categories as those found useful (See Table 3.3), this method was used successfully by Weeber in his own automated LBD system in 2001 (Weeber et al., 2001).

3.2 Methodology

The pipeline used in this project is as follows:

1. Passed each title/abstract of each compressed corpus through an Natural Language Toolkit (NLTK) parser to generate bigrams/trigrams of each unigram based on a minimum occurrence count value.

2. Employed a Skip-gram word2vec model with initial parameter values as in Table 4.1 to generate word vectors for all words and phrases in each corpus.

3. Scan through all generated word vectors to discover variations of the "raynaud" C-concept (e.g. Raynaud's disease, Raynaud syndrome, primary Raynaud, etc).

4. Utilised a grid search on the architecture, dimensionality, epoch, learning rate, downsampling, context window and minimum word count parameters to find the model with the optimum performance in each corpus used.

5. Using the optimally derived word2vec model, we repeated STEP 4 to estimate cosine similarity of all terms in the corpus with Raynaud variation terms from STEP 3.

6. Placed the most semantically similar terms , as defined as those with the closest cosine similarity, from STEP 5 into a list.

7. Mapped every term from the list saved in STEP 6 via MetaMap to UMLS ontologies. Then using a semantic filter (see Table 3.3) we excluded from

further analysis all mapped terms which were not semantically related to the semantic types in the filter.

8. These results are then compared to Weeber's found terms to see if the tool provided acceptable results.

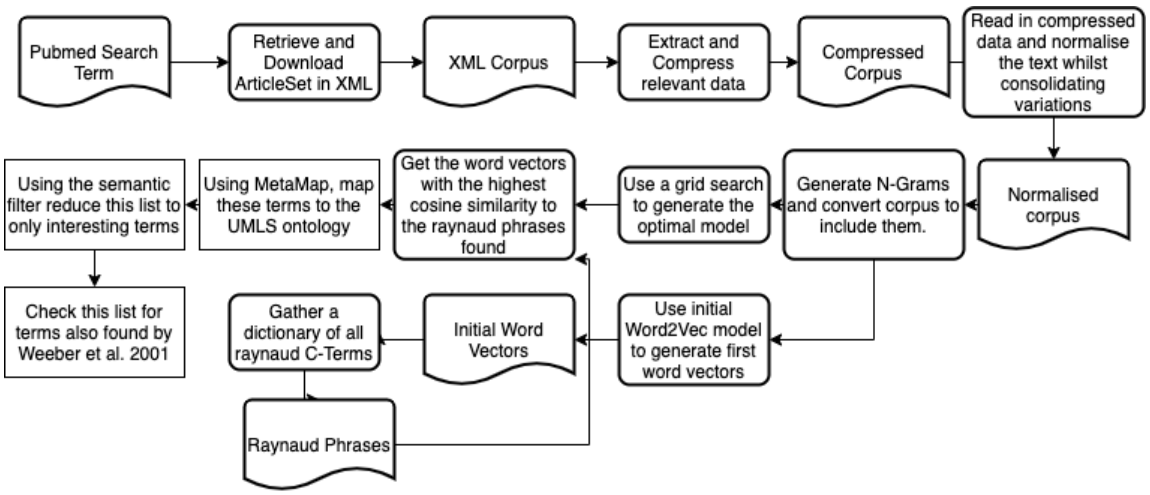

Figure 3.3: Overview of Pipeline from start to finish

3.2.1 Data Retrieval and Parsing

Data selection and pre-processing is a critical part of any research project; for this experiment, the datasets utilised were all curated and downloaded from the MEDLINE database by utilising the PubMed search engine. As the majority of the experiments are based upon replicating the Raynaud-Fish Oil experiments previously utilised, many of the corpora formed are representative of this fact. These search terms are shown above next to the type of discovery they were utilised within Figure 3.1. Once the query has been handled by PubMed the data is then downloaded and returned in a XML file format. As shown in Figure 3.1 a PubMed Article Set is a very large and comprehensive list of relevant information to each article retrieved in the search. Due to this fact the size of a query that encompasses all literature on a disease over a period of years, as used in this research will end up being an extremely large file. Once this file has been downloaded and is saved locally, a script known as *BioParser* was developed which is able to read through a directory of files, retrieve all XML files found (if the user wants to combine multiple) and then find all MEDLINE citations in

said XML files, with an example shown in Figure 3.1 for further processing. For each MEDLINE citation found, the script then only extracts the necessary information required to lower the memory usage of the program as seen in 3.1.

This triplet of information is then iterated through a process where filtering occurs, removing any items missing information to make sure that the project would only be utilising complete datasets with these being initially saved as .txt files, as shown in Figure 3.4. However, the functionality is there to include those which are missing parts of the data. Once this filtering has been done the finished articles are then compressed down into a .txt.gz file to save on local disk space.

Figure 3.4: List of Text Files before Compression

Corpora Search Term	Search Type
(Raynaud) AND ("1960"[Date - Publication] : "1986"[Date - Publication]) AND ("english"[Language])	Open Discovery
(Peripheral vascular disease)) AND ("1960"[Date - Publication] : "1986"[Date - Publication]) AND ("english"[Language])	Open Discovery
(Vascular disease)) AND ("1960"[Date - Publication] : "1986"[Date - Publication]) AND ("english"[Language])	Open Discovery
((((((((((((((Raynaud) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fish Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Maxepa) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fatty Acids, omega-3) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Omega-3 polyunsaturated fatty acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Eicosapentaenoic acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Epa-e) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Cod Liver Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fish Oils) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Salmon Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fatty acids, essential) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Dietary Fats) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) AND "english"[Language])	Closed Discovery
((((((((((((Peripheral Vascular Diseases (PVD): (Peripheral vascular disease) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fish Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Maxepa) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fatty Acids, omega-3) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Omega-3 polyunsaturated fatty acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Eicosapentaenoic acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Epa-e) AND ("1960"[Date - Publication] : "1986"[Date - Publication])) OR ((Cod Liver Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fish Oils) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Salmon Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fatty Acids, essential) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Dietary Fats) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) AND "english"[Language])	Closed Discovery
((((((((((Vascular Disease) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fish Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Maxepa) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fatty Acids, omega-3) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Omega-3 polyunsaturated fatty acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Eicosapentaenoic acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Epa-e) AND ("1960"[Date - Publication] : "1986"[Date - Publication])) OR ((Cod Liver Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fish Oils) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Salmon Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Fatty acids, essential) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) OR ((Dietary Fats) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) AND "english"[Language])	Closed Discovery
((('Expert Review of Hematology"[Journal]) OR "British Journal of Haematology"[Journal])) OR "Blood Reviews"[Journal])) OR "Haematologica"[Journal])) OR "American Journal of Hematology"[Journal])) OR ("Blood Cells, Molecules and Diseases"[Journal])) OR "Blood"[Journal]))) OR ("Pediatric Hematology and Oncology"[Journal])) OR "Pediatric Blood Cancer"[Journal]) OR "Experimental Hematology"[Journal]) OR "International Journal of Hematology"[Journal])) AND "english"[Language])) OR (((Fish Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Maxepa) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Fatty Acids, omega-3) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR ((((Omega-3 polyunsaturated fatty acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Eicosapentaenoic acid) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Epa-e) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR ((((Cod Liver Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Fish Oils) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Salmon Oil) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Fatty acids, essential) AND ("1960"[Date - Publication] : "1986"[Date - Publication])))))) OR (((Dietary Fats) AND ("1960"[Date - Publication] : "1986"[Date - Publication]))) AND "english"[Language]))))	Closed Discovery

Table 3.1: The seven search terms utilised within this experiment

3.2.2 Data Pre-Processing

Some of most popular examples of preprocessing which are becoming heavily used include the removal of stopwords from text. However, this is not utilised by Weeber et al. (Weeber et al., 2001), stemming/lemmatization, which is the subject of the BioLemmatizer Paper (Liu et al., 2012) and casefolding. Due to the number of changes that were made regularly to the code and pipeline, the decision was made quite early in the project to utilise as few permanent pre-processing techniques on the data before it was written to disk and to do them when the data is read into memory ready for usage. The first pre-processing step that is utilised is through the Gensim library's *simple_preprocess* function which lower-cases all text. This pre-processing step is taken as a method of normalising the text and to reduce the number of variants of each word due to capitalisation etc, this is a practice that has been found to have a positive impact on the generation of biomedical word embeddings(Chiu et al., 2016). It is also used to remove any any word with less than three characters. A simple graphical overview of the process undertaken in section's 3.2.1 is shown below.

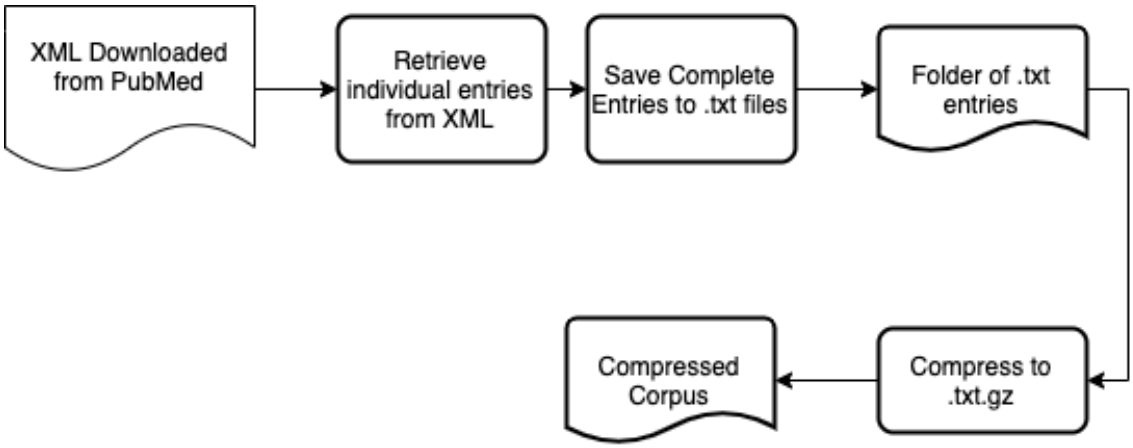

Figure 3.5: Corpus Creation

As mentioned above, the traditional approach taken during the pre-processing of free text is the removal of stopwords however this is not implemented in this research. This is not done for a few reasons, the first being that as defined in the Word2Vec paper, the Word2Vec model utilises a downsampling technique based on how frequently a term appears in the corpus (Mikolov, Sutskever et al., 2013). This means

that the removal of stopwords would not have any significant impact on the results and would be more akin to a waste of processing time and power. There is also the use of Lemmatization which when utilised with Word2Vec can make the vector space sparser, however, it has been stated in previous works that Word2Vec allows semantically similar words to overlay without the need for lemmatization (Major, Surkis and Aphinyanaphongs, 2018).

Text Normalisation

The first of the pre-processing steps taken in this process were to normalise the data. This research utilised a few different steps of text normalisation based on the success rate of those utilised in other pieces of literature. The first was capitalisation normalisation, defined by Gupta and Lehal in their 2009 paper as "casefolding" (Gupta, Lehal et al., 2009). Case-folding is the task of converting all text to either lower or upper case, in this research lowercasing was employed. This form of case normalisation is used to make sure that all variants of a term are treated as the same word vector e.g. Raynaud and raynaud. A secondary step of normalisation was the removal of all tokens with a text length of lower than 3. This removal was done due to many items not providing as much importance as those above it for example many terms under this are preposition terms that do not provide a great deal of information likely valuable to our task. At the same time as this occurs the text is tokenized at a word level thus returning a list of all words at a length greater than three, all in the same lower-cased form and with no punctuation or digits.

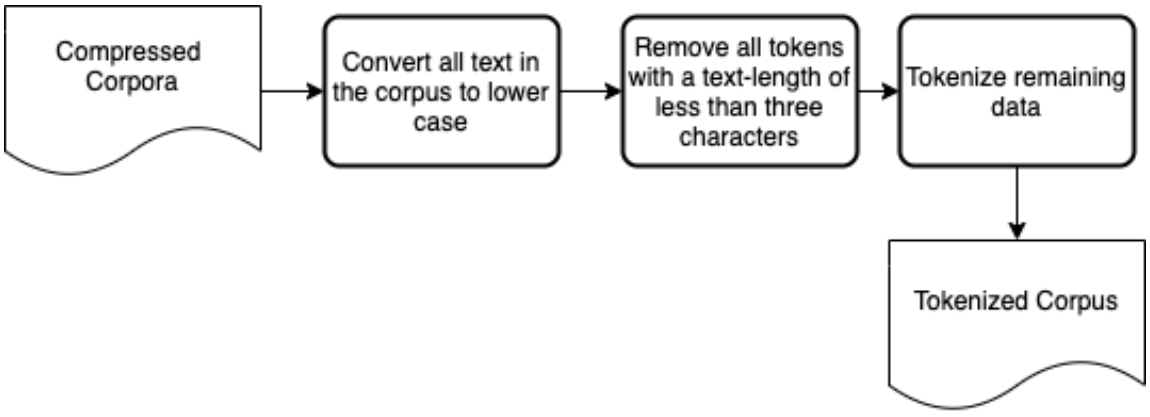

Figure 3.6: The process taken to normalise the text used in this process

Generation of N-Grams

Due to the fact that Word2Vec was developed to find representations of "words" it became clear that to allow the model to utilise N-Grams a process would have to be implemented within the pre-processing pipeline that found any n-grams in the corpora and transforms the text to include them. This was implemented using the scoring equation found in Mikolov's 2013 paper.

$$score() = \frac{count(count_i, count_j) - \delta}{count(w_i) \times count(w_j)} \tag{3.2}$$

The above equation is used to generate a score for each potential n-gram found in the corpus, the delta within this equation is used as a discounting coefficient, thus making sure that there are not an erroneous number of phrases are generated with infrequent words. After these scores have been generated for all possible phrases, the only phrases taken into consideration are those with a score meeting a threshold. For this research the process was run twice to generate both bi-grams and tri-grams due to the success of other of this technique in experiments (Ye et al., 2016). Additionally, as a method of increasing the size of our corpora without sacrificing its specificity which was a large concern as shown by the paper "Bigger does not mean better! We prefer specificity" (Dusserre and Padró, 2017).

After the pre-processing stage of the project, the normalised corpora was then plugged into a Word2Vec model which utilised multiple hyper-parameters to yield the optimal word embeddings. The decision to utilise Word2Vec word embeddings was made early on in the project given to the success of papers published in other fields. (Tshitoyan et al., 2019)

3.2.3 Word2Vec Model Creation

Once the chosen corpus has been pre-processed and been transformed to include its phrases a Word2Vec model is created using this text data as an input. The other parameters for this model are based upon the results of a grid search which was run

for each corpus with the results being outlined in Chapter 4 of this thesis. The usage of optimised hyper-parameters allows for the best possible word embeddings to be created for each corpus thus allowing the next steps in the pipeline to perform at the best they could.

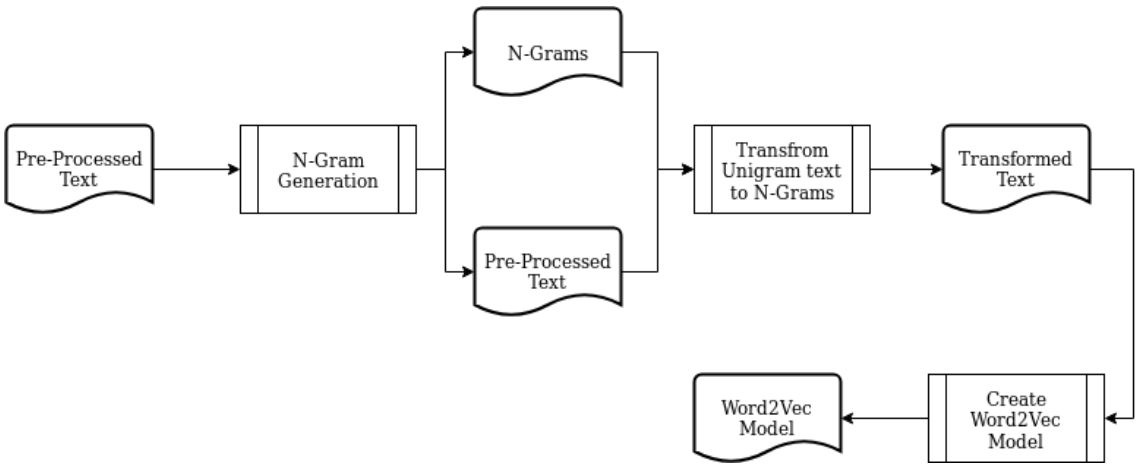

Figure 3.7: The process taken to generate Word Embeddings in this project

Figure 3.7 shows there is a large number of steps taken to transform the text from the saved pre-processed text into the n-gram transformed text that is placed into the model.

Hyper Parameter Analysis

Parameters	Options
Epochs	10, 25, 30, 50
Minimum Word Count	1, 10, 15, 25, 50
Context Window Size	1, 10, 30
Learning Rate / Alpha	0.01 to 0.1 with jumps of 0.005
Downsample Rate	1e-03, 1e-04 ... 1e-08, 1e-09

Table 3.2: Parameters for Grid Search

The parameters as seen in table 3.2 will all have their own effects as to the performance of the model. For example, the number of epochs is the number of iterations over the corpus the model performs, with many large corpora only requiring 1-3 iterations it may be found that this project needs to go beyond that due to the size. The minimum word count is the minimum number of occurrences a word must have in the corpus to be taken into account, whilst this project has tried to consolidate all variations of words into as few as possible this parameter could remove potentially interesting information if too high or could keep too much background noise in the corpus if too low. The context window size is how many words either side of the target word is taken into account, thus too low a value could struggle to find the context of a word due to lack of information. The learning rate is the magnitude to which word updates are shifted along a gradient (Chiu et al., 2016). The final hyperparameter changed throughout the grid search is the downsampling rate which as discussed in Section 3.2.2 is a method used by Word2Vec to dilute the most-frequent words in a corpus to levels similar to the rarer words.

For this experiment we utilised a 256 vector model, there were a few reasons for this. The first being that many papers have found that a word vector size of approximately 200 performs better than those of 100 (Tshitoyan et al., 2019; Gu et al., 2018; Chiu et al., 2016), furthermore, it has also been found that when the size of a vector hits 300 the effect it has is limited (Li et al., 2017).

3.3 Mapping

Once the word embeddings had been created and those with the highest cosine similarity to all of the generated Raynaud Phrases were retrieved, it became necessary to map these terms to their UMLS concepts as mentioned above. This was done to allow for a meaningful reduction of terms whilst also categorising those remaining into relevant biomedical categories ready for filtering in the next section of the process. The process of mapping a term to its UMLS concepts is a complex one, with the inclusion of many different text mining techniques to get the optimal result. In this section the method utilised and described will be taken from the MetaMap tool as that is used in this project. The first sections of the tool are based primarily in the area of lexical and syntactic analysis with the input text initially having the sentences detected, the text tokenised and the identification of acronyms and abbreviations. The tool then moves onto its POS Tagging module, these words are then searched for within the SPECIALIST lexicon and ran through a parser called the "SPECIALIST minimal commitment parser" as defined by McCray in 1993 (McCray, Alan R Aronson et al., 1993). These found phrases are then further analysed through a deeper, more thorough process. This process starts through the generation of as many variant phrases as found, each phrase is automatically searched to identify any candidate phrases, a score of these candidate phrases is also generated which details how closely they are related to the input text. These are then combined into phrases which are then compared to the input text to find those that match it closest. In cases where multiple phrases match the input text, MetaMap can utilise a word-sense disambiguation to bolster the confidence of its choices (Alan R Aronson and F.-M. Lang, 2010).

3.3.1 Semantic Filtering

Once the text is mapped to the UMLS metathesaurus, the next step taken is to reduce the number of terms found to only those found to be potentially significant. This is done utilising the filtering based on the semantic type method found in Weeber et al. 2001 paper (Weeber et al., 2001). The similarity types used can be found in section

3.3 in section 3.2.3 of this thesis. This list was formed with the majority of terms being chosen due to the fact that they were used in very similar experiments by Weeber as stated above to allow for some consistency when comparing the method defined here to their method.

Semantic Type (Long Form)	Semantic Type Acronym
Biological Function	biof
Body Location or Region	blor
Body Part, Organ, or Organ Component	bpoc
Body Space or Junction	bsoj
Cell Function	celf
Laboratory or Test Result	lbtr
Molecular Function	moft
Organism Function	orgf
Organism Attribute	orga
Organ or Tissue Function	ortf
Pathologic Function	patf
Phenomenon or Process	phpr
Sign or Symptom	sosy
Physiologic Function	phsf
Lipid	lipd*
Vitamin	vita
Element, Ion, or Isotope	elii

Table 3.3: Table of the Semantic Types that are used in the filter, those with a star are only found in older of MetaMap.

Chapter 4

Results

4.1 Hyper-parameter Analysis: Grid Search

To make sure that the model was performing sufficiently, an optional step was built into the pipeline that allows the user to utilise the grid search hyper-parameter tuning method. Due to the time and computational costs of the grid search method an analysis was made of the different parameters available to the model and the effects each parameter has on the different results outputted by the model. This was done with a two pronged method, the review of previous similar experiments and also the investigation of certain changes on our own corpora. Each model created throughout these options had its effectiveness tested through the use of average cosine similarity score of all Raynaud's Phrases and the other found significant B-Terms found in the model's vocabulary. Both of these equations are visualised in the below equations:

$$Average\ Similarity\ Score = \frac{\sum Similarity\ Scores}{Number\ of\ Similarities}$$

With the optimal model being the model with the highest similarity score. Initial experiments with this equation initially focused on one manually chosen word and the similarity between that word and the phrases found to include the term "Raynauds". The decision was finally made to include all found "significant phrases"

instead of just the one due to the fact that an average will give a more representative view of the models chance of being successful. The initial experiments were ran on the parameters shown in table 4.1, however, these did change as a grid search was performed and replaced with a new value which will be stated in its corresponding section. These experiments have been run in two variants, the first only searching for B-Terms and the second searching for a combination of B and A-Terms. This decision was made as it allows the reporting of whether performance weakens due to the inclusion of the A-Terms which are more likely to be weak relationships.

Parameter	Option
Epochs	50
Minimum Word Count	1
Context Window Size	10
Learning Rate / Alpha	0.0199999
Downsample Rate	1e-09

Table 4.1: Parameters for initial Skip-gram model utilised

4.1.1 Architecture

One of the first parameters that was put through the grid search was the architecture of the model. As previously discussed in Section 3.1.2 the Word2Vec group of models are based on two different architectures, Skip-gram and Continuous Bag-of-Words to make sure we were utilising the optimal architecture a grid search was ran to see whether there was a clear-cut correct choice for all three of our corpora. Below are two sections detailing each discovery types corpora and where the resulting experiments have been ran and ranked.

Open Discovery

As the table shows, there was not an outright best architecture for the open discovery experiments with the PVD corpus performing better with Skip-gram, whereas the other two performed better with a CBOW model. One thing that should be noted is the largest drop in performance is also between the PVD's two experiments

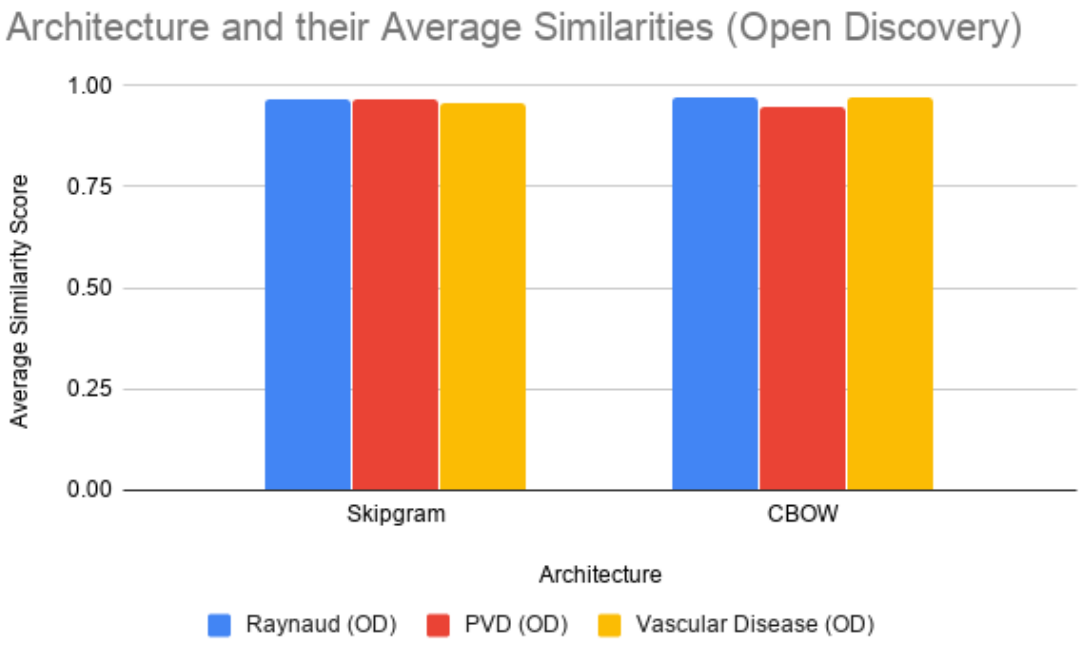

Figure 4.1: Comparison of different Architectures with an Open Discovery Method

with a decrease in average similarity of 0.018. Performing these experiments again with the inclusion of any found A-Terms finds that the best performing architecture differs dependent on the type of discovery being undertaken, with the open discovery performing best with a skip-gram architecture and the closed performing better with a CBOW architecture.

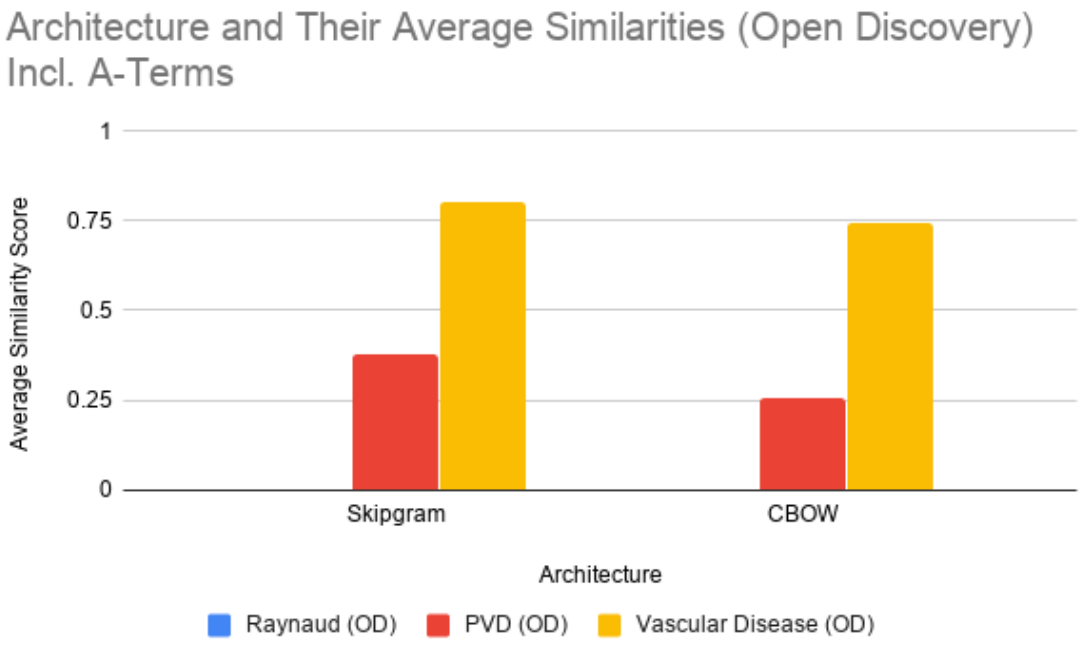

Figure 4.2: Comparison of different Architectures with an Open Discovery Method and inclusive of A-Terms

Closed Discovery

As seen in Figure 4.3 both architectures perform well with each one being optimal for 50% of the corpora. A pattern that has formed here that was not necessarily found was the fact that the two smallest corpora were found to perform best with a skip-gram model whereas the larger perform much better with a CBOW model.

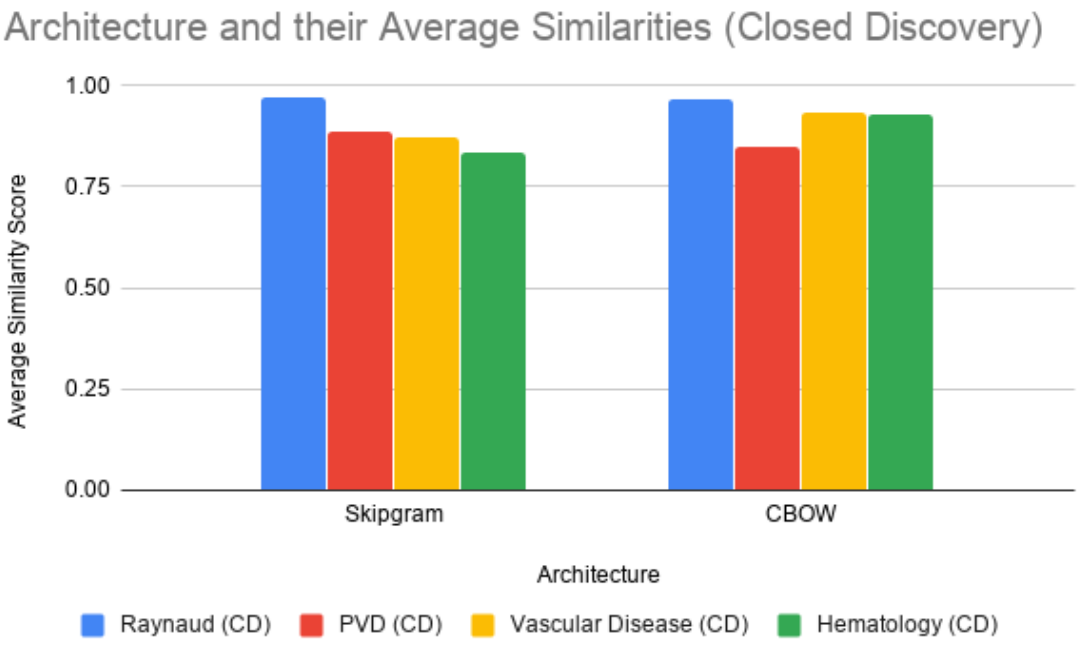

Figure 4.3: Comparison of different Architectures in a Closed Discovery Experiment

Figure 4.4 shows the results for these experiments indicate the best performing architecture for these experiments when taking into account A-Terms is the CBOW architecture.

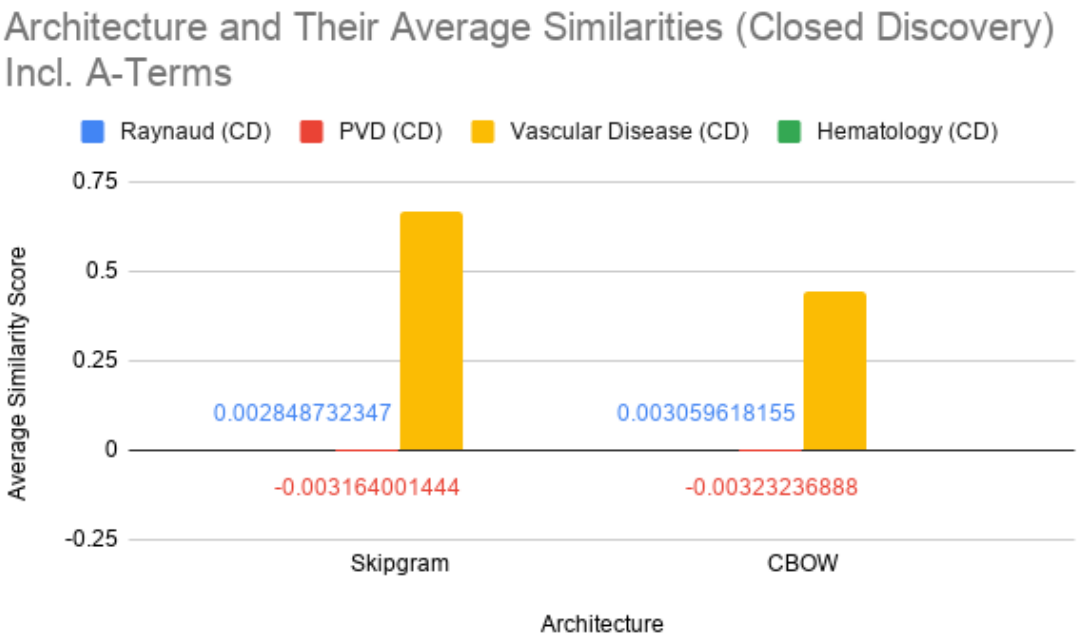

Figure 4.4: Comparison of different Architectures in a Closed Discovery Experiment Inclusive of A-Terms

4.1.2 Dimensionality

For this experiment a model was created utilising a dimensionality ranging from 100 to 950 in jumps of 50. This was experimented with to not only test the effect of a smaller dimensionality on the corpora but to also test the effect a larger dimensionality has on the results.

Open Discovery

The dimensionality of a word vector has been found to not result in a consistently optimal value in our experiments. One experiment found that a dimensionality of 100 was enough to generate optimal results, Raynaud Disease when only looking for B-Terms, which is a massive difference when compared to the PVD corpus which needs a vector dimensionality of 850 to find its most robust relationships between the Raynaud phrases and the B-Terms (See Fig. 4.5).

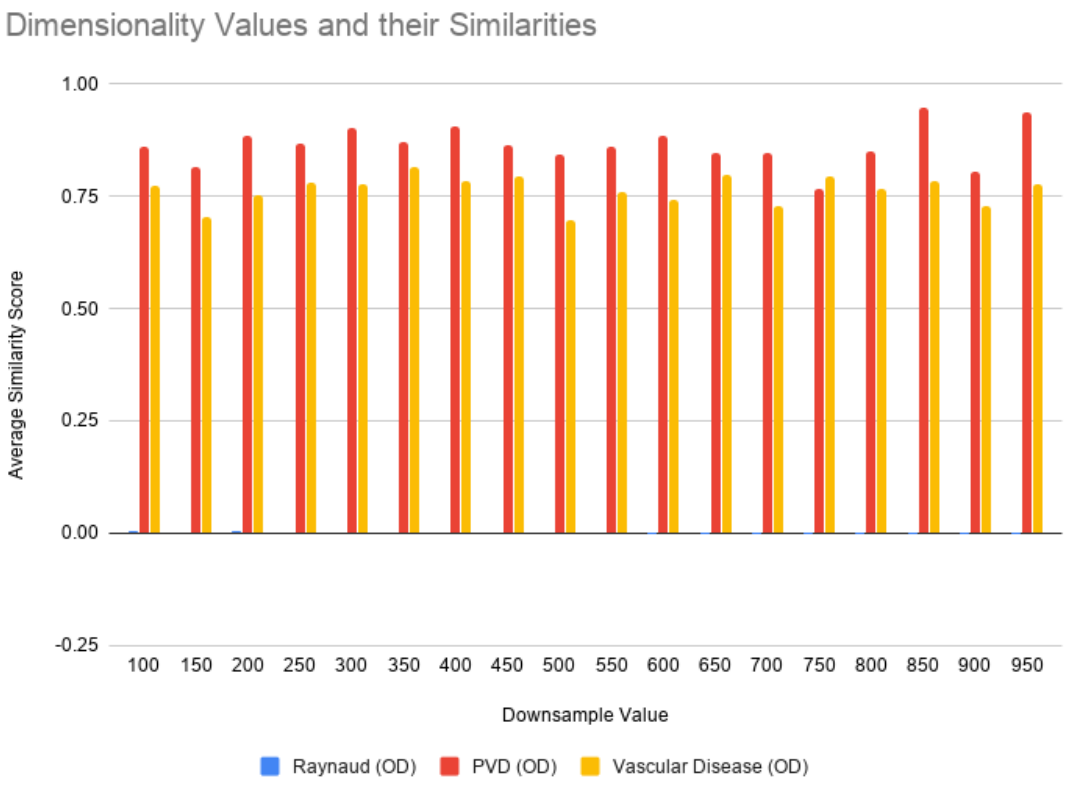

Figure 4.5: Dimensionality Grid Search results for the Open Discovery Corpora

These results do not become any clearer when including the A-Terms in the search, with both the PVD and VD corpora having different results again (Raynaud was not tested due to the lack of A-Terms found), 750 and 450 respectively (See Fig. 4.6).

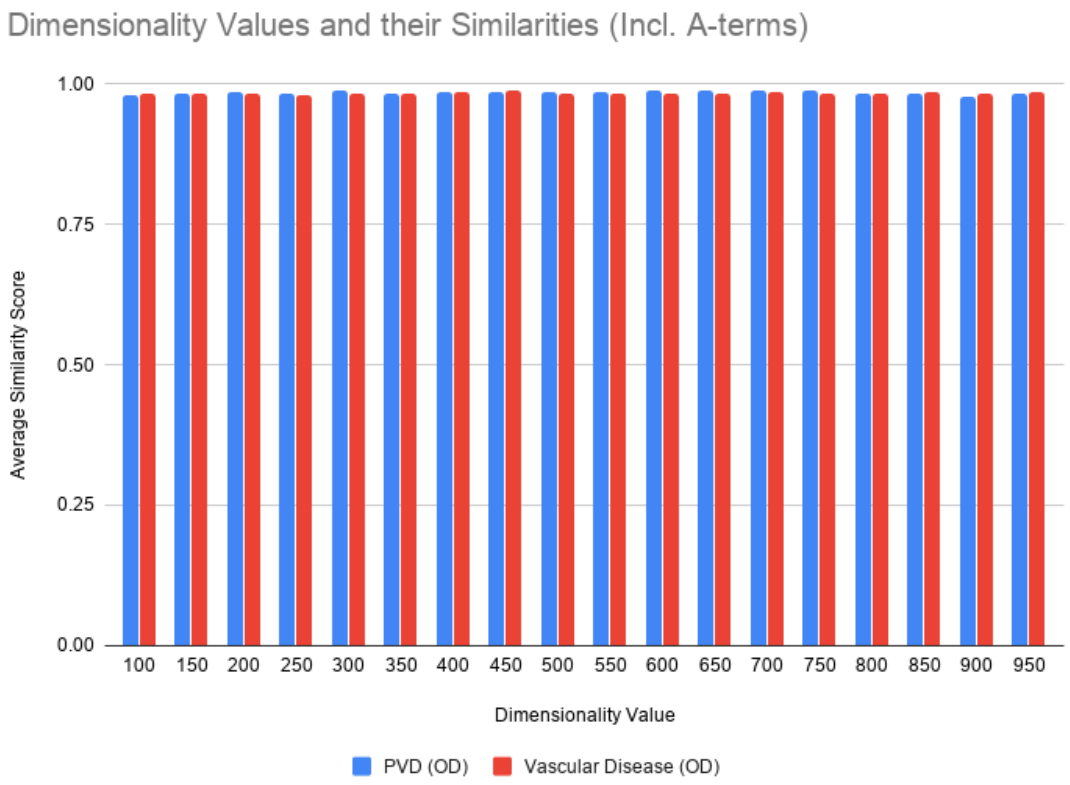

Figure 4.6: Dimensionality Grid Search Results for the Open Discovery Corpora Inclusive of A-Terms

Closed Discovery

The findings in the open discovery experiments for this parameter struggled to find any one individual strongest parameter and whilst the closed discovery experiments also do not indicate an exact correct value they do help determining a seemingly best performing range of between 700 and 950 with all but one corpus, the under-performing Peripheral Vascular Diseases, having an optimal dimensionality between this range (See Fig. 4.7).

When the A-Terms are included in this experiment the two closed discovery corpora where free text A-Terms are found the best, PVD and Vascular Diseases the results show that the best performing dimensionality values are 950 and 450 respectively (See Fig. 4.8).

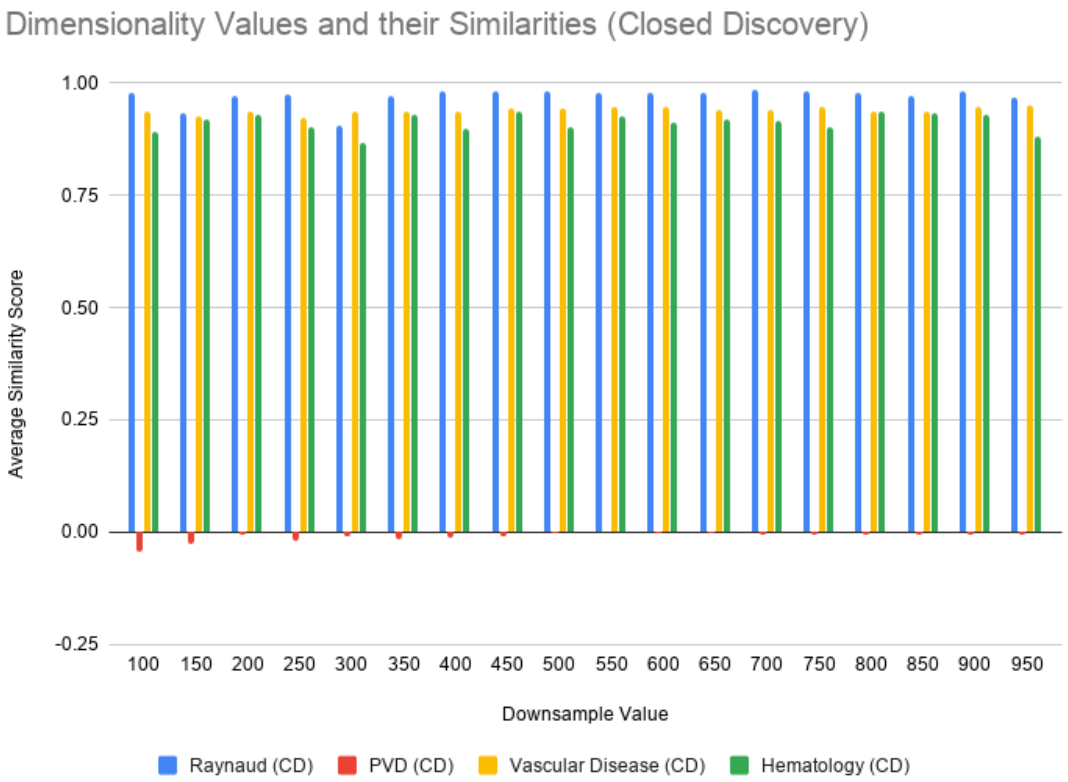

Figure 4.7: Dimensionality Grid Search results for the Closed Discovery Corpora

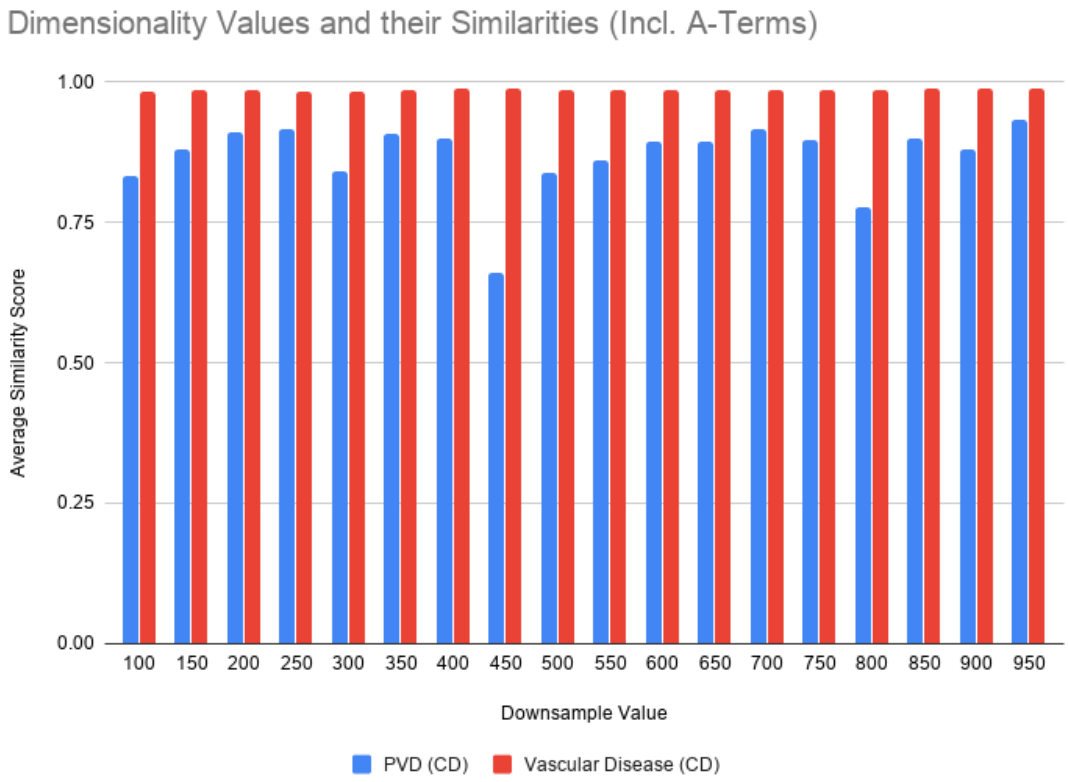

Figure 4.8: Dimensionality Grid Search results for the Closed Discovery Corpora when A-Terms are included

4.1.3 Epoch

Open Discovery

To find the optimal number of epochs used by the model to train there were four different options were tested ten, twenty-five, thirty, and fifty epochs. These numbers were chosen for a few reasons, with one of the main being that one of the downfalls in word2vec being that more epochs alone cannot solve the problems found in a smaller corpus (Gu et al., 2018). As seen in Figure 4.9, this is not necessarily the case for the open discovery experiments detailed in this thesis.

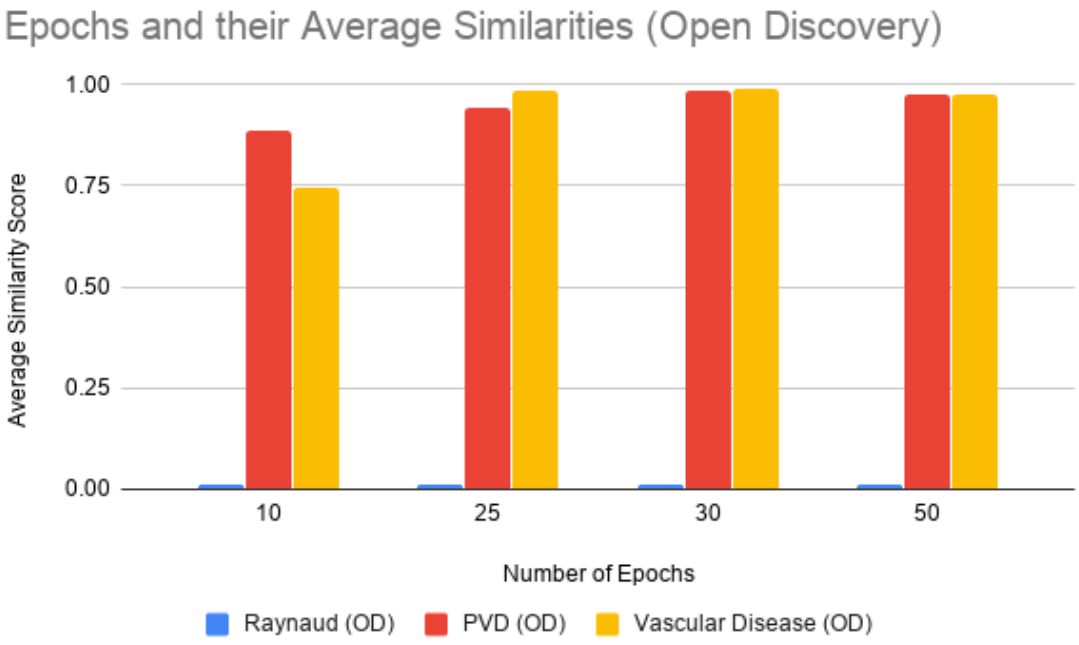

Figure 4.9: Graph showing epoch grid search results

As the above graph shows, the general consensus for the optimal number of epochs the models need to run for is usually on the higher side of the potential values. One important factor to note is that the smallest corpora by far needs the most epochs to get its best results whilst the other two corpora actually performed better with less epochs. However, when the A-Terms are included within the equation the optimal found epoch rate for both the PVD and Vascular Disease corpora is to up all epoch rates to 50. Due to this, the final experiment will be ran with an epoch number of 50.

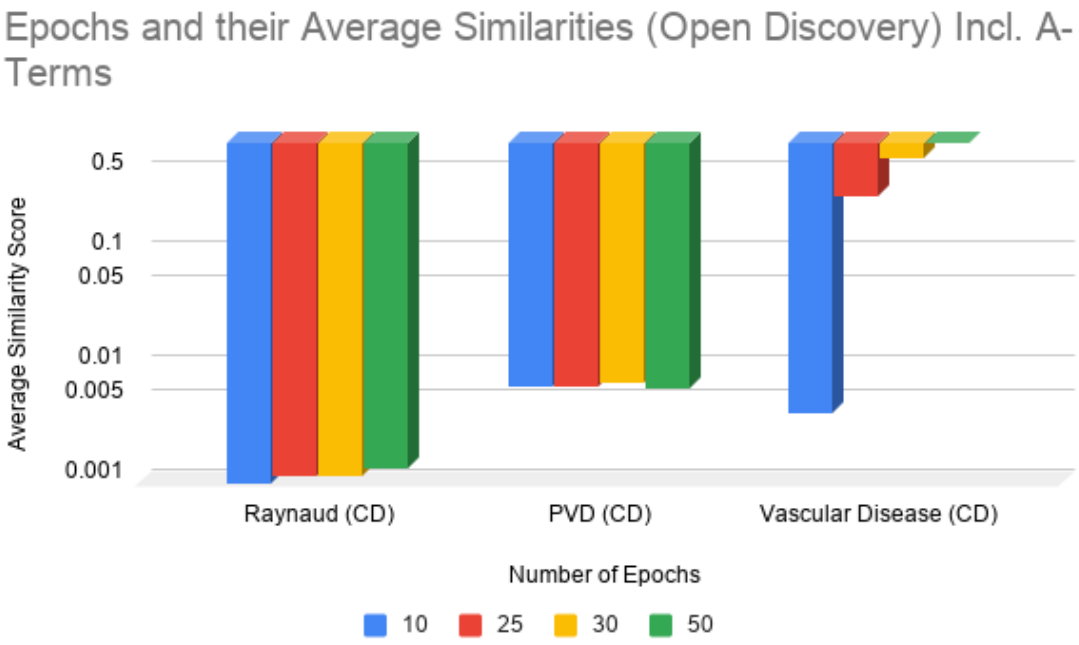

Figure 4.10: Average Similarity of each model inclusive of A-Terms

Closed Discovery

One thing that Figure 4.12 shows is that the more data in a corpus the less epochs that are found to be necessary with the smaller Raynaud corpus performing best with 50 epochs but the Hematology corpus performs at its best when only ran for one epoch. This is likely due to the fact that the amounts of data in the Hematology corpus is starting to contain an acceptable level for a Word2Vec model thus requiring fewer run throughs.

This figure of 50 epochs also stands for the Raynaud / Vascular Disease corpora when inclusive of A-Terms whereas the PVD Corpus was found to run best with 30 epochs. Due to the fact that the majority of corpora run best with 50, this is the figure that shall be used.

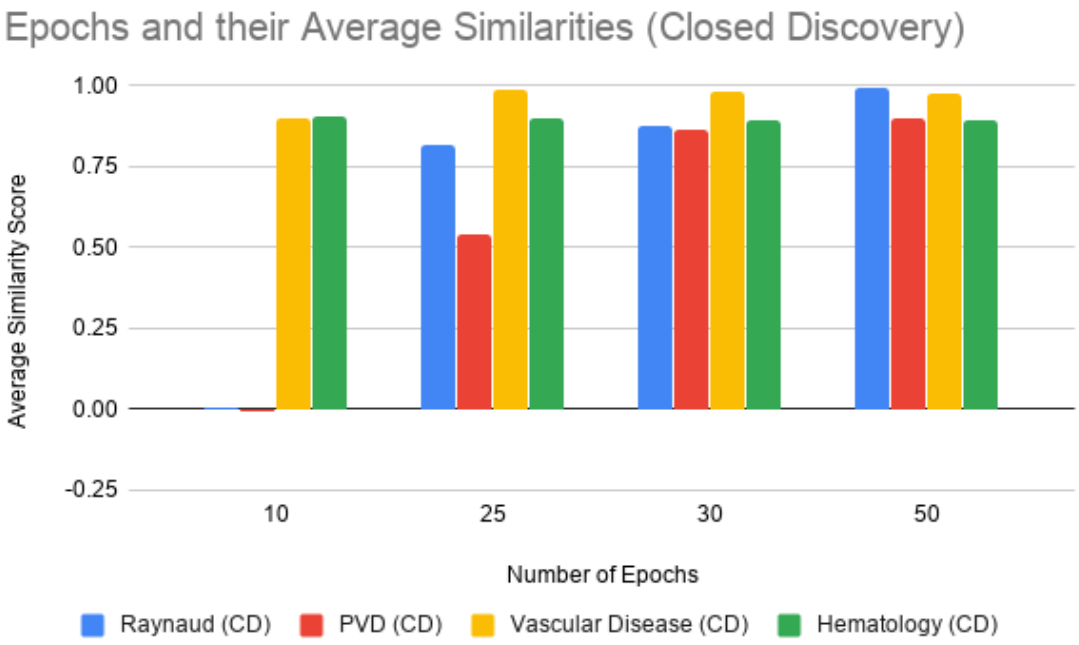

Figure 4.11: Result of the Closed Discovery Epoch Grid Search

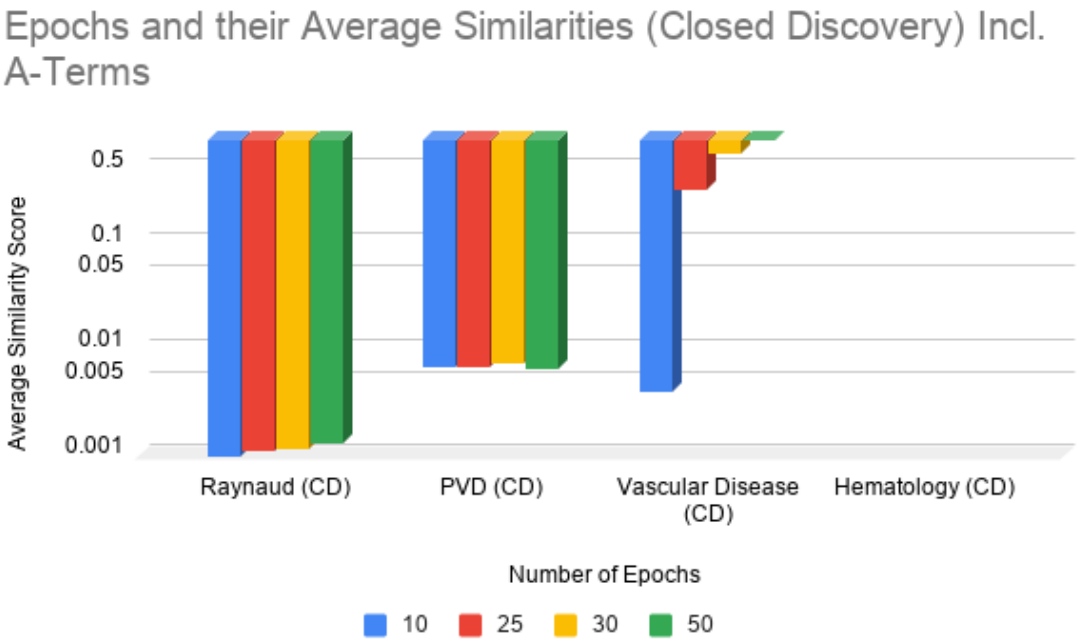

Figure 4.12: Result of the Closed Discovery Epoch Grid Search

4.1.4 Learning Rate

To find the optimal value for the learning rate models were created using all learning rates in a range from 0.01 to 0.1 with jumps of 0.05. Whilst these experiments did not find a huge increase in any learning rate value for any of the corpora it did show that certain learning rates do perform better than others. However, one thing these results did show is exactly how much of an impact the wrong learning rate can have on all corpora.

Open Discovery

As mentioned previously, none of the used open discovery corpora have been found to have a grouped optimal value for the learning rate, which is shown in figure 4.13. However, one interesting thing is how all three corpora have a matching worst learning rate, a value of 0.01. As mentioned in Gu. et al. 's paper, those words earlier in the corpus usually have a higher learning rate and thus have a greater impact on the word vectors (Gu et al., 2018). This is likely due to the smallest of the three corpora, the Raynauds set of documents, performs better with the largest corpora. As mentioned previously, while all results don't have a top-performing result they do have some similarities at the other end of the table with both the PVD and Raynaud corpora's bottom three performing results including a rate of 0.015 and all three corpora's worst performing learning rate was a value of 0.01. The full results for this experiment are shown in figures 4.13 and 4.14:

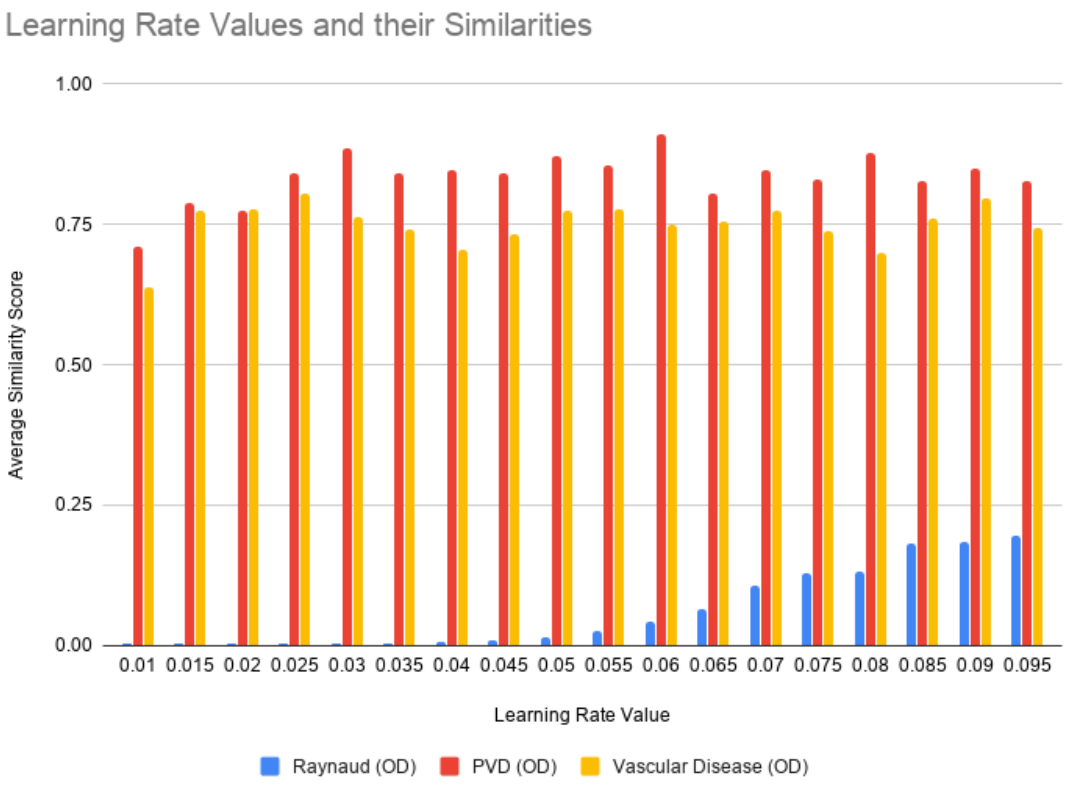

Figure 4.13: The Average Similarity Score for each Learning Rate Experiment using the Open Discovery corpora

As is shown in the above data the best performing result is the PVD dataset when run with a learning rate of 0.06 which achieves an average similarity score of 0.91 As shown in figure 4.13 this corpus has the highest range in average values of 0.201 compared to the lowest range of the Raynaud corpora which is 0.0033. However, it should be noted that whilst the PVD range is the highest so is its minimum result whereas Raynaud which technically has the most consistent results has the lowest best performing score. As is shown in Figure 4.14 the optimal learning rate for the PVD corpus whilst including the A-Terms is a smaller learning rate of 0.01 which is the exact same for the vascular disease open corpora.

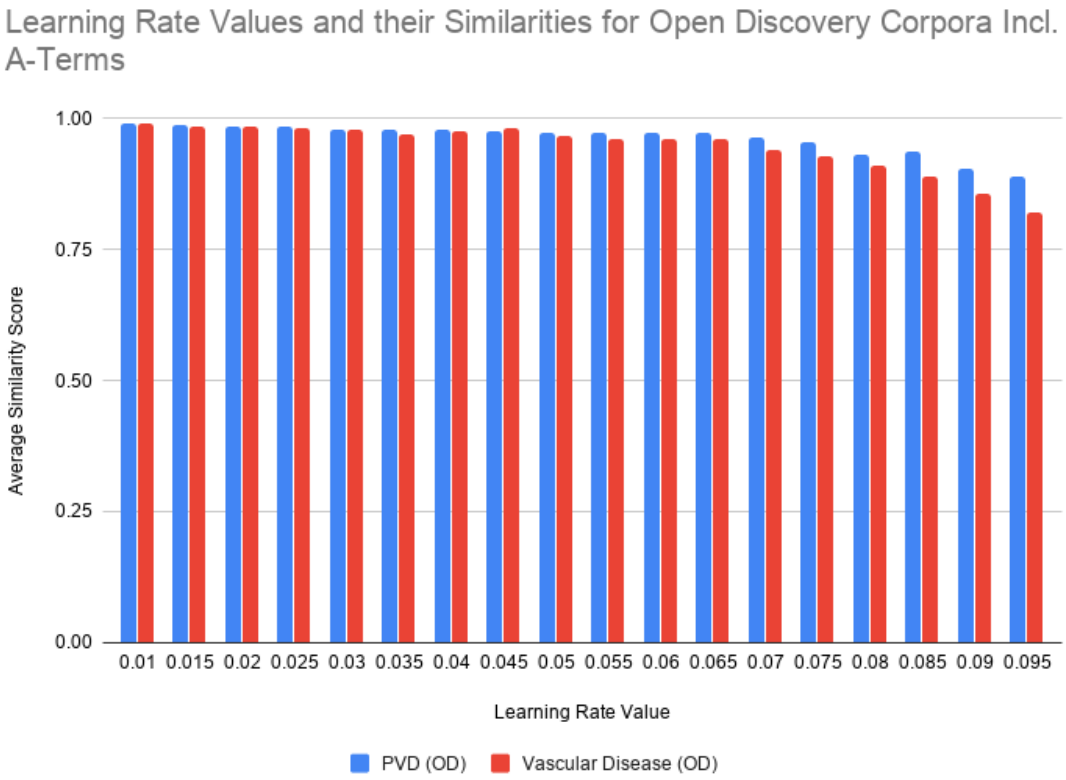

Figure 4.14: Learning Rate Open Discovery when Inclusive of A-Terms

Closed Discovery

When the same experiments were run for the Closed Discovery corpora it was also found to not include an optimal parameter for all four corpora used.

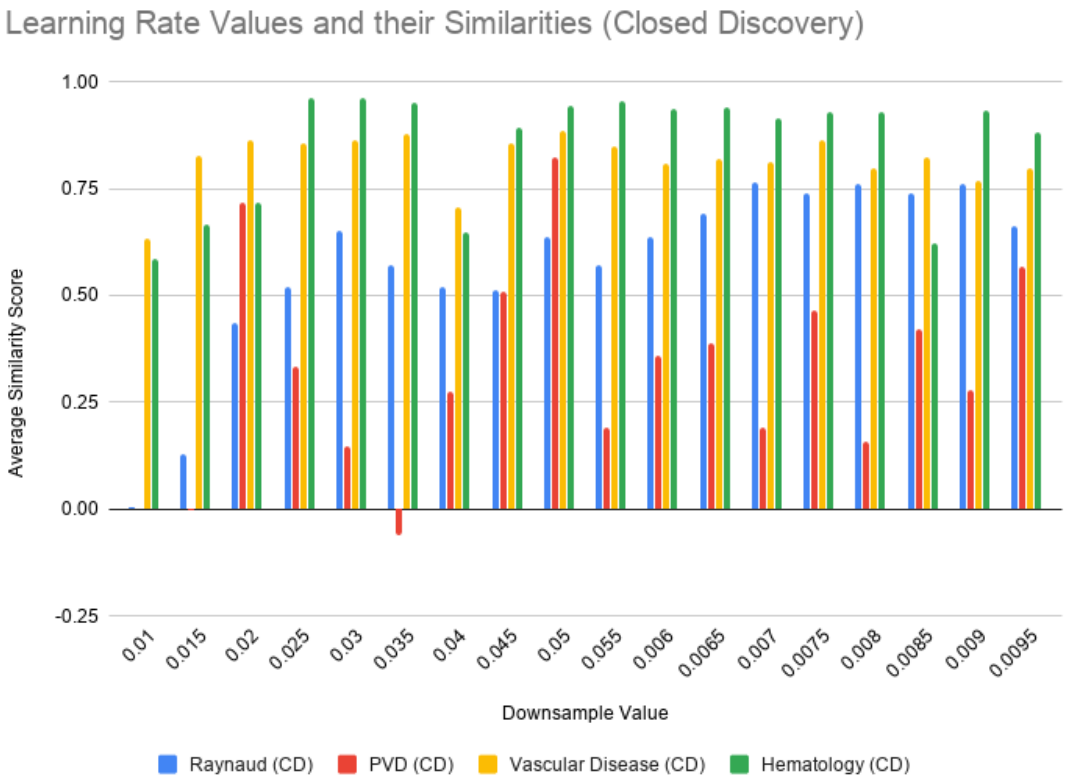

Figure 4.15: Learning Rate Grid Search Results for the closed discovery corpora

Furthermore, unlike the previous experiments there was found to be an optimal value to be used by half of the corpora, 0.05. Which was shared by both the PVD and Vascular Disease corpora as is shown in figure 4.15. The learning rate of 0.05 also performs very well for the Haematology corpus where it is the fourth best performing corpus. This is also the case for the worst performing learning rate, three of the closed discovery corpora (Raynaud, Vascular Disease and Haematology) operate worst with a learning rate of 0.01 whereas the PVD dataset performs worst with a learning rate of 0.035. Since 0.05 has been found to be optimal for half of those experiments in the closed discovery experiments and that 0.06 was found to perform well a value of 0.05 will be used for the learning rate. When inclusive of A-Terms the worst performing learning rate is stuck at 0.01 for two of the corpora where A-Terms are

found, Raynaud and PVD however the worst for the VD corpus has been found to be 0.095 (See Fig. 4.16).

Figure 4.16: Learning Rate Grid Search Results for the closed discovery corpora when inclusive of A-Terms

4.1.5 Downsampling

A grid search was run on a total of 7 different possibilities to make sure that the optimal downsampling value was found, these options were increments from 1e-03 up to 1e-09.

Open Discovery

Out of the seven models that were generated and tested for this experiment the best performing open discovery model was the PVD corpus downsampled by a factor of 1e-08, interestingly all three of the Open Discovery corpora were downsampled within one level of scientific notation, 1e-08 for PVD and Vascular Disease and 1e-07 for the Raynaud corpus.

When these same experiments were ran with the inclusion of found free-text A-Terms the results were extremely similar (See Fig. 4.18)

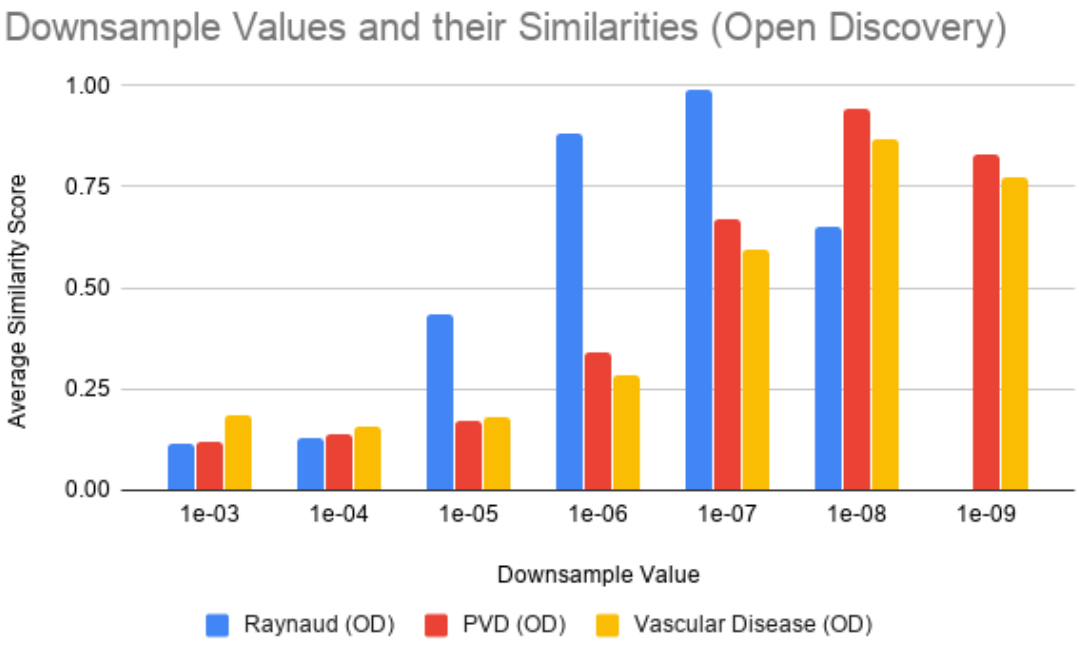

Figure 4.17: The Average Similarity of The Open Discovery Corpora with each Down-sampling Option

This graph also shows the same dominance found in initial open discovery experiments with both the PVD and Vascular Disease corpora tested having a best performing downsampling value of 1e-08 and because of this the final experiment will be ran utilising this down-sampling value.

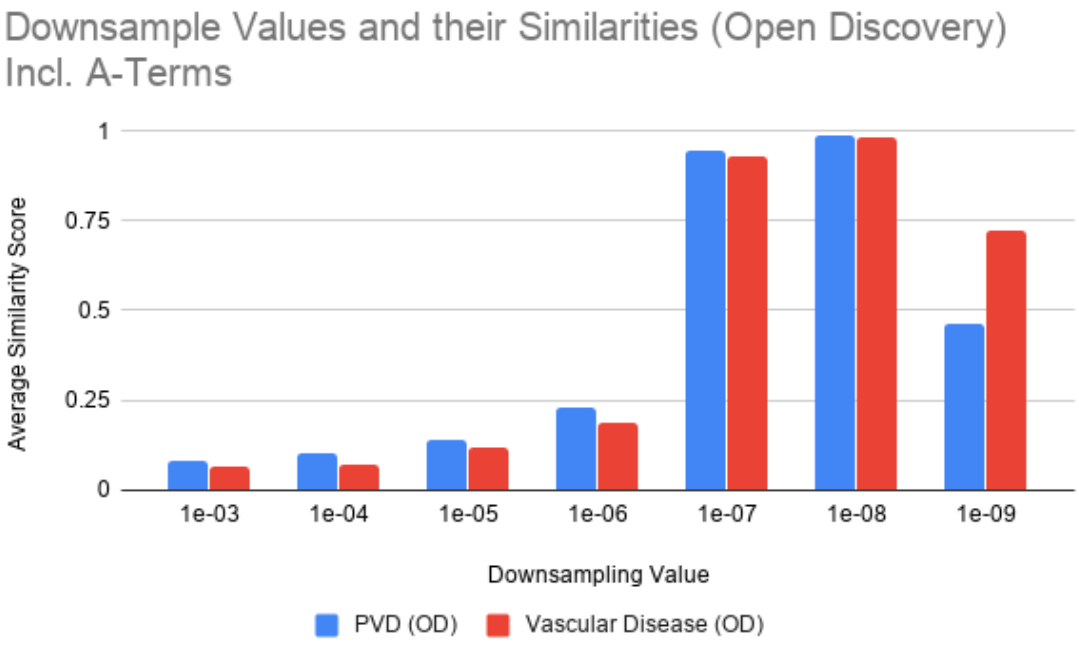

Figure 4.18: Average Similarity Score of Open Discovery Corpora when taking into account the 5 A-Terms found

Closed Discovery

The results of the closed discovery experiments were interestingly very similar to those of the open discovery experiments with all four of the corporas best results being found within the same three optimal values found. However, whilst instead of the fact that the 1e-08 value being the most common optimal value in the larger closed discovery corpora it has been found that the most common words need to be downsampled to a level of 1e-09 which is likely to make up for the total number of terms being increased.

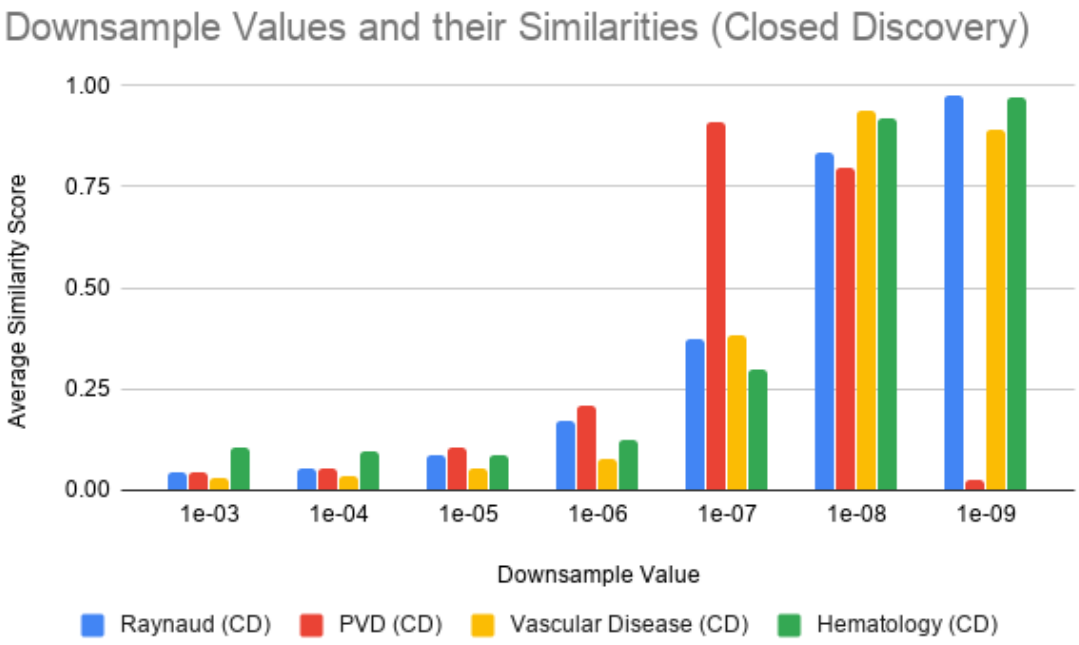

Figure 4.19: The Results for the Downsampled Grid Search performed on the Closed Discovery Corpora

What should be noted as displayed on the above graph is the finding that there has been not one corpus that performs better with a downsampling rate of lower than 1e-07. With the Vascular Disease having a almost sequential increase of its downsampling value in each experiment until it peaks at 1e-08 and the PVD corpus performing having its second best performance at 1e-08 followed by its worse performance at 1e-09. With results in the closed discovery corpora once again being very similar, these experiments have found that the best performing results in the closed discovery experiments when taking into account the A-Terms found is using a down-sampling value of 1e-08.

What these two experiments demonstrate is how important an optimised down-sampling value can be with a value too high or too low being able to have a massive effect on the performance of a model as shown with the gap in performance when comparing 1e-03 and 1e-09 for open discovery and the performance of the closed discovery PVD model where it drops between 1e-08 and 1e-09.

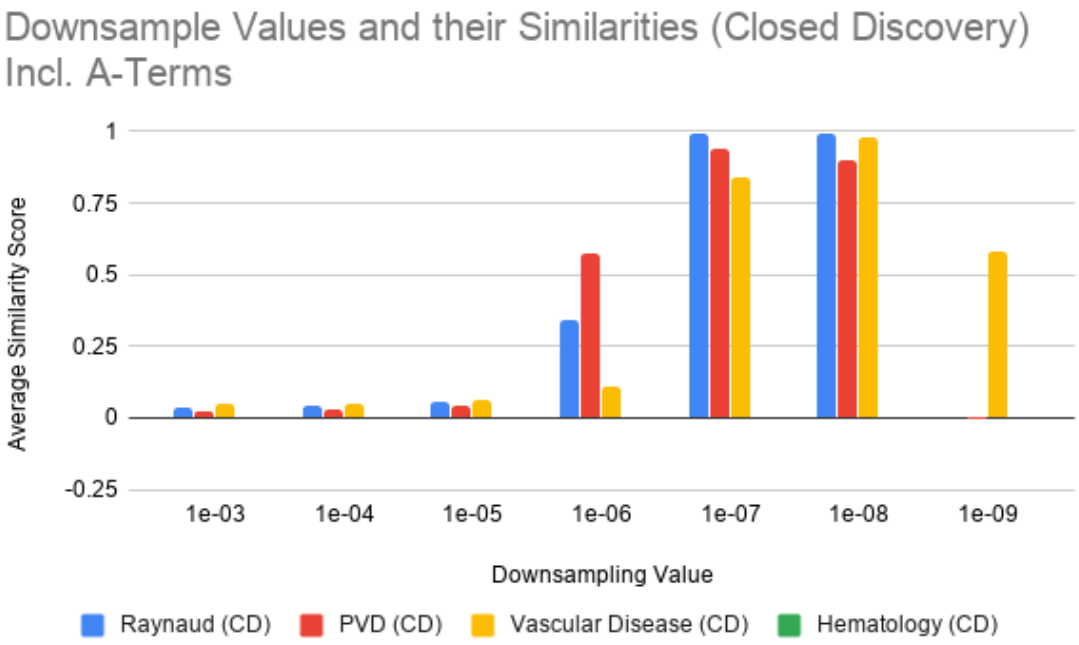

Figure 4.20: The Results for the Downsampled Grid Search performed on the Closed Discovery Corpora Including A-Terms

4.1.6 Context Window

To find the optimal context window size in all experiments a grid search was ran with the following parameters - 1 word, 10 words and 30 words.

Open Discovery

As with most of the grid searches performed in these experiments there is not one outright best result, however a context window of just one word produces the best results as shown in Figure 6:

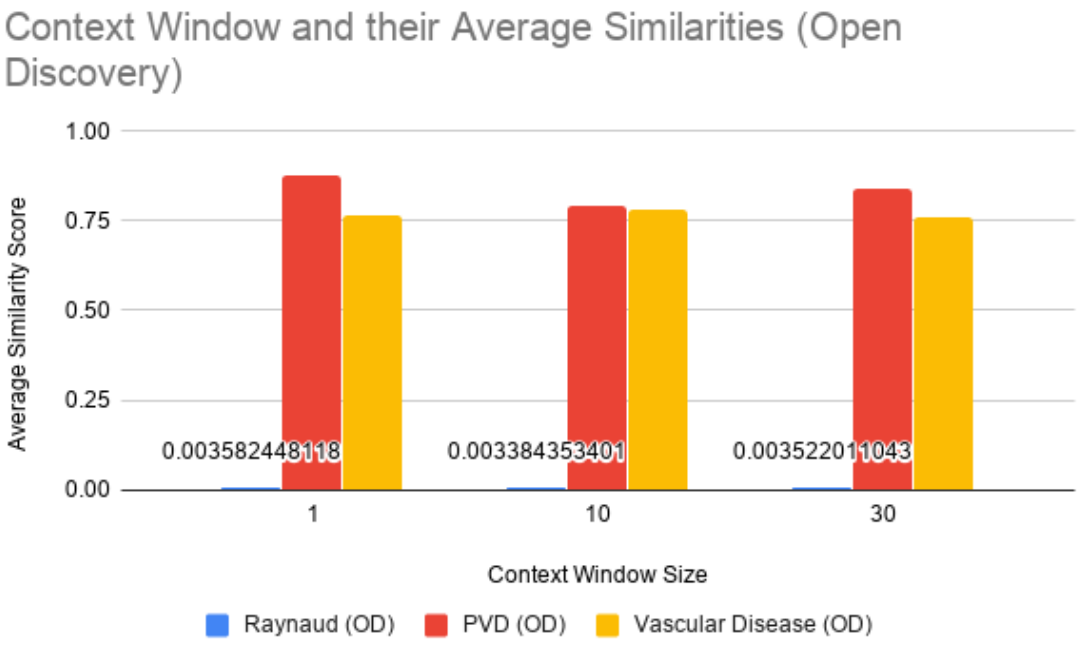

Figure 4.21: Graph showing the results for the Open Discovery Experiments

As the graph in Figure 4.21 the most consistent performing context window size for an open discovery corpora is a window that only takes into account the n-grams directly surrounding each word by using a context of only one. The table shown in Figure 6 allows easier viewing of the Raynaud's results when in comparison to the other corpora. This open discovery model was then rerun on the two corpora where A-Terms were found, PVD and VD with one and four terms found respectively, the results were similar with the only change being the PVD corpus actually having a slightly higher average on the 10 word context window (See Fig. 4.22).

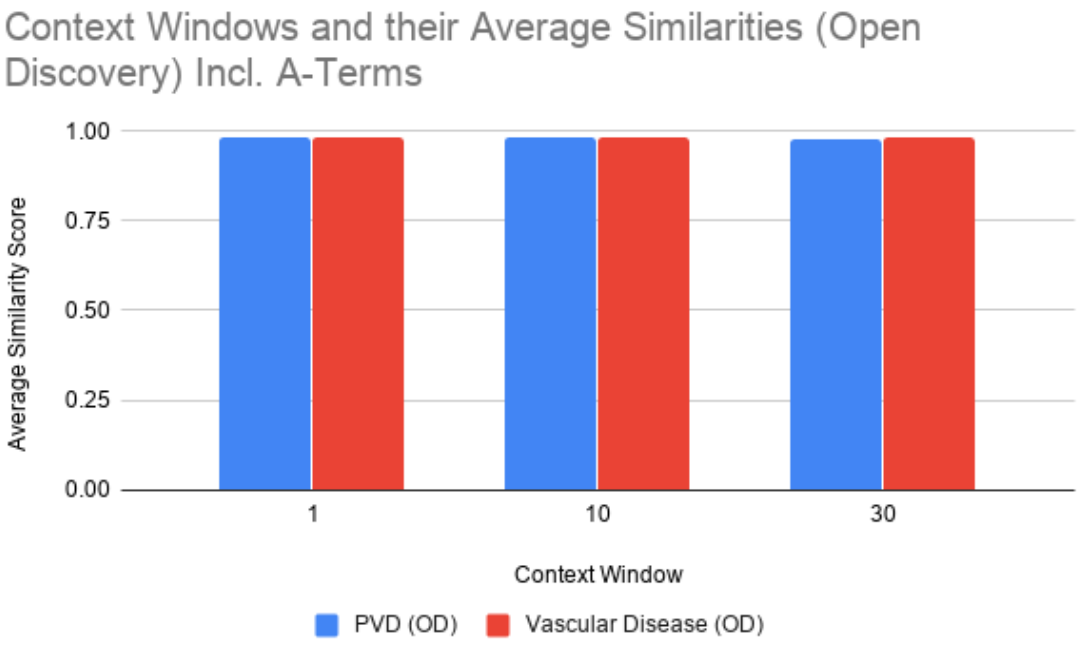

Figure 4.22: Graph showing the results for the Open Discovery Experiments when taking into account the found A-Terms

Closed Discovery

Whilst the open discovery experiments have been found to utilise a much smaller context window for its experiments it has been found that the closed discovery experiments perform better with a large window.

As seen in Figure 4.23 instead of the much smaller windows performing better they actually have the worst performance for all four corpora with the two larger options having the best performance with an optimal performance value of ten. This is because a larger context window means the model has a better ability to detect semantic relationships instead of a smaller context window which uses just terms that occur in close proximity to the main word which has more of a benefit to closed discovery relationships because the desired A-Terms will very rarely appear in the context of Raynaud's Disease since the link had not yet been formed.

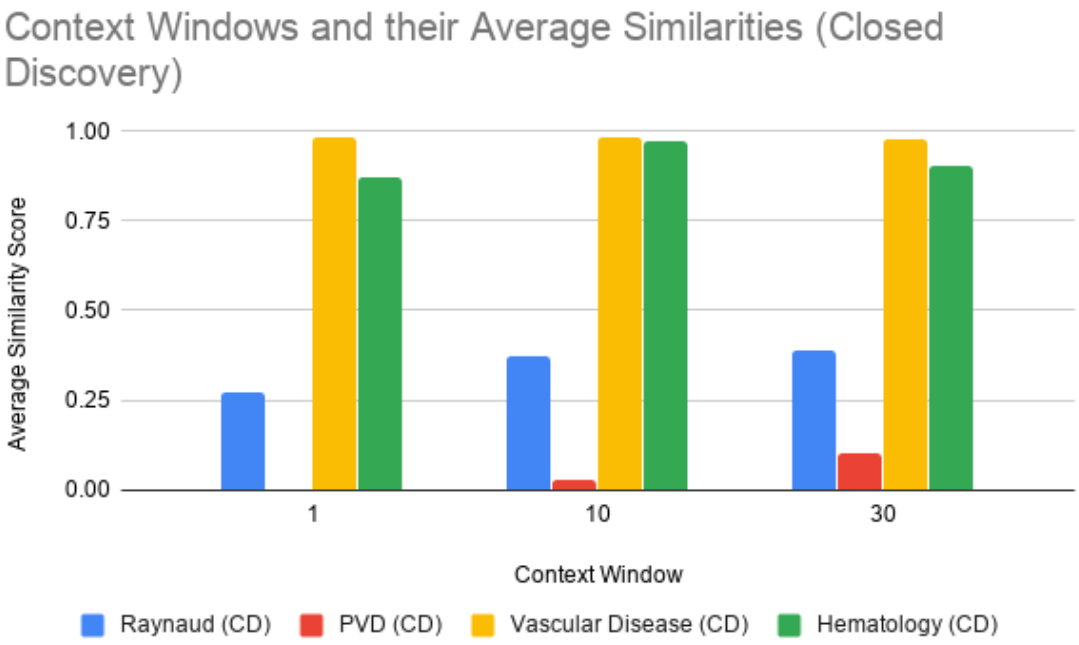

Figure 4.23: Graph showing the results for the Closed Discovery Experiments when taking into account the found B-Terms

When inclusive of A-Terms the closed discovery results are completely different with the most optimal results being a context window of 30 for the PVD and Vascular Disease corpora with the Raynaud corpus proving to have an optimal value of 10. As this is very similar to the closed discovery results without the A-Terms included the optimal model will include a context window of 30.

4.1.7 Minimum Word Count

The minimum word count of a Word2Vec model is used to exclude any words that can be seen as too rare to appear in a corpus. Due to this fact the importance of selecting a correct value cannot be understated with it being too high it could exclude some potentially valuable information, but being too low could potentially include words that appear only once and thus reduce the quality of the results.

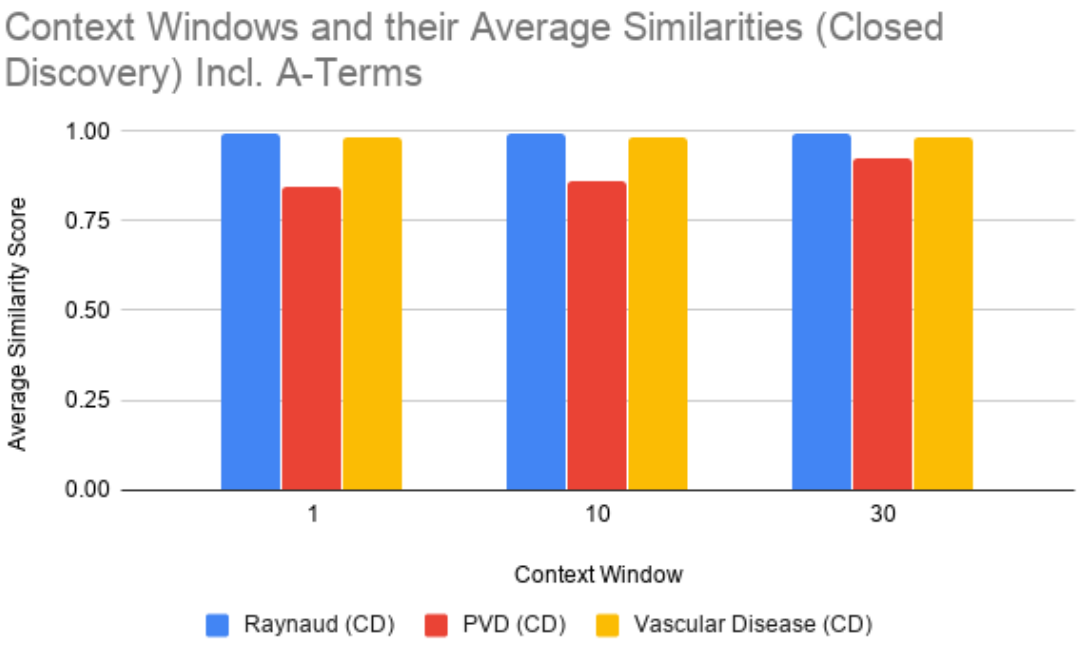

Figure 4.24: Graph showing the results for the Closed Discovery Experiments when taking into account the found A-Terms

Open Discovery

The best performing result for the open discovery corpora was found to be the inclusion of every word that occurs in the corpus. This was likely because the corpora were a lot smaller than those in the closed which meant that the performance of Word2Vec would have suffered due to a lack of large amounts of training data. This is also backed by the finding that the results do get weaker as more words are discarded.

Most corpora have a large drop the more words that are excluded from the corpus (See Fig. 4.25), however it should be noted that the Raynaud and PVD corpora do have a small increase in performance with the jump from 25 to 50 words discarded. When the inclusion of A-Terms were taken into the mix however the results differ slightly with the larger corpora needing more words removed to perform at their optimal level (See Fig. 4.26) as shown by the experiments which found that the best results for the PVD and VD corpora is 10 words and 15 words respectively.

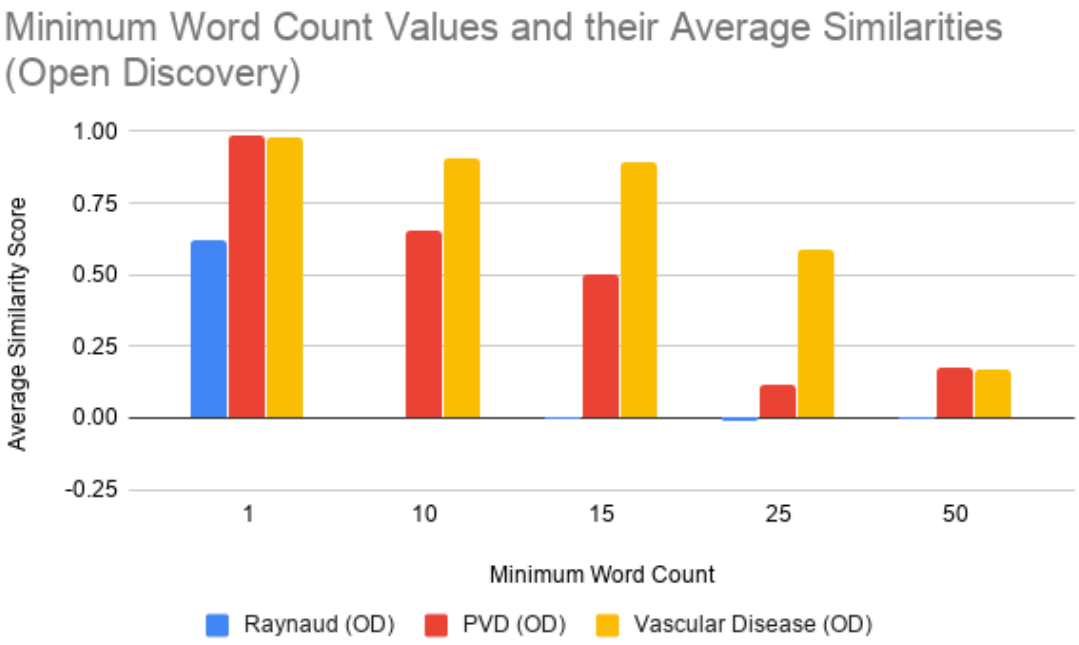

Figure 4.25: Results from the Minimum Word Count Grid Search for the Open Discovery corpora

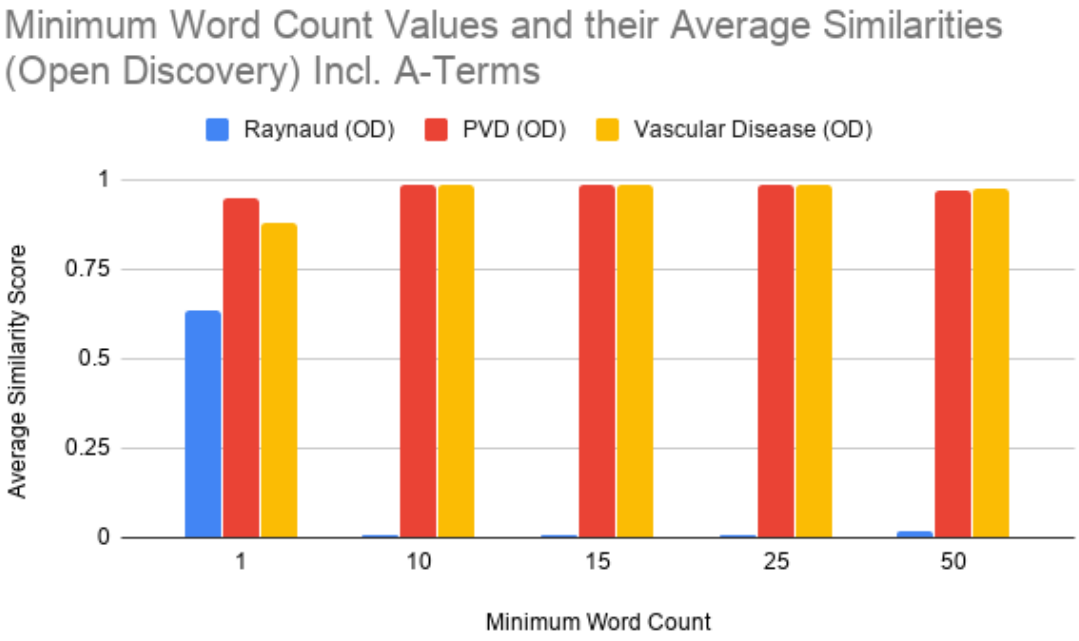

Figure 4.26: Results from the Minimum Word Count Grid Search for the Open Discovery corpora

Closed Discovery

One large difference in the results of these experiments from those undertaken on the Open Discovery is found within the inclusion of the Haematology corpus. During

the experiments on this corpus it is found that only TWO of the potential values, one minimum word and ten minimum words, find any of the potential B-Terms in the corpus. As is shown in Fig. 4.27 the best performing minimum word count on average for this experiment was found to be a context window of one minimum word.

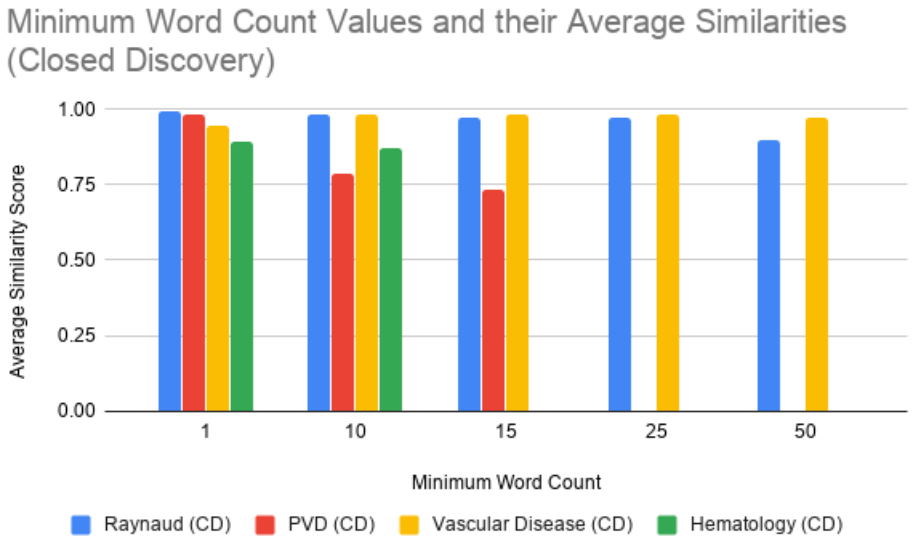

Figure 4.27: Results from the Minimum Word Count Grid Search for the Closed Discovery corpora

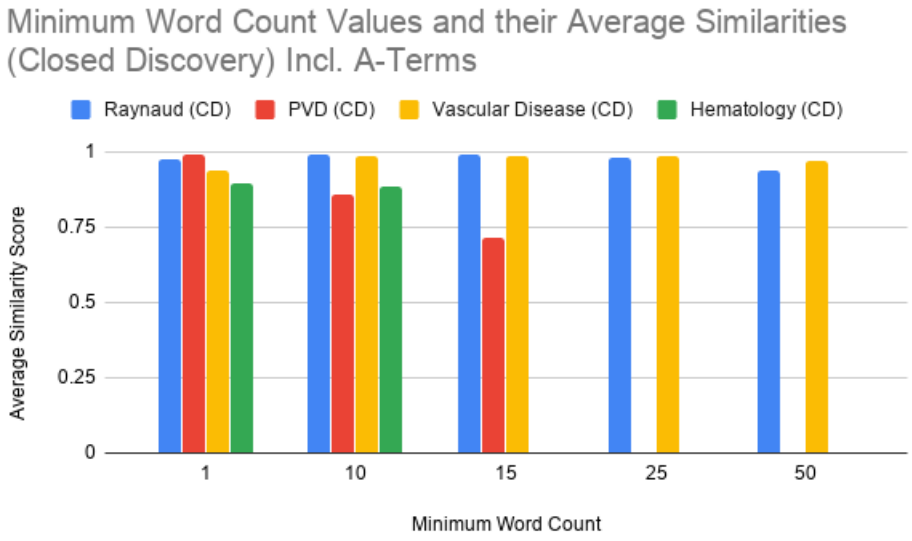

Figure 4.28: Results from the Minimum Word Count Grid Search for the Closed Discovery corpora when including A-Terms

4.1.8 Optimised Models

In this section we will be discussing which results are found to be optimal for each corpus and then discuss the final models that are used for the finished experiments.

4.1.9 Optimised Model: B-Terms

As shown in Section 4.1, the optimal model architecture with a majority in the best performing models was the Continuous Bag of Words architecture which outperformed the skip-gram architecture. The models will utilise a minimum word count of 1 word due to the fact that 71.42% of experiments outlined in Section 4.1.7 performed at their best with this parameter. The next hyper-parameter that has been set is a context window. However, as the graphs in section 4.1.6 show this parameter had extremely inconsistent results when comparing on a corpus by corpus basis with the best performing parameter, a context window of 1 only achieving a success rate of 42.85%. The model will also downsample all terms by a rate of 1e-08 because this performed best for the majority of experiments see Section 4.1.4. This model will also be trained for a total of fifty epochs since this number was found to produce acceptable results (See Sec. 4.1.3).

4.1.10 Optimised Models: A-Terms

Unlike the model described in Section 4.1.9 the best performing model architecture is the skip-gram architecture with a majority of 60% (when disregarding the lack of A-Terms found in the Raynaud open discovery or Haematology closed discovery corpora). However, as also displayed in this research the optimal value for a minimum word count when looking for A-Terms is found to be 10 with a minimum word count of 1 also performing well. When deciding upon a set learning rate there is a strong case for a learning rate of 0.01 as this is optimal not only for two of the open discovery corpora (See Fig. 4.14 but it also the optimal value for one of the closed discovery corpora. There are two joint top performers for the Minimum Word Count parameter, and that is the usage of both one minimum word and also ten minimum words. To allow for simplicity at the model generation stage, the

minimum word count used will be ten words to help minimise the differences in the models. This is because a comparative test that was run of all minimum word count options alongside all context window size options the best performing results for all experiments including A-Terms had a minimum word count of 10 and a context window of 1 respectively. All words will be downsampled by a factor of 1e-08 as this was found to be successful in 80% of experiments (See Section 4.1.4). As described in Section 4.1.3 when including the need to find A-Terms in the text the optimal number of epochs is 50 with only one corpus, the Peripheral Vascular Disease corpus, requiring less than this (30 epochs).

As seen in both Sections 4.1.9 and 4.1.10 there are certain parameters that are found to perform optimally on both the open and closed discovery methods, through all experimented corpora which has allowed these experiments to find a generic model for potential usage in further experiments. It should however be noted that these experiments do show certain discrepancies and indicate that an exhaustive hyper-parameter search would be preferable for any new experiments, this is due to the slight differences in many different corpora.

4.1.11 Breakdown of Corpora

The graph below contains a number of lexical statistics from each of the three corpora:

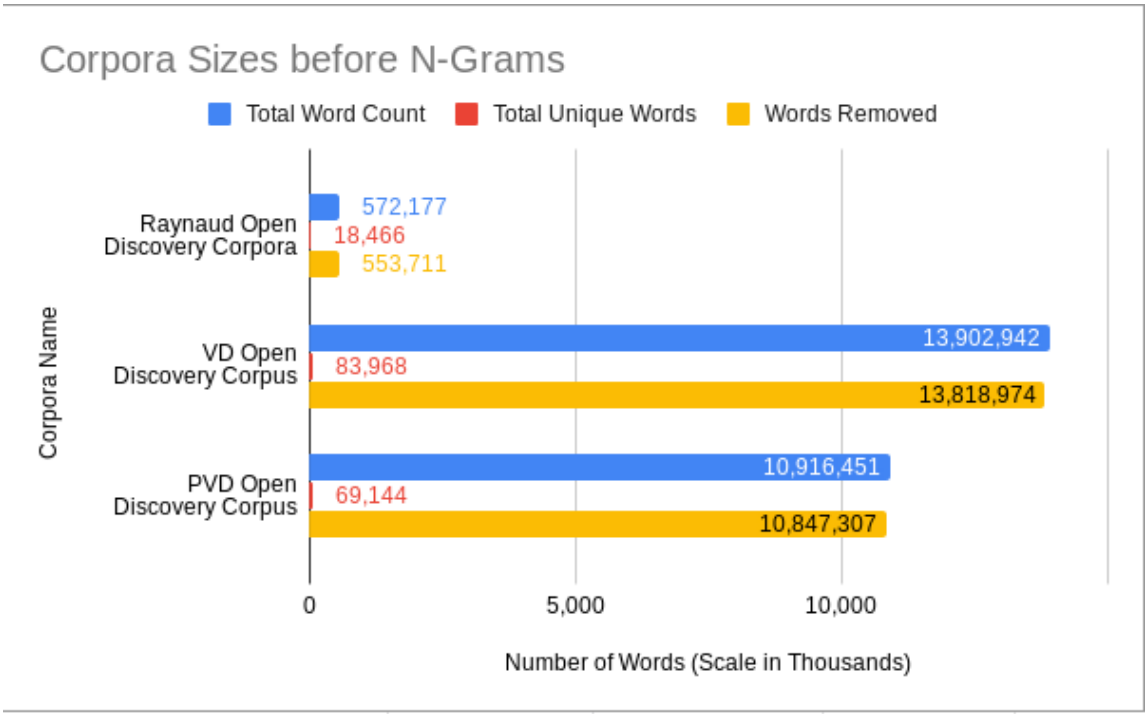

Figure 4.29: Size of the three closed corpora before N-Grams were generated

As is expected the more specific a corpus is, the fewer unique words are found in the entries since there are less entries. This is shown by the decrease of 95% from our largest corpus, comprising of articles on general Vascular Diseases over the corpus comprising of those purely comprising of articles based on the highly specific Raynaud's Disease. Once this idea is taken to the corpus after N-Gram generation, the corpora sizes do have a large increase on generation on bi-grams however this is not continued into tri-gram creation. This would be down to the fact that many of the tri-grams created utilise and thus replace many of the previously found bigrams.

One thing that should be noted is that the decrease in unique trigrams is very close to the decrease shown in unigrams with a 93% decrease from smallest corpus to largest. The closest margin is still quite the separation with the bigram generation having a slightly smaller decrease of 86% in unique term number (See Fig. 4.30).

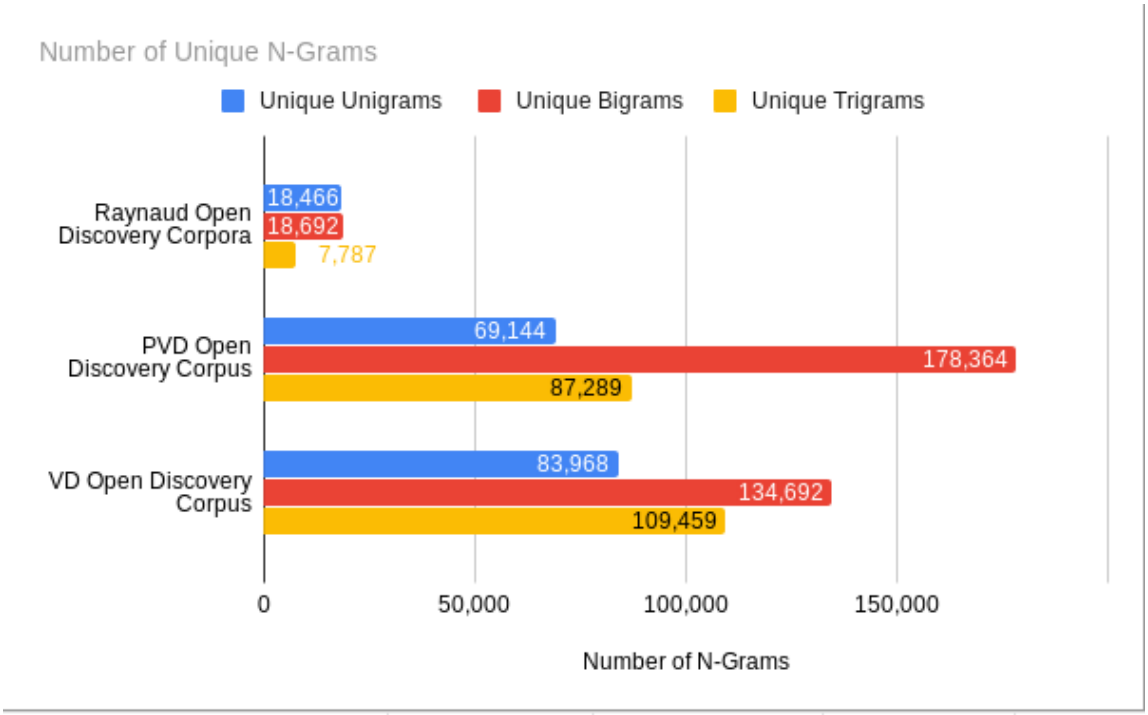

Figure 4.30: Number of Bigrams and Trigrams

As mentioned throughout this research, there are predominately two different types of Discovery in the field of Literature Based Discovery, open and closed. In this section of the thesis there will be a presentation of the results of the experiments taken. Whenever a new corpora was experimented on, there were two parameters that were changed to see the effect they had on the results. These parameters were the minimum word count and also whether we took the most similar words, defined as those with the closest cosine similarity to the target phrases, or whether we took the least similar words, those with the furthest cosine similarity. In this section we will be displaying which of the A-Terms and B-Terms found in the 2001 paper by Weeber, as shown in Appendix 6.1 and 6.2, are also found by the model.

4.2 Open Discovery

As has been mentioned throughout this thesis, the open discovery method requires there to be no A/B literature to be involved in the search. As the search terms have differed between these searches and the closed discovery. For this experiment, there were also three different corpora used in the same vein as those in the closed discovery

with the most specific being only data related to Raynaud's, the second corpora being on Peripheral Vascular Diseases and the third being all Vascular Diseases. In the graph below, we can see how many entries are utilised when the system works with each corpus.

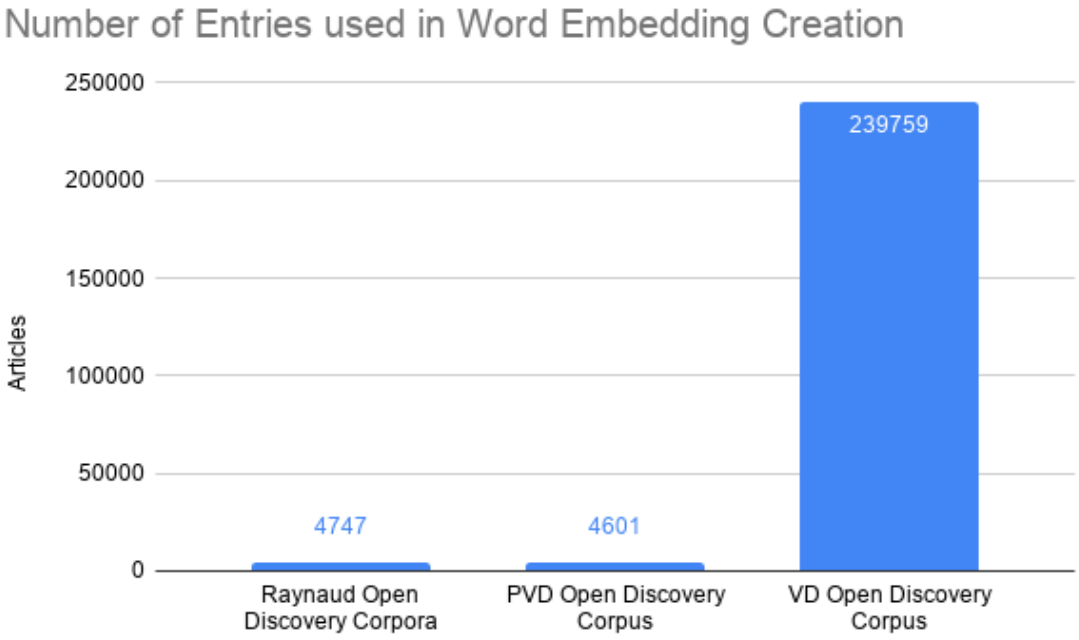

Figure 4.31: Number of Entries used in Word Embedding Creation

As the above graph shows, there are massive increases in dataset size dependent on the specificity of each corpus. When the number of documents increases, it is expected that the number of found significant terms would rise. As the graph below does show, many of these terms were found, and the trend shows that the more data we have utilised, the more of these words that do appear.

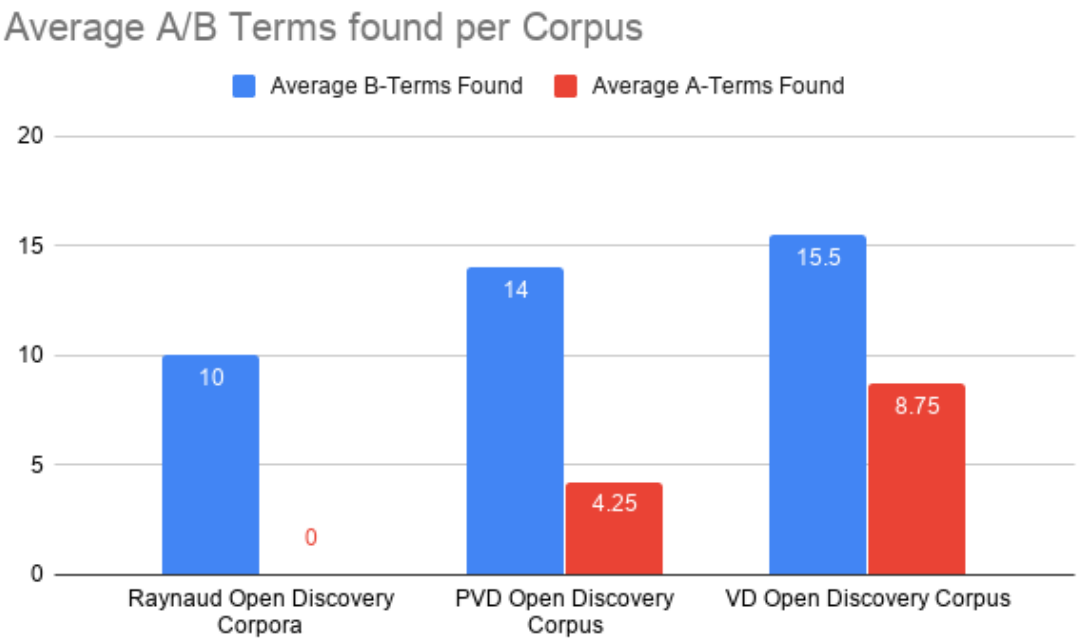

Figure 4.32: The Average number of A/B Terms that are found per Open Discovery corpus.

One thing that was found through these experiments is that certain search terms were found consistently, whereas others were found rarely, if at all. The table below consists of all the significant terms that were looked for and the percentage of corpora they were found within. These tables do not take into account the different parameters used e.g. the varying minimum word count or whether the least or most similar words were used and class a term as found if it appears in any one of those experiments. However, the table does show yellow for those terms which are not found in every variant of that corpus and red for those that never appear in any of them.

As is shown in the above table the larger the corpora the more significant phrases are

	Vascular	Peripheral Vascular	Raynaud
Blood Viscosity			
Platelet Aggregation			
Vascular Reactivity			
Erythrocyte Deformability			
Plasma Viscosity Level			
Hemorheology			
Decreased Vascular Flow			
Hyperviscosity			
Fibrinolysis			
Thrombosis			
Platelet Adhesiveness			
Effects, blood coagulation			
Vasodilatation			
Vasodilation			
Vasospasm			
Vasospasm Mechanisms			
Vasomotion			
Decreased Vascular Resistance			
Total Found:	**94.44%**	**88.88%**	**83.33%**

Table 4.2: Performance of finding the B-Terms of each Open Discovery Corpora

usually found which is likely due to both the number of new entries allowing for new connections to be formed but also due to the inclusion of new phrases completely.

Table 4.3: Performance of finding the A-Terms of each Open Discovery Corpora

	Vascular Disease	**Peripheral Vascular**	**Raynaud Disease**
Fish Oils			
Maxepa			
Fatty Acids, Omega-3			
Omega-3 polyunsaturated fatty acid			
Eicosapentaenoic Acid			
Epa-e			
Cod Liver Oil			
Salmon Oil			
Fatty Acids, Essential			
Dietary Fats			
Total:	80%	70%	0%

As the above tables show the open discovery corpora as a whole perform rather well in retrieving the B-Terms with at-least, one experiment ran per corpus finding most words. There were however terms not found in any corpus at any parameter mix. On the other hand some of these terms, highlight the need for someone with biomedical training to look over the lists generated by programs like this. This need was found especially necessary with the b-term "Vasospasm Mechanism" which in itself may not be seen. However, there are many different mechanisms such as "vasoconstriction" and "oxidative stress" which to somebody within the field are seen to be "Vasospam Mechanisms".

It should be noted that some terms like these have been found in specific experiments for our tool but not counted to keep parity with Weeber's list) which may be missed when someone with the correct knowledge looks at the file. However, while each corpus performs well with the B-Terms, there is a much more significant drop in performance when it comes to finding the A-Terms. As expected there was a drop in all corpora in regards to number of a-terms that appear in each list, what is surprising is the fact that the smaller, more specific corpora, which is based purely on those articles that are found when Raynaud is the search term finds 0 linkage terms between Raynaud itself and Fish Oil. As shown below in Table 4.4 out of a maximum 15 a-terms the highest average number found was 8.75 which equates to

58% of the a-terms found, note these averages do not show the total number found per experiment.

Raynaud Disease	Peripheral Vascular Disease	Vascular Diseases
0	4	8.75

Table 4.4: The average number of A-Terms found per Corpus

As is to be expected with these experiments the number of documents has a large effect on the total number of significant phrases found in each experiment with the trend being that the more documents in the corpus the more likely one of the A-Terms defined by Swanson and Weeber are to be found.

4.3 Closed Discovery

As seen in Table 3.1, for this experiment there were three different corpora generated all with a different size and specificity, ranging from Raynaud's Diseases all the way up to general hematological journals. The difference being that because this is a closed discovery the corpora included all data published on the found Fish Oil terms between 1960 and 1986. The experiments that were undertaken for this corpora were the same as those previously detailed however the results did differ.

Utilising the same method as in Section 4.2 we found that the closed discovery corpora had the following numbers of A/B-Terms. Bear in mind it should be noted that the hematology corpus did not find any Raynaud terms in its experiments with a 15 minimum word count so their was no relationships detected.

The average number of A/B-Terms found per the first two smaller corpora does see an increase when utilising the closed discovery corpora. The largest increase will be an increase of 11.75 when taking into account the fact that the average A-Terms found in the open discovery raynaud corpora was zero but is now found to be an average of 11.75. There is also an increase of 3.25 for A-Terms when comparing the two PVD corpora. It should be noted that all three corpora experience a decrease in B-Terms found. This is experienced most dramatically with the Raynaud and PVD corpora which experience a drop of 5 terms dropped on average with the VD corpus

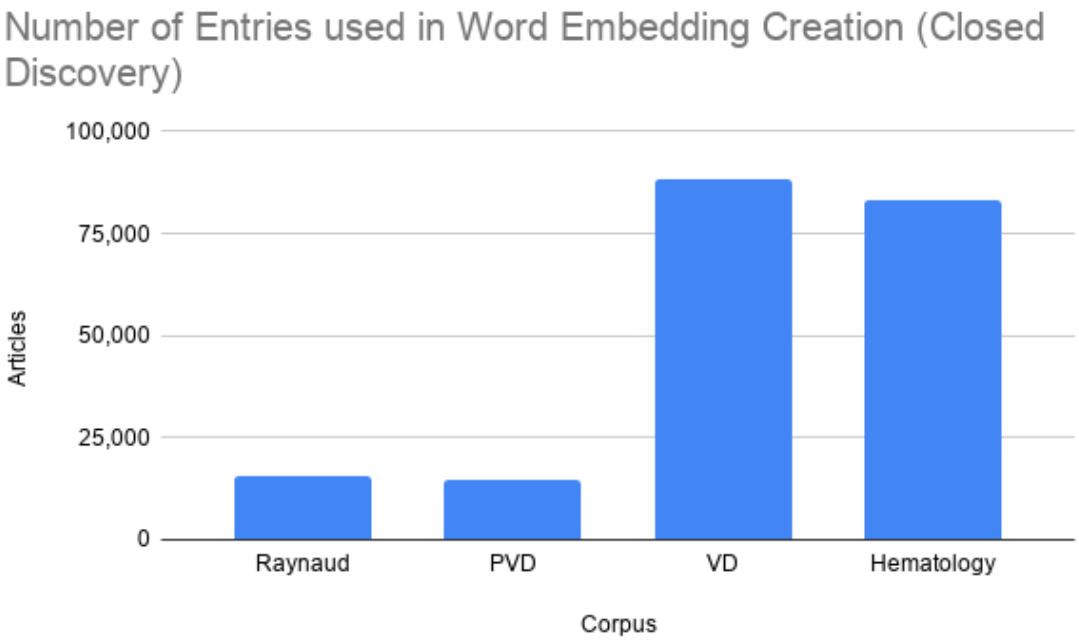

Figure 4.33: Number of Entries used in Word Embedding Creation for the Closed Discovery

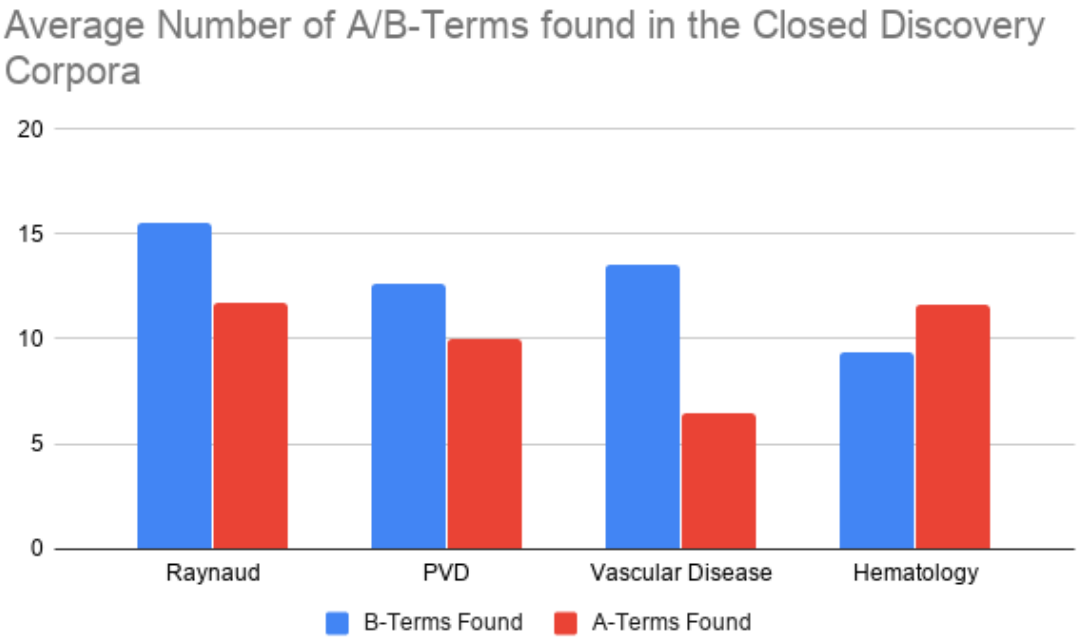

Figure 4.34: The Average Number of A/B Terms found in the closed discovery corpora

has experienced a smaller decrease of 2 B-Terms, however that these results will be skewed as that there were also no Raynaud terms found in the 15 minimum word

count least similar experiments. The number of A-Terms found in these experiments has massively increased from finding no A-Terms in any of the Raynaud experiments to every A-Term being found in atleast one experiment, See Table 4.5. Nonetheless, the number of A-Terms found in the Vascular Disease corpus has dropped by 10% and the PVD corpus has experienced an increase of 20%. The new corpus included in these experiments, the hematology corpora performed exceedingly well finding all A-Terms in at least one experiments with 70% of terms being found in all experiments.

	Raynaud Disease	Peripheral Vascular	Vascular Disease*	Hematology*
Fish Oils				
Maxepa				
Fatty Acids, Omega-3				
Omega-3 polyunsaturated fatty acid				
Eicosapentaenoic Acid				
Epa-e				
Cod Liver Oil				
Salmon Oil				
Fatty Acids, Essential				
Dietary Fats				
Total:	100%	90%	70%	100%

Table 4.5: Performance of finding the A-Terms of each Closed Discovery Corpora

	Raynaud	Peripheral Vascular	Vascular*	Hematology
Blood Viscosity				
Platelet Aggregation				
Vascular Reactivity				
Erythrocyte Deformability				
Plasma Viscosity Level				
Hemorheology				
Decreased Vascular Flow				
Hyperviscosity				
Fibrinolysis				
Thrombosis				
Platelet Adhesiveness				
Effects, blood coagulation				
Vasodilatation				
Vasodilation				
Vasospasm				
Vasospasm Mechanisms				
Vasomotion				
Decreased Vascular Resistance				
Total Found:	94.44%	66.66%	100%	61.11%

Table 4.6: Performance of finding the B-Terms of each Closed Discovery Corpora

Chapter 5

Discussion

5.1 What we have learned from this thesis

In this project, word embeddings are generated for a total of seven different biomedical corpora taken from the PubMed database with the first set consisting of all articles related to Vascular Diseases, Peripheral Vascular Diseases or Raynaud's Disease published between 1960 and 1986 and the second set is all articles related to the aforementioned diseases but also those of the found A-Terms as found in Weeber's paper (Weeber et al., 2001) for the closed discovery we also utilised a corpora based upon all articles from a selection of the major Hematological journals. These datasets were preprocessed and normalised through the removal of too short words and case-folding. Once these techniques were complete, the preprocessed text was taken and used to generate both bigrams and trigrams before this transformed text was fed into a Word2Vec model to create a set of word embeddings for each corpus. Once this model has been created the most and least similar phrases to the found Raynauds terms are taken and mapped to the UMLS metathesaurus. Once this has been done a semantic filter is then ran which removes all terms that are not found to be in any potentially interesting semantic types leaving a shorter list of terms for manual curation to find any potential terms of interest. This pipeline was utilised in both an open and closed discovery context and as explored in Chapter 4 both of these methods have their own strengths and weaknesses with the closed discovery corpora

providing better results in regards to found A-Terms however this is likely due to the inclusion of the relevant articles more than it is the ability of the Word2Vec model. Whereas the open discovery corpora performed better when searching for B-Terms than the closed discovery which can be attributed to the specificity of the corpora to the diseases in question. The fact that the smaller corpora manage to retrieve the B-Terms more successfully than the larger corpora can also be linked to other works in the field which have found that more specific corpora can outperform more general, potentially unrelated corpora (Dusserre and Padró, 2017).

5.2 Comparison with other pipelines

As has been discussed throughout this thesis, a large amount of the pre-processing section of this pipeline was adapted from the one defined from Marc Weeber's 2001 experiments. This is visible in the fact that both processes utilise phrases, made up of bigram and trigrams, instead of a system focusing solely unigrams. The pipeline defined in this thesis does however differ to the system developed by Gordon and Lindsey in 1996 due to the fact that their system utilises a system where stopwords are detected and removed (Gordon and Lindsay, 1996). Whilst Weeber have also implemented such a system, this module was experimented with in early experiments however it was not implemented due to the size of the corpora used and the fact that removing these terms from play was unnecessary due to the fact that Word2Vec downsamples all commonly occurring terms. All three pipelines have a common factor of the fact that they all utilise the MetaMap tool to map the free-text to the UMLS metathesaurus. After the text were mapped to the UMLS the pipeline then uses the same broad semantic filter defined by Weeber in 2001 (Weeber et al., 2001). This method however was not developed or investigated at the time Lindsey and Gordon published their works and in future works, other knowledge bases have been used such as the Semantic MEDLINE Database (SemMedDB) (Hristovski et al., 2006) with some works deciding not to implement UMLS concepts at all and opting to use MeSH headings instead (Cheng et al., 2014) due to the fact these terms are assigned by humans with specific training at this task.

5.3 Limitations and Future Work

One large limitation of the work displayed in this research is the amount of time it takes to find the optimal parameters of the Word2Vec model. Whilst a grid search is exhaustive and easily expanded upon for new experiments it is an extremely time consuming method of hyper-parameter analysis due to the fact it tests all different parameter combinations. Due to the amount of time this method takes this research only ran the analysis on free-text where it could have been beneficial to run the experiments on the mapped text which could have reinforced the accuracy of the model.

It has also become clear throughout the result gathering and analysis section of this project that one of the main limitations of this work was the precision of the output. Even with a semantic filter the resulting output from the pipeline was extensive and would require a large amount of manual effort to filter through if the pipeline is to be used in a non-controlled environment where the results can be scanned for specific terms. Future works in this project would likely include the narrowing of the significant words lists as the size of these are currently dependent on the number of Raynaud phrases found with it not uncommon to have thousands of potentially significant words found, which could be achieved through the tweaking of the current filters and occurrence numbers. This could also be done by upping the thresholds for a Raynaud phrase to be found but due to the lack of documents in each corpus it became clear early on that limiting the model to a small number of terms is unlikely to provide good results.

There could also be future works on the ranking of the returned terms. This method has been explored in other experiments and has since become a mainstay of many different LBD pipelines, such as Linking Term Association and Minimum Weight Association (Henry and McInnes, 2019). The inclusion of one of these metrics would also allow for more exhaustive filter as the terms that meet the semantic filters criteria could then be filtered based on this value, thus reducing down the found terms further.

To improve the accessibility of these technologies research would likely have to be done in the generalisation of the models due to the fact these experiments have found a large degree of variability in result quality even when using slightly different parameters on the same corpus. There would also likely be a large amount of work necessary on the improvements of data availability and potentially the use of multiple types of data in the same model e.g. both biomedical literature and also clinical notes as a method of solving the data sparsity problem. One other focus of research that is beyond the scope of this work is the expansion of the relationships found, for example there has been research that has formed more comprehensive relationships by forming links between concepts through the identification of drugs and diseases within the same sentence which has been the focus of some relationship extraction research (Xu and Q. Wang, 2013). There has also been large amounts of work in the field of relationship explanation whilst still utilising co-occurance alongside the implementation of neural nets (Spiro, Fernández García and Yanover, 2019). However as also mentioned in this thesis it is possible that the integration of the SemMedDB could be a potential method of expanding the existing relationships found.

Whilst this work has provided a baseline system to the extraction of significant phrases from a biomedical dataset it is hoped that this can serve as an insight into the potential pitfalls when developing a literature based discovery system with word embeddings as its basis. It should however perform as a platform for new research and developments in not only replicating old discoveries as shown by the Raynaud-Fish Oil experiments but in also hopefully generating new hypotheses. This is due to the fact that as shown by other research (Pyysalo et al., 2018; Meng et al., 2018; Tshitoyan et al., 2019) that LBD systems are not only of interest to those studying Raynaud's disease. Through the systems put in place in this thesis there is a possibility that through optimisation and automated tweaking to the parameters of the model that it could be deployed and tested on the results reported by the aforementioned research.

References

Aronson, A. R. and F. M. Lang (2010). 'An overview of MetaMap: historical perspective and recent advances'. In: *J Am Med Inform Assoc* 17.3, pp. 229–236 (cit. on p. 19).

Aronson, Alan R (2001). 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.' In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 17 (cit. on p. 20).

Aronson, Alan R and François-Michel Lang (2010). 'An overview of MetaMap: historical perspective and recent advances'. In: *Journal of the American Medical Informatics Association* 17.3, pp. 229–236 (cit. on p. 30).

Ayyadurai, Shiva (Jan. 2014). 'The control systems engineering foundation of traditional Indian medicine: The Rosetta Stone for Siddha and Ayurveda'. In: *International Journal of System of Systems Engineering* 5, p. 125. DOI: 10.1504/IJSSE.2014.064836 (cit. on p. 2).

Azvolinsky, Anna (Jan. 2017). *Repurposing Existing Drugs for New Indications*. URL: https://www.the-scientist.com/features/repurposing-existing-drugs-for-new-indications-32285 (cit. on p. 3).

Bodenreider, Olivier (2004). 'The unified medical language system (UMLS): integrating biomedical terminology'. In: *Nucleic acids research* 32.suppl_1, pp. D267–D270 (cit. on pp. 20, 21).

Cheng, Liangxi et al. (2014). 'Enhancing the accuracy of knowledge discovery: a supervised learning method'. In: *BMC bioinformatics* 15.12, S9 (cit. on p. 76).

Chiu, Billy et al. (2016). 'How to Train good Word Embeddings for Biomedical NLP'. In: DOI: 10.18653/v1/w16-2922 (cit. on pp. 14, 25, 29).

Detailed Indexing Statistics: 1965-2017 (n.d.). URL: https://www.nlm.nih.gov/bsd/index_stats_comp.html (cit. on p. 4).

Dialani, Priya (Oct. 2019). *Using AI for Accelerating Drug Discovery*. URL: https://www.analyticsinsight.net/using-ai-for-accelerating-drug-discovery/ (cit. on p. 2).

Digiacomo, Ralph A., Joel M. Kremer and Dhiraj M. Shah (1989). 'Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study'. In: *The American Journal of Medicine*. ISSN: 00029343. DOI: 10.1016/0002-9343(89)90261-1 (cit. on p. 4).

DiMasi, J. A., R. W. Hansen and H. G. Grabowski (Mar. 2003). 'The price of innovation: new estimates of drug development costs'. In: *J Health Econ* 22.2, pp. 151–185 (cit. on p. 1).

Dusserre, Emmanuelle and Muntsa Padró (2017). 'Bigger does not mean better! We prefer specificity'. In: *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*. URL: https://www.aclweb.org/anthology/W17-6908 (cit. on pp. 27, 76).

Gordon, Michael D and Robert K Lindsay (1996). 'Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil'. In: *Journal of the American Society for Information Science* 47.2, pp. 116–128 (cit. on p. 76).

Gu, Y. et al. (2018). 'Optimizing Corpus Creation for Training Word Embedding in Low Resource Domains: A Case Study in Autism Spectrum Disorder (ASD)'. In: *AMIA Annu Symp Proc* 2018, pp. 508–517 (cit. on pp. 29, 43, 46).

Gupta, Vishal, Gurpreet S Lehal et al. (2009). 'A survey of text mining techniques and applications'. In: *Journal of emerging technologies in web intelligence* 1.1, pp. 60–76 (cit. on p. 26).

Health, National Institutes of (2019). URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html (cit. on p. 20).

Henry, Sam, Clint Cuffy and Bridget T McInnes (2018). 'Vector representations of multi-word terms for semantic relatedness'. In: *Journal of Biomedical Informatics* 77, pp. 111–119. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2017.12.006. URL: http://www.sciencedirect.com/science/article/pii/S1532046417302769 (cit. on p. 14).

Henry, Sam and Bridget T McInnes (2019). 'Indirect association and ranking hypotheses for literature based discovery'. In: *BMC bioinformatics* 20.1, p. 425 (cit. on p. 77).

Hristovski, Dimitar et al. (2006). 'Exploiting semantic relations for literature-based discovery'. In: *AMIA annual symposium proceedings*. Vol. 2006. American Medical Informatics Association, p. 349 (cit. on p. 76).

Humphreys, B. L. et al. (1998). 'The Unified Medical Language System: an informatics research collaboration'. In: *J Am Med Inform Assoc* 5.1, pp. 1–11 (cit. on p. 20).

Kaji, Nobuhiro and Hayato Kobayashi (2017). 'Incremental skip-gram model with negative sampling'. In: *arXiv preprint arXiv:1704.03956* (cit. on p. 19).

Li, Quanzhi et al. (2017). 'Data sets: Word embeddings learned from tweets and general data'. In: *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*. ISBN: 9781577357889. arXiv: 1708.03994 (cit. on p. 29).

Liu, Haibin et al. (2012). 'BioLemmatizer: a lemmatization tool for morphological processing of biomedical text'. In: *Journal of biomedical semantics* 3.1, p. 3 (cit. on p. 25).

Major, Vincent, Alisa Surkis and Yindalon Aphinyanaphongs (2018). 'Utility of general and specific word embeddings for classifying translational stages of research.' In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association, p. 1405 (cit. on p. 26).

McCray, Alexa T, Alan R Aronson et al. (1993). 'UMLS knowledge for biomedical language processing.' In: *Bulletin of the Medical Library Association* 81.2, p. 184 (cit. on p. 30).

McCray, Alexa T, Suresh Srinivasan and Allen C Browne (1994). 'Lexical methods for managing variation in biomedical terminologies.' In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, p. 235 (cit. on p. 19).

Meng, Guilin et al. (2018). 'Adopting literature-based discovery on rehabilitation therapy repositioning for stroke'. In: *BioRxiv*, p. 422154 (cit. on pp. 12, 78).

Mikolov, Tomas, Kai Chen et al. (2013). 'Efficient estimation of word representations in vector space'. In: *arXiv preprint arXiv:1301.3781* (cit. on pp. 18, 19).

Mikolov, Tomas, Ilya Sutskever et al. (2013). 'Distributed representations of words and phrases and their compositionality'. In: *Advances in neural information processing systems*, pp. 3111–3119 (cit. on pp. 18, 25).

Prasad, V. and S. Mailankody (Nov. 2017). 'Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval'. In: *JAMA Intern Med* 177.11, pp. 1569–1575 (cit. on p. 1).

Pratt, Wanda and Meliha Yetisgen-Yildiz (2003). 'LitLinker: Capturing connections across the biomedical literature'. In: *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*. ISBN: 1581135831. DOI: 10.1145/945645.945662 (cit. on p. 11).

Pyysalo, Sampo et al. (Oct. 2018). 'LION LBD: a literature-based discovery system for cancer biology'. In: *Bioinformatics* 35.9, pp. 1553–1561. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty845. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/35/9/1553/28557393/bty845.pdf. URL: https://doi.org/10.1093/bioinformatics/bty845 (cit. on pp. 12, 78).

Schnabel, Tobias et al. (Sept. 2015). 'Evaluation methods for unsupervised word embeddings'. In: *Proceedings of the 2015 Conference on Empirical Methods in*

Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, pp. 298–307. DOI: 10.18653/v1/D15-1036. URL: https://www.aclweb.org/anthology/D15-1036 (cit. on p. 14).

Smalheiser, Neil R (2019). 'Ketamine: a neglected therapy for Alzheimer Disease'. In: *Frontiers in aging neuroscience* 11, p. 186 (cit. on p. 12).

Smalheiser, Neil R and Don R Swanson (1998). 'Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses'. In: *Computer methods and programs in biomedicine* 57.3, pp. 149–153 (cit. on pp. 10, 11).

Spiro, Adam, Jonatan Fernández García and Chen Yanover (July 2019). 'Inferring new relations between medical entities using literature curated term co-occurrences'. In: *JAMIA Open* 2.3, pp. 378–385. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooz022. eprint: https://academic.oup.com/jamiaopen/article-pdf/2/3/378/32298793/ooz022.pdf. URL: https://doi.org/10.1093/jamiaopen/ooz022 (cit. on p. 78).

Srinivasan, Padmini (2004). 'Text Mining: Generating Hypotheses from MEDLINE'. In: *Journal of the American Society for Information Science and Technology*. ISSN: 15322882. DOI: 10.1002/asi.10389 (cit. on p. 4).

Swanson, Don R. (1991). 'Complementary structures in disjoint science literatures'. In: *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1991*. ISBN: 0897914481. DOI: 10.1145/122860.122889 (cit. on p. 9).

Swanson, Don R and Neil R Smalheiser (1997). 'An interactive system for finding complementary literatures: a stimulus to scientific discovery'. In: *Artificial intelligence* 91.2, pp. 183–203 (cit. on p. 10).

Swanson, Don. R. (1986a). 'Fish oil, Raynaud's syndrome, and undiscovered public knowledge.' In: *Perspectives in Biology and Medicine*. ISSN: 00315982 (cit. on pp. 3, 4).

– (1986b). 'Undiscovered Public Knowledge'. In: *The Library Quarterly: Information, Community, Policy* 56.2, pp. 103–118. ISSN: 00242519, 1549652X. URL: http://www.jstor.org/stable/4307965 (cit. on p. 5).

TH, Muneeb, Sunil Sahu and Ashish Anand (July 2015). 'Evaluating distributed word representations for capturing semantics of biomedical concepts'. In: *Proceedings of {B}io{NLP} 15*. Beijing, China: Association for Computational Linguistics, pp. 158–163. DOI: 10.18653/v1/W15-3820. URL: https://www.aclweb.org/anthology/W15-3820 (cit. on p. 14).

Tshitoyan, Vahe et al. (July 2019). 'Unsupervised word embeddings capture latent knowledge from materials science literature'. In: *Nature* 571.7763, pp. 95–98. DOI: 10.1038/s41586-019-1335-8 (cit. on pp. 12, 27, 29, 78).

Voytek, Jessica B and Bradley Voytek (2012). 'Automated cognome construction and semi-automated hypothesis generation'. In: *Journal of neuroscience methods* 208.1, pp. 92–100 (cit. on p. 12).

Wang, Yanshan et al. (2018). 'A comparison of word embeddings for the biomedical natural language processing'. In: *Journal of Biomedical Informatics* 87, pp. 12–20. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.09.008. URL: http://www.sciencedirect.com/science/article/pii/S1532046418301825 (cit. on p. 13).

Weeber, Marc et al. (2001). 'Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries'. In: *Journal of the american society for information science and technology* 52.7, pp. 548–557 (cit. on pp. 4, 5, 21, 25, 30, 75, 76).

Xu, Rong and QuanQiu Wang (2013). 'Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing'. In: *BMC bioinformatics* 14.1, p. 181 (cit. on p. 78).

Ye, Zhan et al. (2016). 'SparkText: Biomedical text mining on big data framework'. In: *PLoS ONE* 11.9, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0162721. URL: http://dx.doi.org/10.1371/journal.pone.0162721 (cit. on p. 27).

Chapter 6

Appendix

Blood Viscosity	Platelet Aggregation	Vascular Reactivity
Blood Viscosity	Fibrinolysis	Vasodilatation
Erythrocyte Deformability	Platelet Aggregation	Vasodilation
Plasma Viscosity Level	Thrombosis	Vasospasm
Hemorheology	Platelet Adhesiveness	Vasospasm Mechanisms
Decreased Vascular Flow	Effects, blood coagulation	Vasomotion
Hyperviscosity		Decreased Vascular Resistance
		Decreased Vascular Flow

Table 6.1: Found B-Concepts as defined in Weeber's 2001 paper

Blood Viscosity	Platelet Aggregation	Vascular Reactivity
Fish Oils	Eicosapentaenoic acid	Fatty Acids, Essential
Maxepa	Cod Liver Oil	Dietary Fats
Fatty Acids, Omega-3	Fish Oils	
Omega-3 polyunsaturated fatty acid	Maxepa	
Eicosapentaenoic acid	Fatty acids, omega-3	
Epa-e	Omega-3 polyunsaturated fatty acid	
Salmon Oil		

Table 6.2: Found A-Concepts as defined in Weeber's 2001 paper

Architecture	Window	Score
Raynaud	1	0.0035824481182552227
Raynaud	10	0.003384353400797274
Raynaud	30	0.003522011043204793
PVD	1	0.8792117434042528
PVD	10	0.7926014610579173
PVD	30	0.8393686332752044
Vascular Disease	1	0.7648024610640313
Vascular Disease	10	0.7840307721825088
Vascular Disease	30	0.7582937348648169

Table 6.3: Context Window Parameters for Open Discovery Corpora