Exploring the application of Natural Language Processing to scientific medical cannabis publications

Dissertation presented for the degree of Master of Science in the Department of Data Science at the Univeristy of Cape Town

Author - James de Beer ; Supervisor - Juwa Nyirenda ; Co-Supervisor - Joost Heeroma

March 14, 2021

Abstract

Cannabis has become recognised internationally as a powerful medicinal plant. The explosion of clinical research on cannabis has made it difficult for researchers and medical professionals to keep up to date with new findings. Analyzing the large quantities of available text data using natural language processing and machine learning algorithms could improve the speed and accuracy at which cannabis research is processed, as well as expose hitherto unknown correlations between cannabis compounds and the treatment of health conditions. In turn, this would help direct future research and clinical. trials. This thesis aims to develop an appropriate method to extract the key correlations between cannabis compounds, human physiology and disease from the existing medical literature. First, natural language processing techniques (such as document clustering and topic modelling, global vector word embeddings and supervised document classifies) are used to group 500 journal articles from the general literature on cannabis according to broad research topics; analyse the interaction between cannabis compounds, human physiology and diseases; and train a classifier to classify unseen documents. Second, the correlations generated through this quantitative process are assessed qualitatively against those in a manual dataset of research findings from more than 500 studies collated over a number of years and provided by a medical company specialising in cannabis research. The results indicate that the methods and working code developed were able to effectively and accurately demonstrate correlation and reproducibility of results. This was shown by the close similarity of ranked target variable to disease correlations. The unsupervised methods were able to effectively cluster and model topic distributions between the data to group documents by topic, while the supervised learning methods were able to accurately train models based on these suggestions, thereby solving a real-world practical problem in data management and analysis. The approach has potential for directing researchers towards priority areas for new research and clinical trials, optimising their time and reducing redundant research. The validity of the results need to be tested through clinical trials.

1

Contents

List of Tables

List of Figures

Plagiarism Declaration

I know the meaning of plagiarism and declare that all work in the dissertation which is properly acknowledged, is my own.

_____

1 CHAPTER 1 - INTRODUCTION

More than 150 million people regularly use cannabis and its derivatives, making it one of the world's most popular and widely used substances (Lawler, 2019). Cannabis has had a complicated and interesting past as a recreational and medicinal plant in use for over 5,000 years (Bennett, 2010). In the United States (US), cannabis was used as a patent medicine during the 19th and early 20th centuries, and it was described in the United States Pharmacopoeia for the first time in 1850. Federal banning of cannabis use and sale first happened in 1937 and cannabis was dropped from the United States Pharmacopoeia in 1942. Prohibition under federal law occurred with the Controlled Substances Act of 1970 (Bridgeman, 2017), which, to many, started an official 'war on drugs' by the American Government, having a global impact. The prohibition also meant the end of research into the medicinal use of cannabis as funding dried up. This had a direct adverse effect on the running of clinical trials, and generation of data to support medicinal use (Bridgeman, 2017).

Recently, some US states legalized the use of both recreational and medicinal cannabis. Soon after, many more countries around the world followed suit. The perception of cannabis has also shifted from that of a gateway drug towards a greater understanding of its diverse medical benefits. This understanding partly contributed to the unbanning of cannabis in several countries (Bridgeman, 2017). Also many practitioners globally are adopting methods involving use of *Cannabis sativa* in the treatment of ailments and diseases (Hall, 2018). However, these medical effects are still widely misunderstood by the general population (Heeroma, 2016). Furthermore, there are relatively few registered medicines backed by clinical trials which support such treatments and their cannabinoid-derived active medicinal properties (Hall, 2018).

These new legislation's have led to a flurry of research activities into the healing properties of cannabis and has resulted in much of the focus of research in the field being aimed at proving the efficacy and safety of cannabis generally, and specifically in the treatment of ailments. This is because, predominantly in the past, but still to this day, there are controversies over legal, ethical and societal implications associated with cannabis use, safe administration, packaging and dispensing. The occurrence of adverse health consequences attributed to prolonged use in certain cases, and therapeutic indications based on limited clinical data, represent some of the complexities associated with this treatment (Bridgeman, 2017). As discussed later, the complexity of the plant and its constituents means that generating rapid results from clinical trials and standardized pharmaceutical dosages in the correct setting is not a quick process.

The research related to cannabis has led to an explosion in the amount text data produced in the form of academic publications, making it difficult for scientists and the public alike to stay up to date and spot opportunities presented by the new scientific breakthroughs. The trend in the explosion of research publications in the field is forecast to continue as there are still many gaps that require conclusive research (Gong, 2018). These gaps are primarily pharmacological knowledge of the effect of the hundreds of naturally derived plant constituents which occur in the plant's flower, both in their isolated compound form as well as in their complex natural profiles. These natural profiles can comprise of over 400 different

compounds, and the unique effect these compounds have when working together in the body is known as the entourage effect, and is one of the most under-researched areas of medical cannabis (Russo, 2019).

It is going to become harder to digest the rapidly increasing density of literature and studies on cannabis and its medical properties, as the global market continues to grow, unless a way is developed for summarizing the data. The inclusion of current advances in statistical techniques and methods in the analysis of this emerging data will be a critical contribution to understanding the complex plant which is *Cannabis sativa.* It is essential to be prepared to handle this type of data adequately and maximize the chances of identifying the medical potential of cannabis and appropriate treatment opportunities.

One organization that has pioneered the work of summarizing journal papers relating to research in the use of cannabis is Green House Medical. Green House (GH) Medical is a company in the Netherlands comprising a team of scientists and experts that consult, conduct scientific and medical research, engage in product development, and provide quality control services. Their work is predominantly focused on cannabinoids, their biochemical interactions in the human body and how they could be used to help combat disease.

GH Medical has one of the world's largest genetic library of cannabis strains including several unique locally indigenous strains (known as landraces) unknown to the general public and scientific world, which are reported to have therapeutic properties but have not been widely investigated (Heeroma, 2016). These include strains from African countries such as South Africa (Pondoland), Lesotho, Democratic Republic of Congo and Swaziland. GH Medical's mission statement is to merge anecdotal evidence with rigorous scientifically based evidence. A step towards this vision is the method GH Medical is currently pursuing, which is to isolate each individual cannabinoid, test their biochemical and physiological function and most importantly their therapeutic potential, either alone or in combination with other cannabinoids.

Over the past 20 years, GH Medical has read, aggregated key findings and drawn conclusions from the growing pool of medical and academic research being conducted within the medical cannabis sphere, driven predominantly by widespread adoption and acceptance of this natural plant-based form of medicine. Until now, the primary method of investigation has been through years of literature-based research. Their current manual method of reading, summarizing and drawing conclusions has understandably taken a long time to do given the large volumes of text and publicly available data produced over the years, and the numerous findings of existing analyses. This method is inevitably becoming less efficient as the amount of literature produced grows rapidly. GH Medical has provided their library of research papers which is the input text data used to build natural language processing model. GH medical has also provided its collection of manually aggregated correlations between Diseases the way cannabis constituents act with humans physiology. These manually collected correlations are used to qualitatively validate the models built off the published.

This thesis hypothesizes that analyzing large quantities of available text data using natural language processing and machine learning algorithms will improve the accuracy and speed at which cannabis research can be digested, categorized and have correlations drawn between

compounds, their effects and diseases. The thesis also hypothesizes that this type of text analysis will be able to uncover useful but, as yet, unknown relationships between the naturally occurring compounds within cannabis and their interaction with human physiology.

The **primary objective** of this paper is to develop an appropriate strategy using natural language processing techniques to accurately group the medical literature according to broad research topics, to analyse the interaction and correlation between cannabis compounds, human physiology and diseases, and to train a classifier to classify unseen documents.

In this objective there are three main issues to be addressed that are categorized as Objective 1's sub-objectives:

- (1) we seek a technique that we can use to accurately group the medical literature according to broad research topics;

- (2) we seek a way to analyse the interactions and correlations between cannabis compounds, human physiology and diseases; and

- (3) we seek to classify unseen documents into the broad research topics discovered.

The **secondary objective** of this paper, which derives from accurately carrying out the primary objective, is to;

- compare the correlation results from objective 1 sub-objective 2, to the manually derived correlation results drawn up by GH Medical.

This aim is focused on making the process of investigating and deciphering new research by both medical professionals and the general public more targeted, accurate and accessible.

The rest of the thesis is organized as follows. Chapter 2 of this paper describes the literature review which investigated the different processing techniques currently available for natural language processing aimed at solving similar tasks in the field of text analysis in medicine. It also includes a section on the pharmacological properties of cannabis as this was necessary to understand in order to identify appropriate search terms and key variables. Drawing on the findings of Chapter 2, Chapter 3 describes the data-sets and the methods used to understand the characteristics of the data, as well as to get it into a workable format suitable for analysis. Chapter 4 describes the methodology used and followed in the working code to build the models. The combination of methods was unique to this paper and did not follow a classical approach described in the related work, but rather built on existing methods to demonstrate how they can be used to achieve these objectives. Chapter 5 presents the results generated from the models, while Chapter 6 discusses the results and their significance, their limitations and the scope for further work. Chapter 7 concludes the paper and ties the success of the developed methods to the aims, hypothesis and results.

2 CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

This literature review investigates existing knowledge in the field of natural language processing, specifically with respect to medical and scientific texts. The purpose of the literature review is to broaden understanding in the field specific to answering the aims and objectives and hence direct the methodologies in a practical and efficient manner. This is important to build on existing research and pave the way for new methods and research conducted. The scope of the literature review includes: the feature selection and processing required to deal with large quantities of text data in workable formats; the reviewed methodologies suitable to processing text data to achieve the objectives and interpret the results; and to identify the limitations of these methodologies, given the research questions at hand.

The task presented in this paper addresses a few practical problems which each have unique use cases. The literature review is aimed at specifically investigating the aggregation of these techniques to yield a productive output for Green House Medical and the application they require. The literature required for review has been broken into four main classes which ultimately attempt to answer the first research objective, which is to develop an appropriate strategy using natural language processing techniques to accurately group the medical literature according to broad research topics, and to analyse and classify the interaction between cannabis compounds, human physiology and diseases. The strategy required to answer this has been broken into four main research topics required for review.

- Accurately categorizing literature into discrete topics/classes to make the process of targeting specific types of research easier for researchers

- Accurately reproduce the manual summarized findings on the correlation between cannabis compounds, human physiology and diseases

- Build a classification model which can categorize future literature into the discrete topics identified in the first point.

- Investigate how to form recommendations for the further understanding of unknown relationships between these compounds and the effect on human physiology and diseases.

The literature review is ordered methodologically and looks at algorithms and techniques used in analysing scientific text published in papers dealing with similar problems to this thesis, and is ordered in a way which tackles the specific objectives.

These methods have a wide range of application in statistics and machine learning in general, but for the for the purpose of this study, these methods have been investigated explicitly in relation to the application of Natural Language Processing and text analysis in the field of medicine.

The review starts by defining natural language processing as found in literature. It then goes on to review the most ideal methods for data preprocessing and exploratory data analysis which aims at making downstream processing more efficient. It then looks at the methods of supervised and unsupervised learning techniques and their previous applications, to help

develop a strategy to answer the research objectives, and prevent time spent on 'reinventing the wheel'. Lastly, the review investigates the pharmacological properties and relevant compounds, receptors and enzymes associated with medical cannabis: this is important for identifying the search terms to be used in the algorithms and to gain a deeper understanding of the problem at hand.

2.2 Related work in the field of:

Natural Language Processing and Text Mining

Text Mining (TM) is defined as the task of extracting what we to consider useful information and knowledge from data that is in the form of text (G, 2009). Text data mining needs to be able to deal with highly unstructured input data, which can be cleaned and standardized. The process of making an unstructured input more ordered, allowing it to be transformed into representable models, and generating information discovery is the primary goal of text mining (Kim JD, 2003) . Text mining in general is considered a difficult task which gathers many dissimilar sub-tasks to support different applications (C. Luque, 2018). These sub-tasks can include text-based prediction, pattern identification, hypothesis development, relationship extraction, document classification and summarization (Jingbo, 2014).

Natural Language Processing (NLP) falls into the world of AI and 'Computational Linguistics' which is looking at the analysis of natural language. A critical challenge of Natural Language Processing is to surpass the rate of information digestion and implicit complexity language interpretation presents. Because of this dynamic complexity, numerous incongruent techniques need to be considered (C. Luque, 2018). The literature review of techniques below looks at information retrieval techniques and their ability to access important features, improving the effectiveness of a user's need to easily request and receive specific information. Information Extraction aims to extract only the relevant entities, relationships and events from a set of documents (Leaman, 2009). The development of labelled data yields the potential for supervised Machine Learning (ML) algorithms which learn automatically from data input and allow for future unlabeled prediction output.

Unsupervised Learning (UL) is a type of algorithm which groups unlabeled data according to similar components, features or patterns that they may share. Supervised Learning on the other hand is characterized by the knowledge obtained due to labelled training data and the prediction on future unseen data.

As discussed in (Seeger, 2001) it is possible to deduce training labels from exploratory hypothesis testing using unsupervised learning, and feed this new found information to a supervised methodology (Kurama, 2020). The majority of new research on the ECS and cannabinoids is in text format. Either Journals, medical documents, surveys, academic papers or academic research and more recently patient feedback. Much like most 'big data', it is not always digestible through manual interpretation. Text mining systems follow three phases generally to achieve the aim of automatically extracting and discovering new knowledge from the unstructured input (C. Luque, 2018). Figure 1 below illustrates this process.
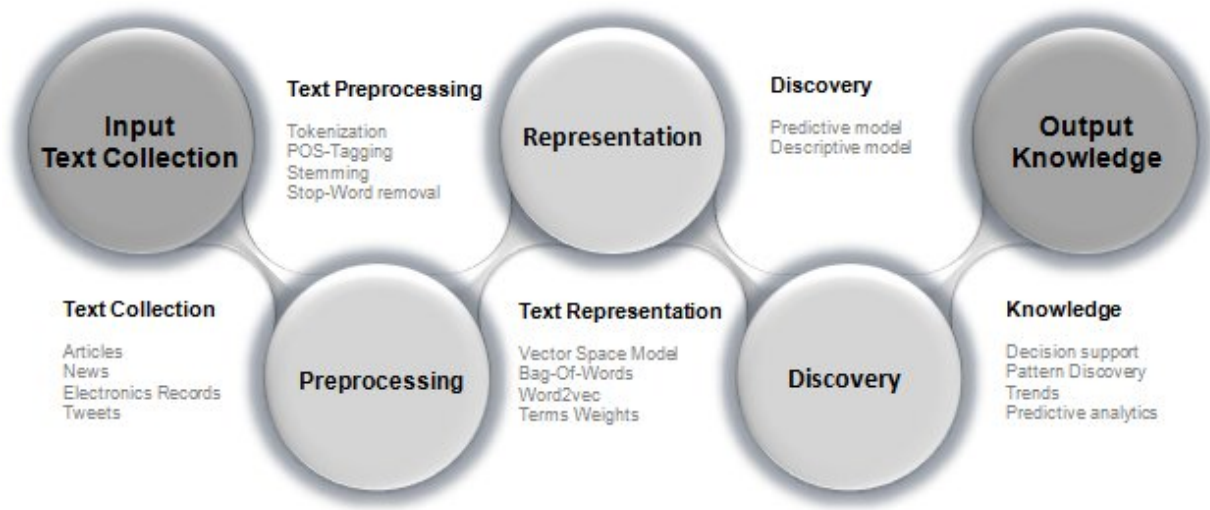
Figure 1: Text Mining Process (C.Luque,2018)

Text Data, Pre-Processing and Exploratory Data Analysis

In raw form, a text document is a string. It can also be considered multidimensional – based on the representation of the data. As a string, it is a dependency-orientated data type (Aggarwal, 2015). Each string corresponds to the document as a sequence of terms or characters; however, text is rarely represented as strings because it is difficult to directly use and leverage the ordering of terms when trying to scale these operations. According to Aggarwal (2015) it is useful practice to use a vector-space representation, where the frequencies of terms occurring in documents is used. This does result in a loss of semantics, as the the ordering of the terms is lost. When working with frequencies, it is normal to normalize the data with statistics, such as the length of the documents or the ratio of occurrence between documents. In a vector space the text data is converted to multidimensional quantitative data (Mysiak, 2019) with attributes corresponding to the terms themselves, and the frequencies or inverse frequencies correspond to the values of these attributes. A phenomenon called data sparsity is introduced in this method of text representation, where most values that comprise each vector are zero due to the nature of how terms-in-documents relate to terms-in-corpus. This sparsity brings new challenges to quantitative analysis, and introduces stringent modification requirements to the data type (Cothenet, 2019). This is the primary reason text mining is considered a separate avenue to data mining in general (Charu & Zhai, 2018). When working with text as a Dependency-orientated data type, one can think of all terms as vertices that are connected together by a set of edges, which hold the relationship. These data types are more challenging to work because of the complex arrangement contained in language. However, these dependencies if captured correctly can yield their own unique benefits to analytical results drawn (Romanov, Lomotin, & Kozlova, 2019).

Text data is inherently highly disordered, and means of ordering and reducing the dimensionality of this data into workable formats is imperative to initiating natural language processing

11

techniques (Romanov, Lomotin, & Kozlova, 2019). (Uysal, 2013) notes that the Natural Language pre-processing channel is comprised of the following components:

- White space, punctuation, symbol and stop word removal
- Tokenization
- Feature word removal, including stemming
- Encoding harmonization
- Synonym aggregation

Charu & Zhai emphasizes that data preprocessing is the most crucial phase of data mining, and mentions that too much focus is often put on the analytical component instead. It is recommended that this process makes use of descriptive statistics and visualisations when engaging in textual EDA (Mysiak, 2019).

This component is said to be the backbone to the analysis, which gets the highly unordered format of text into a workable format and the importance is emphasized as a necessary preparation before starting to extract features from the data-set which will be used in the machine learning techniques (Slavazza, 2019)

According to (Romanov, Lomotin, & Kozlova, 2019) the most appropriate combination of preprocessing is tokenization, stemming and lower-casing, and that lower-casing increases classification F1-score accuracy significantly. In fact, Uysal's (2013) paper mentions that preprocessing is as important as feature extraction in text classification. He notes that there is no unique combination of preprocessing tasks to provide successful results and that experimentation is required to select the correct preprocessing tasks for the problem rather than entirely/individually applying or ignoring them (Uysal, 2013).

General concepts are able to be explored through good visualisations much quicker than quantitatively, and it is recommended that the extraction of features which can be visually demonstrated should be included in EDA (M Moradi, 2020).

Feature extraction and its importance in developing accurate models

Due to text data being highly disordered and of high dimensionality, feature selection is a fundamental task prior to any unsupervised or supervised text classification problem. Text data has a high dimensionality of features and 'noisy variables. This noise is often non-relevant to the data being investigated and needs to be removed in order to extract the most important features (Charu & Zhai, 2018).

Feature extraction helps to reduce computational load when working with the data-set and drastically improves the accuracy of analysis by further selecting or combining variables, which reduces the dimensionality while accurately describing the data-set (Kothari, 2017). Machine learning (ML) and statistical algorithms are workable with numerical vectors and not with text. Useful characteristics to preserve from the corpus are most commonly either statistical or semantic characteristics (Claus Asmussen, 2019). This is necessary when working with text as some words are more likely to be correlated to a class distribution than other words. Literature has proposed a wide variety of methods for determining important features and the most relevant have been selected.

More traditional methods of feature extraction include Term Frequency (Tf), Term Frequency – Inverse Document Frequency ($Tf-idf$) , and Bag-of-Words which are all statistical approaches (M Moradi, 2020). A more modern approach to extracting features is through continuous word representation language models. Three most relevant and backed by widely cited literature in recent times is that of Word2Vec and Stanford's Global Vectors for Word Representation (GloVe) (M Moradi, 2020). These more modern approaches tackle a problem which the more traditional methods do not capture as well, which is the preservation of semantic meaning in the text as a direct representation of the word in the corpus.

Term Frequency simply describes how often a word has occurred in the data-set or subset and can be thought of as a probability of finding said word, whereas Tf-Idf is a measure of how frequently words occur in a subset of documents in the entire corpus and provides significant improvement to the feature space (Kothari, 2017).

According to (Slavazza, 2019), ($Tf - idf$) is commonly regarded as the best method for machine learning input features for text processing. However, in (Romanov, Lomotin, & Kozlova, 2019) they describe ($Tf - idf$) features to be extremely computationally expensive and thus reverted to solely using semantic vector models such as GloVe as features.

In Selivanov (2018), his vignette specifies that because of 'copy-on-modify' semantics used in programming languages, it is tricky to iteratively grow document term matrices. This can lead to a bottleneck when scaling the size of data-sets. He goes on to say that vectorized texts are unique in that they save computational memory as they are not stored in massive matrices. They also have the ability to hold semantic meaning. Vectorized features allow for a further use case, where entire documents can be vectorized based on the sum of their parts i.e. the vectors (words) that make up a document (Jeffrey Pennington, 2014). This allows for entire documents to be stored in a single vector. The accuracy of this application is suggested to outperform conventional methods (Jeffrey Pennington, 2014)

Vector Space embedding models – a multipurpose solution to feature extraction and correlating data

Deep Learning has been a fundamental pathway to development of Natural Language Processing efficiency, and has added to the dynamic area if research (Cothenet, 2019). As mentioned there are 2 widely used types of modern NLP embeddings, namely Word2Vec and GloVe. Word Embeddings help to make the English language, or any for that matter, 'readable' by machines by capturing relationships between the words. These relationships could be contextual, semantic or even morphological (Cothenet, 2019). One-hot encoding is a simple form of embedding, in which all words in a corpus get converted to integers, and the integers used as the indices to one-hot encode each word. The one-hot encoded vectors are all orthogonal to one another, and do not represent the semantics in a vector space (Khoshgoftaar, 2020).

Unlike Latent Semantic Indexing (LSI) which is a matrix decomposition model, which needs the entire decomposition to be recalculated when any part of the dictionary changes, word2vec and GloVe are able to build onto the existing vector space and increase the accuracy of words vectorized positioning based on the new added data (C. Luque, 2018). Representing words from text as vectors numerically based on their contextual meaning has become the

'de facto' way of analysing text in machine learning (Khattak, 2019). Khattak (2019) aims to create a blueprint for clinicians and healthcare works who want to incorporate text features in their own models and application, very similarly to the objectives of this paper. Khattak (2019) explores types of word representations, different clinical text corpora and make use of pre-trained clinical word vector embeddings, as well as the applications to these approaches. It is noted in this paper that successfully embedding a feature of word embeddings to be critical to the success of the interpretability of these results.

Word embeddings have an increased application in the range of biomedical Natural Language Processing (bioNLP) tasks which can range from drug discovery to diagnosis of automated diseases (Jha, 2018). Jha (2018) goes on to describe that even though word embeddings have syntactic and semantic meaning regularities, this meaning can remain elusive. This can be a problem in sensitive domains such as bio-medicine where the interpretation of the results is crucial to widespread understanding of the methods. Jha (2018) addresses this issue by creating transformation matrices that transforms text into input embeddings, where they are both interpretable and retain their expressive features.

Ghosh (2018) investigates the use of word embeddings to automatically create discrete taxonomies, and uses a disease vocabulary driven word2vec model (Dis2Vec) to evaluate the model against a corresponding human annotated data-set of taxonomies. The findings from this study show that their Dis2vec model outperforms distributed vector representations in the ability to capture attributes across different classes of diseases.

Wu (2018) notes that even though there has been celebration over Word2Ve as a technique to yield semantically rich representations, there has been relatively less success extending this to generate unsupervised sentence or document embeddings. Recent work has shown that a distance measure between documents called Word Movers' Distance (WMD) which aligns semantically similar words is able to yield unprecedented classification accuracy. However, WMD is very computationally expensive and is hard to extend passed a KNN classifier. Wu (2018) goes on to emphasize that this technique can be replaced by other techniques such as GloVe global vector embedding model.

GloVe which stands for 'The Global Vectors for word representation is introduced by by Jefferey Pennington et al (2014) and is said to be a very efficient and effective way to learn vector representation of words. Shi (2014) compares the word2vec and GloVe models and shows that the skip-gram with negative-sampling (SGNS) technique which is implemented in word2vec is similar to the objective of GloVe even though their cost functions are defined differently. While Glove explicitly factorises a co-occurrence matrix and SGNS implicitly factorises a point-wise mutual information matrix, they share similar objectives. The differences come from their cost functions and weighting structures (Shi, 2014). Shi (2014) describes the GloVe model as more general and more suited for a wider domain optimization.

Jeffrey Pennington (2015)'s paper on GloVe explains their creation of a new global logbillinear regression model. This model combines te advantages of the two major model families which is the global matrix factorisation and a local context window method. The GloVe method leverages statistical information by training the nonzero elements in the co-occurance matrices as apposed to on the entire sparse matrix or on the the context windows alone (Jeffrey

14

Pennington, 2015). Even though the comparison quantitatively of GloVe and word2vec is complicated, through means of controlling the configuration of certain variables, Pennington was able to test comparably. It was also found that word2vec's performance decreased if the number of negative samples increased over 10. For the same corpus, vocabulary, window size and training time GloVe managed to achieve more accurate results in a shorter time frame.

The result they concluded was that GloVe outperformed other models on word analogy, word similarity and named entity recognition tasks (Jeffrey Pennington, 2015)

Baroni et al. (2014) substantiates this claim in noting that the methods are not fundamentally different as they both probe the underlying co-occurrence statistics in a corpus, but the count-based methods which capture global statistics in GloVe can be advantageous.

Unsupervised Learning Methods to categorize and label documents

Topic modelling is one of the preferred and more successful methods for categorizing sparse, highly disordered text data (Selivanov, 2018). This is especially preferred when a user's review is required to be highly specific, or targeting niche topics within the field research being modeled. Claus Asmussen (2019) has pointed out that defining an input to topic modelling algorithms can be a limitation, and his work highlights the use clustering as form of unsupervised learning which help point out a 'trusted estimation' of the number, or size topics as a preliminary step to topic modelling.

A cluster is a subset of data which are similar. Clustering, which is a form of unsupervised learning, is the process of dividing a subset into groups such that:

- The members of each group are as similar as possible to one another.
- Different groups are as dissimilar as possible from one another.

Clustering can uncover previously undetected relationships in a data-set, and there are four broad categories of clustering methods which are utilized when looking at the structure of text data in an unsupervised manner (Aggarwal, 2015), namely Distance Based Methods, Density based methods, Probabilistic methods and Neural Net based methods. A good clustering of text requires effective feature selection and a proper choice of the algorithm for the task at hand (Charu & Zhai, 2018).

Among these different algorithm classes, the distance-based methods are some of the most popular in a wide variety of application, and specifically within text clustering (Charu & Zhai, 2018).

Aggarwal (2015) mentions that many clustering algorithms require significant adjustments to address the special structure of text data.

Table 1 below lists some popular text clustering algorithms.

For a clustering method to be good, it needs to be able to determine within-cluster similarity and between cluster dissimilarity, it needs to be able to handle high dimensionality, handle various types of attributes, deal with noise and outliers and be scalable and interpretable (Charu & Zhai, 2018).

Table 1: Popular Text Clustering algorithms

Distance.Based	Distance.Based.1	Probabilistic.Based	Neural.Network	Count.Based
Partitioning	Hierarchical	Mixture of Gaussians	Self-Organizing Maps	GloVe
K- Means	Agglomerative	Probabilistic Latent Sematic Allocation	Word2Vec	
K-Medians	Divisive	Probabilistic Latent Dirichlet Allocation		
K-Medoids				

According to (Ambigavathi, 2020) who's paper investigates clustering algorithms in Machine Learning for medical healthcare data, they conclude that is hard to decide in advance which clustering algorithms would be the most suitable for a particular data-set, and the number of optimum clusters.

Applying clustering methods blindly can often lead to meaningless and unhelpful results (Ambigavathi, 2020). The clustering algorithms will naturally form clusters: this does not necessarily mean they are correct, or meaningful in context. (Q Zhang) expresses the importance of a dissimilarity matrix, which shows the similarity levels between observations, and this helps one visualize the tendency for clusters to form.

A popular method in literature using probabilistic document clustering in fact a topic modelling. Topic modelling and specifically Latent Dirichlet Allocation (LDA) aims to create a probabilistic generative model for the text documents in the corpus, which represents the corpus as a function of hidden variables, which have parameters estimated using the particular corpus (Charu & Zhai, 2018). It is in essence a dimension reduction approach, but has a rich interpretive quality too. Features are typically reduced to a few topics. In this sense LDA is comparable to discrete PCA. When using probabilistic generative models, the categorization of documents in clusters is less important to the latent clusters of topics and this way of clustering in the scientific and statistic area of research has branched into its own subject, 'topic modelling' (Selivanov, 2018).

The most popular topic modelling approach in recent literature and studies is that of Latent Dirichlet Allocation (LDA) (Charu & Zhai, 2018). In Stefano Sbalchiero's (2012) paper he concludes that LDA topic modelling as a generative probabilistic method proves to be a superior method for textual topic modelling. LDA requires a distribution prior and is therefore a Bayesian version of LSI, and is advantageous over LSA as it is resistant to over-fitting.

A look at supervised learning methods to build a classifier for labelled documents

Brendal (2020) explores how to leverage unsupervised learning on supervised learning problems and suggest that once unsupervised learning tasks have produced labels as an output, they can be used to train classifiers. From the literature, what have shown to be most useful and popular in text classification algorithms and have high accuracy with text classification include Multinomial Naive Bayes (MNB), k-nearest neighbor (KNN), multinomial logistic regression (MLR), gradient boosted machines (GBM) and support vector machines (SVM) and neural networks (NN) (Khoshgoftaar, 2020). Each classifier has pros and cons naturally (Slavazza, 2019), and each behaves differently with text data as it would with numerical data with regular distributions of the data, and the usefulness of a classifier will always be dependent on the task (Zhang, 2017).

Naïve Bayes has a very low time complexity, and its assumption is shown to work well in real world situations, including document classification (Aggarwal, 2015). The problem of increased dimensionality is dealt with in the Naïve Bayes algorithm as each feature is estimated as a one-dimensional distribution (Aggarwal, 2015).

Multinomial Logistic Regression (MLR) is a classification algorithm that generalizes a logistic regression to multi-class problems. The package GLMnet is a package is highly recommended for fitting generalized linear model, making use of a penalized maximum likelihood (Trevor Hastie, 2014). Trevor explains in his vignette that this method of predictive models works well thanks to its fast algorithm which is able to take advantage of the sparse input matrix, thus making it highly efficient for text analysis.

Unlike the more conventional approaches to text classification which are heavily dependent on empirical data, support vector machines are models which are able to perform well with high dimensional text data (Joachims, 2001). In Joachim's paper on SVM in text classification, he outlines why support vector machines are able to handle such large feature spaces, how doing that is relatable statistically to text properties, and how text data needs to be poised in order to be a successful classifier. Joachim's paper analyzes from a theoretical perspective why SVM's are a unique and powerful tool, and demonstrates how previous benchmark-based success on learning methods is not sufficient justification.

Support Vectors hold to their name, as they support the position of the maximal-margin hyper-plane. If a support vector data-point was to move, so would its support, and the maximal margin hyper-plane would move. There are properties of Support Vector Machines which describes the flexibility support vector machine algorithms hold with text data, as it is able to be effective in high dimensional spaces, including spaces with greater dimensionality than the number of samples themselves. Due to the ability to map high-dimensional feature spaces of data through kernels, SVM's are also memory efficient (Joachims, 2001).

There is significant evidence in (Zaghloul, 2009) which support the idea that SVM is better suited to small text-based classification problems than neural networks due to its unique properties, however it is also recommended that these methods are tested against one another.

2.3   Medical Cannabis and its Pharmacology

In order to utilize text analysis, understanding the potential results is fundamental, and therefore a good understanding of the pharmacology of medical cannabis is required prior to interpreting the results of this analysis. It is important for the reader to familiarize themselves with the key aspects of the complex pharmacology of medical cannabis, in order to maximize the utilization of these results practically. I will attempt to not go into too much detail, and keep the summary to less than two pages. This may seem to be going in depth but it is contrast to the thousands of journals all focused on this incredibly deep field. All the scientific terms discussed below are key metrics and serve as the 'target variables' in later analysis, hence the need to explore them and their functions meticulously.

Research surrounding the Endocannabinoid system as triggered enormous interest of the effect of physiological functions as well as the promising potential for drugs to modulate the cannabinoid receptor activity, especially to do with the nervous system (Stephen D. Skaper, 2012). Vincenzo di Marzo goes on to remind us that the brain is the 'last great frontier of science'. Disorder relating to the nervous system account for more chronic suffering and causes for people visiting hospitals than any other disorder. The ECS is involved in the development of the nervous system, the bodies homeostasis, the energy balances and the more knowledge and research there is done about the ECS the greater advantages humans have to capitalize on therapies based of the endocannabinoid system for the context of many diseases. (Stephen D. Skaper, 2012)

All people are equipped with an endocannabinoid system (ECS), which is a cell-signaling system in the body and is extremely complex.

Figure 2 below shows an example of how endocannabinoids promote balance within the body.
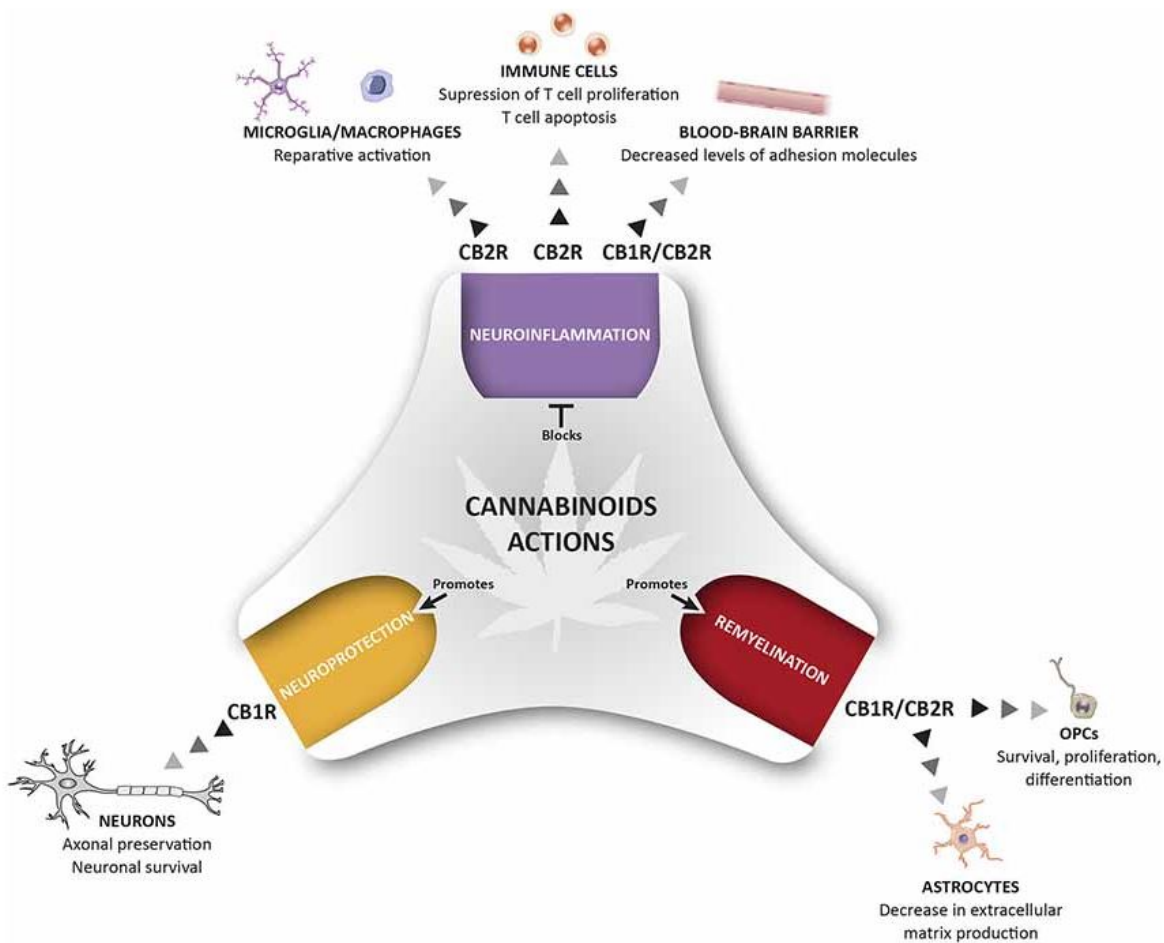
Figure 2: Cannabinoid Primary Actions (Mecha,2018)

The ECS involves three core components: endocannabinoids, receptors, and enzymes. Scientists have identified two key endocannabinoids so far, namely anandamide (AEA) and 2-arachidonoylglyerol (2-AG).

They have also identified two main endocannabinoid receptors: CB1 receptors, which are mainly occurring in the central nervous system, and CB2 receptors, which are mainly occurring in the peripheral nervous system, especially immune cells. These receptors are found throughout the body and ECB's bind to them in order to signal that the ECS needs to take action. The effects that result depend on where the receptor is located and which endocannabinoid it binds to. Figure 3 below illustrates where the CB1 and CB2 receptors are located all over the body.
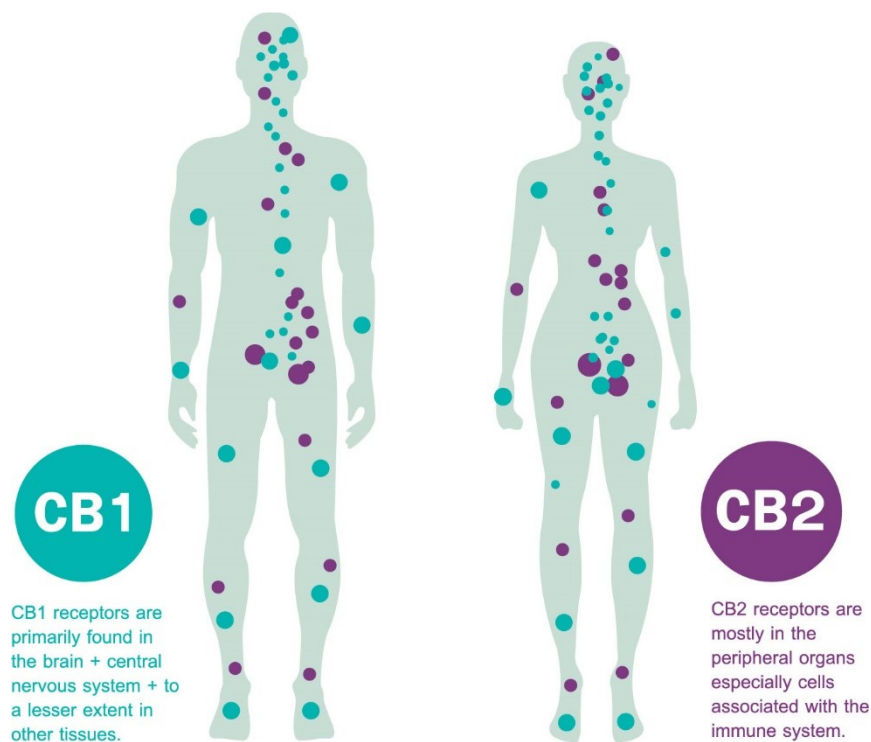
Figure 3: Cannabinoid Receptors CB1 and CB2 around the body(C.Luque,2018)

It has been projected that migraines, fibromyalgia, irritable bowel syndrome, and related conditions represent clinical ECB deficiency syndromes (CEDS) (McPartland JM, 2014). It is also proposed that incorrect and deficient ECB signals could be involved in the development of depression. Schizophrenia, multiple sclerosis (MS), Huntington's disease, Parkinson's disease, anorexia, chronic motion sickness all also seem to be related to ECB deficiencies (McPartland JM, 2014).

The CB1 receptor is the most abundant, and it is a G-protein-coupled receptor, which gets expressed in the central nervous system (CNS), found all over the body as seen in Figure 3. CB2 receptor is particularly associated with immune function and the cells that control it, however they can also be seen to be expressed in the CNS Enzymes also play a role in the ECS. There are two main enzymes responsible for this: fatty acid amide hydrolase, which breaks down AEA and monoacylglycerol acid lipase, which typically breaks down 2-AG.

Experts believe that maintaining homeostasis is the primary role of the ECS, which is linked to the following processes: appetite and digestion, metabolism, chronic pain, inflammation and other immune system responses, mood, learning and memory, motor control, sleep, cardiovascular system function, muscle formation, bone remodeling and growth, liver function, reproductive system function, stress, skin and nerve function.

Cannabis Sativa (marijuana) contains over 100 naturally occurring cannabinoids, and current research indicates that there are many still to be discovered. This allows the consumption of cannabis to have a range of effects on the body and mind as they interact with the ECS, some more desirable than others.

Using cannabis medicinally is a rapidly evolving space, and the quantity of research is increasing rapidly. This increase in research is partly due to the increase in use by medical professionals and the legalization of the most commonly occurring cannabinoid tetrahydrocannabinol (THC) around the world. Cannabinoids are currently used to treat a range of conditions, which include: anxiety, inflammation, nausea, tumour growth, multiple sclerosis, appetite stimulation.

With regards to enzymes, the eCB ligands are anandamide or AEA and 2-AG. These enzymes are released when there is demand from a cell membrane phospholipid. The eCB system has grown in depth after discovery of secondary receptors ligands and ligand metabolic enzymes. Without going into too much depth, these enzymes can be enhanced by compounds known as 'entourage compounds. These compounds interact with other cannabinoid molecules that inhibit hydrolysis, and extend the action through a synergistic effect.

N-palmitylethanolamide (PEA), N-oleoylethanolamide (SEA), and cis-9-octadecenoamide (OEA or oleamide) are entourage compounds and there is evidence to suggest that it represents a route for molecular regulation of endocannabinoid activity, they are key 'target variables' seen in the next subsection (Ben-Shabat S, 1998).

THC, which is psychoactive, and the major component in cannabis sativa, is mediated by activation of the CNS via the Cb1 receptor, however due to the sometime undesired psychoactive effects this mechanism is limited. There are other phytocannabinoids with a weak or zero psychoactive properties which show a lot of promise to be therapeutic agents. (Bridgeman, 2017). For example, unlike THC, CBD provokes its pharmacological effects without stimulating the CB1 and CB2 receptors. There are antiepileptic, antipsychotic and anti-inflammatory effects from the use of CBD which has been proven via clinical trial and has started occurring in registered pharmaceutical drugs. CBD combined with THC has been given approval in many EU countries and is being studied by the FDA in registered trials.

There is an exceptional amount of depth to which this topic can be explored, hence the thousands of journals investigating them, and hence the reason for this thesis. We won't go into more depth on the pharmacology of cannabis, but rather explore the scientific methods surround the analysis of all the text data being produced studying these topics. The next subsection, GH Medical Results, primarily deals with all of the topics covered above, and are the results of manual research which has been taking place over the course of decades.

2.4   Concluding from Related Work

It is clear from the above related work, that all of these ideas and methods are relevant to testing the paper's objectives and they may all hold some significant relevance. However, practically it makes the most sense to develop a strategy which uses the main advantages of various methods to achieve the desired outcome.

The development of the correct feature space is noted as a key variable in the success of a method, and the representation of words in vector space holds the potential for more than one use. Words in a vector space generated through Glove and word2vec models can replace the conventional Tf or TfIdf formats in being as used as features for model inputs and it

is logical to adopt this advanced method. They can also however act as a vehicle to draw correlation between like terms. These types of relationships are the exact relationships desired for output between the variation of target variables provided by GH Medical as this will allow for the correlation of various target terms to be computed efficiently over a large range and therefore there is potential to adapt the use cases of this feature. The reproduction of highly specific correlations between compounds and their pharmacological effect is thought to be the primary resource for regenerating the manually derrived findings between compunds and their physiological effects, and also seems to be the most effective way to be used as an input to both Supervised Learning (SL) and Unsupervised Learning (UL) techniques. The way these two methodologies feed into one another is investigated and elaborated further in the methods.

Word attributes are high in dimension, very sparse and have low word frequencies relative to the corpus of words they occur in. It goes without saying that the design of clustering and classification algorithms need to work effectively with text data and account for its characteristics is perilous. And therefore, algorithms that leverage the nonnegative sparse features of text are usually preferable. Kmeans and Hierarchical were recommended in the literature to be the most effective, and due to the high similarity in type of data analysed this would be a convenient point to start. Since the clustering is a preliminary step to topic modelling there is some leeway or forgiveness in the accuracy of the results as it is acting as an initial starting point to the final topic models. Topic modelling is one of the preferred and more successful methods for categorizing sparse, highly disordered text data and suits the objective for being able to inform based on the types of categories their collection of research and literature may fall into.

When defining clusters with nonprobability-based methods the cluster number is definitive, with clean cut document divides. Bayesian models reduce the number of parameters to be estimated and work better for smaller datasets which the working corpus can be assumed to be based on its tight specificity and has grounds to be a good method to adopt. In LDA topic modelling, each document in a corpus is modelled as a mixture of topics. LDA needs to have a predefined number of topics as to generate the various distributions, and therefore it is important that the user has a good initial estimate on the number of topics which are to be distributed. This is where the use of clustering algorithms comes in handy before carrying out topic modelling.

The format of these features is usually described as being term frequency or Tf-Idf, but as a novel approach the idea of using the vectorized feature space created through word2vec or GloVe appears to be an interesting route to increasing accuracy.

The topics generated will be used to train the classifiers seeing as the dataset is unlabeled to start with. It is noted that manual labelling by the researchers themselves would be preferred and would ultimately be the most accurate, however the unsupervised learning techniques hold the potential to serve two purposes. They are able to provide useful insights about the data, and display the categories most prominent within the data. The insights can then be used in their own right as GH Medical has expressed the need for the organizing of literature into discrete topics. The categories suggested can be used to train classification algorithms – the purpose of this would be for the researchers to automatically group data (research papers)

without having to train any unsupervised models to discover which cluster new research may lie in.

In this case, a new observation to a classifier would be a newly published medical text journal, which has not been used in the dataset which trained the classification model. With the various learning algorithms, we will need to experiment and decide which are most effective for the task at hand.

The insight available through this review is hypothesized to be sufficient in developing the strategy to answer the research objectives.

3 CHAPTER 3 - DATA AND PREPARATION

There are two datasets which are referred to in the following sections.

The first is Dataset 1, referred to as the corpus, is the collection of documents made up of 250 papers provided by Dr Heeroma at GH Medical and another 250 peer reviewed papers downloaded off the internet. This is the dataset that all analysis and NLP models in this paper are built off.

The second, is Dataset 2 which is a collection of manually agregated results in table form. This dataset represents the correlations between key target variables and diseases. These results have been collected by researchers at GH Medical over a decade. This dataset was not used to build any models, it was used to qualitatively compare the results from the NLP models and determine their accuracy and effectiveness as models.

The titles of the documents used in dataset 1 and the full dataset 2 can both be found in the appendices.

The structure of the Data and Preparation chapter consists of three main parts:

- Presenting the Datasets
- Data Pre-Processing
- Exploratory Data Analysis

3.1 Presenting the Datasets

Dataset 1 – Document Corpus

The dataset which makes up the entire corpus comprises 500 medical and scientific peer reviewed journals, articles, reports and books, all of which investigate the very specific topics which make up the field of understanding cannabinoids and their medicinal effect on the body. These documents were retrieved primarily through portals such as that of The United States National Library of Medicine (NLM) which is a department in the American National Institute of Health.

Areas of interest to GH Medical and the primary focus of these medical papers include the array of diseases which cannabinoids and the plant constituents treat, the cannabinoid receptors in human and mammals, the endocannabinoid systems in humans and animals, plant cannabinoids, synthesizing and degrading enzymes, metabolites, disorders and therapeutic uses, which we will confirm with the aid of EDA in the next section. The dataset is in English only. Understanding the characteristics of the data prior to building models is important in aiding the understanding, and hence helping draw sound conclusions.

Dataset 2 - GH Medical results

Dataset 2 categorizes over 60 diseases and their relative attributes such as the **Receptors**, **Endocannabinoids**, **Enzymes**, **Phytocannabinoids** and **Terpenes** that all contribute to the multifaceted entourage effect within the body as discussed in the previous subsection. These 5 variables are reffered to later as the *target variables*. These are all extremely important,

and form the basis of the second objective, which is to determine whether these manual results which have been aggregated by members of GH Medical over years, can be reproduced by Natural Language Processing. The full dataset of the below sample can be found in the appendices.

Table 2 below shows a sample of Dataset 2. It presents 3 diseases, namely Functional Gastro Disorder, Pain disorder and Parkinson's Disease. Each of these has information on which Receptors, Endocannabinoids, Enzymes, Phytocannabinoids and Terpenes are associated with these diseases and disorders. The complete version of Dataset 2 has information on 60 different diseases.

It is evident from this sample that some diseases have more data inputted than others. This is representative of the whole dataset. Some diseases have as little as 2 attributes, whereas others are fully populated. It is also noted that there is no quantitative results associated with the correlations between target variables and diseases, only that there is correlation.

Table 2: A sample of Dataset 2 - The Manually aggregated findings

Diseases	Functional Gastro Disorder	Pain	Parkinsons
Receptors	CB1	CB1	CB1
	CB2	CB2	CB2
	GPR55	GPR55	PPARÎ±
	TRPV1	PPARÎ±	PPARÎ³
	TRPV2	TRPA1	TRPV1
	TRPV3	TRPM8	
	TRPV4	TRPV1	
	TRPA1	TRPV2	
	TRPM8	TRPV3	
	PPARÎ³	TRPV4	
		Î±2r	
Endocannabinoids	2AG	Anandamide	2AG
	OEA	PEA	Anandamide
	PEA		OEA
	Anandamide		PEA
Enzymes	FAAH	FAAG	
		MAGL	
Phytocannabinoids	CBD	CBD	CBD
	THC	CBG	THC
	CBG	THC	THCA
	THCV		THCV

Data Pre Processing

Since dataset 1 is the only dataset used to build models, it is the only one which has required pre-processing and EDA performed. Dataset 2 is for assessing results qualitatively.

As with any NLP task, the first step is careful preprocessing of the data to get it into a workable format, which ultimately determines the success of the analysis. To load the data into R, for analysis, the first step is to create a uniform set of data, and to transform all documents into a portable document format (PDF). This allowed the function *VCorpus* in R to read all documents and store it in a *Corpus* object.

Once stored as this object, preprocessing of the data into a workable format can commence. This preprocessing uses the package text mining (tm) to map various actions over the *Corpus* object. These actions include the removal of punctuation, transforming the content to lower case, the removal of numbers (i.e. section numbers, page numbers, numerical data) from the corpus, the removal of stop words derived from the tm stop-word library combined with a derived local list, the stripping of white space, and the removal of all characters except alphanumeric. This part is extremely important to get correct, as any exclusion of text data from the dataset which is important will skew results.

For example, the need to keep certain terms which do contain numbers attached to the word, such as for various receptors in the body such as CB1 and CB2 is critical to this study. In order to keep these presents in the dataset, prior to the removal of numeric characters, a transformation mapping is done. This transformation changes the numeric attachment to the word to alphabetic. For example, the receptor CB1 is transformed to CB'ONE'. And the receptor TRPV1 will be transformed to TRPV'ONE'. This allows the removal of all numeric 'noise' in the dataset while still preserving the important meaning which will be used in the analysis. Noisy data is further removed using a list of stop words which hold very little meaning contextually, and white space is removed to improve processing time and accuracy.

Another transformation technique used is the aggregation of synonyms. For example, the word cannabis is of dominant occurrence in the dataset, which is referred to by synonym, depending on the region or type of language used. In South Africa, it is commonly referred to as 'dagga'. It is also referred to as 'marijuana'. These terms all mean 'cannabis' in the context of the data, and thus aggregating these words and replacing them all with the word 'cannabis' greatly helps the unsupervised and supervised learning techniques. Another example is the mention of the primary cannabinoid 'thc'. THC is referred to in the corpus using more than 5 different terms, such as delta9−tetrahydrocannbinol, d9thc, Δ 9thc, tetrahydrocannabinol etc. This effect is more pronounced on the less common words, which have a higher inverse document frequency. For example, the cannabinoid 'cannabichromene' is a lesser-known cannabinoid, and thus is mentioned more sparsely overall, but has a higher inverse document frequency in the documents which it is mentioned. To pronounce these findings, the various synonyms for the word are condensed to the shortly and easily digested term 'cbc'. The same goes for all the cannabinoids investigated. These abbreviations are used in the papers once the full version of the word is used, and thus condensing the format helps to increase the relevance of these features.

Another transformation important to this dataset is the harmonization of character encodings. In some papers, the use of Greek symbol Delta Δ is changed in the Volatile Corpus Unicode utf-8 encoding which reads as (U+0394) and where Unicode Character δ reads as (U+03B4). Due to this happening in the R environment, after converting the corpus to Unicode UFT-8 encoding, we can reintroduce the transformation which allows (U+0394)-9-tetrahydrocannbinol

(i.e Δ 9thc) to be transformed to 'thc'. This is the final method to the pre-processing step, giving our dataset the desired workable format to begin exploratory data analysis (EDA).

Finally, we scale and normalize the data to improve the stability of model learning. This is done to smooth the gradients for the models when undergoing optimisation. The global vector features, term frequency and tfidf featured are standardized with a zero mean and a standard deviation of one.

The boxplot below (Figure 4) shows the corpus in it's vector form, condensed to 20 dimensions for the purpose of this graphic.

Boxplot of unscaled Data   **Boxplot of scaled Data**
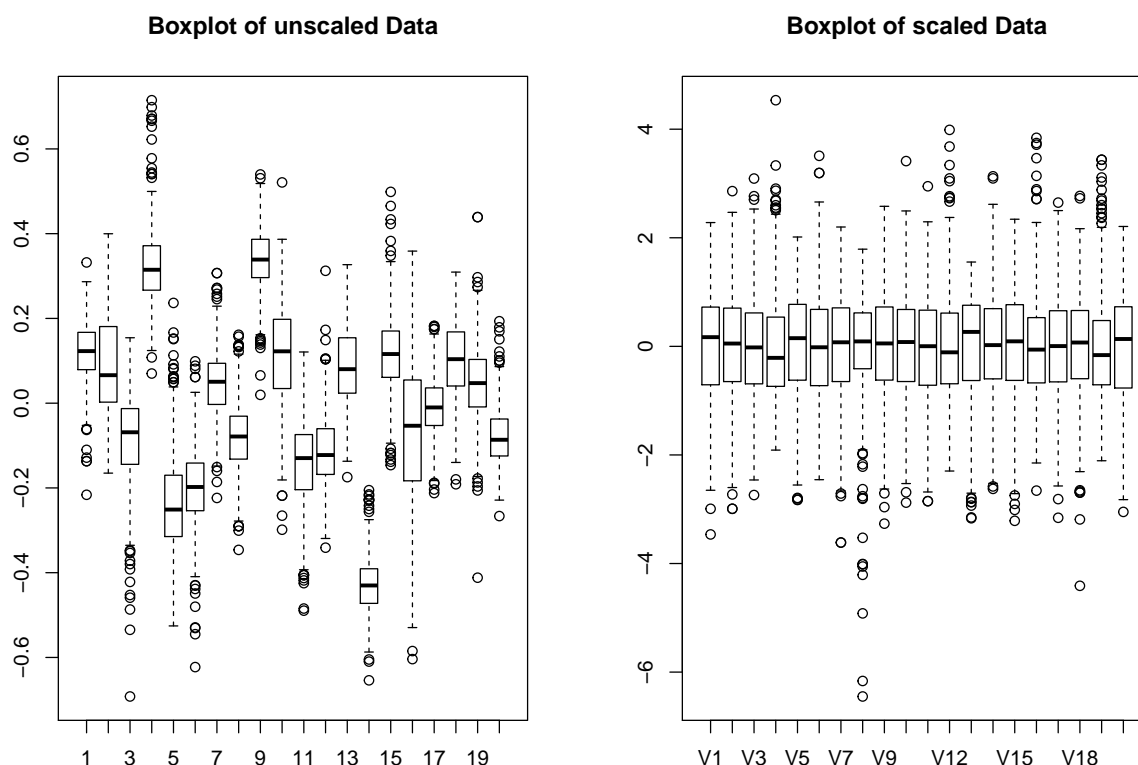
Figure 4: Unscaled vs Scaled data representation

Exploratory Data Analysis

In this chapter, the methods and results from the Exploratory Data Analysis of the corpus are described. The results of the EDA are not included in the model results in Chapter 5, as the EDA is used to direct the hypothesis testing and is not used to answer the objectives. The feature selection phase is not considered to be a part of the EDA as the feature selection is dependent on the model and problem being solved for.

As described in the pre-processing phase, a Corpus object was created and stored. When looking to preform EDA, the format for analysing the data was decided to be in the form

of a Document Term Matrix. This format of storing the corpus allows all the documents to be allocated a row, and every column is the dictionary of terms left after preprocessing. The value in this matrix is the number of times word occurs in the specific document. There are *2.5 million tokens* (words) in the 500 papers, comprising of *112 344 unique terms* used in the Corpus of documents. Considering the Corpus to be an extremely refined collection of papers, all tightly confined to the niche topic of the cannabinoid and endocannabinoid system, suggests the high potential for meaningful analysis considering the relatively small database of 500 papers.

To increase the speed of the EDA processing, the sparsity of the matrix is reduced from 99% to 90%, and the dictionary reduced to only words occurring at least 10 times in the corpus. This reduces the number of unique words to 11 850 words.

Table 3 below gives the key corpus attributes:

<div align="center">

Table 3: Key Corpus Attributes

Documents	482
Unique Terms	112 344
Unique Terms Condensed	11 850
Sparse Entries Full	772075/53377733
Sparse Entries Condensed	581563/5130137
Sparsity Full	99%
Sparsity Condensed	90%
Weighting	term frequency (tf)

</div>

To give the reader a broad understanding of the type of medical paper being analyzed in this report, Table 4 below shows the titles of 5 sampled papers. We can see that the papers belong to highly technical medical cannabis attributes.

Table 4: Sample of Dataset Paper Titles

Title	Author
Cannabidiol, a novel inverse agonist for GPR12	Kevin J. Brown
Beneficial effect of the non-psychotropic plant cannabinoid cannabigerol on experimental inflammatory bowel disease	Francesca Borrelli
Cannabinoids increase lung cancer cell lysis by lymphokine-activated killer cells via upregulation of ICAM-1	Maria Haustein
ACEA (a highly selective cannabinoid CB1 receptor agonist) stimulates hippocampal neurogenesis in mice treated with antiepileptic drugs	Marta Andres-Mach
Alzheimer's disease –mechansims-cause-factors-prevalence	Compaq
Cannabinoid 2 receptor is a novel anti-inflammatory target in experimental proliferative vitreoretinopathy	Anna-Maria Szczesniak
Cannabidiol increases survival and promotes rescue of cognitive function in a murine model of cerebral malaria	A.C. Campos

Principle Component Analysis

We carry out principal component analysis to try reduce the number of dimensions and see if we can explain the variance in the dataset with much fewer dimensions. This allows us to draw meaningful interpretations about the multivariate dataset in 2D, as seen later in topic modelling via the utilisation of Latent Dirichlet Allocation. Figure 5 shows that about 44% of the data can be shown on two dimensions. The Biplot shows that dimensions of the of the document vectors are reasonably well spread throughout the two pricipal components.The Biplot indicates the relationship the variables have on the pricipal compontents.
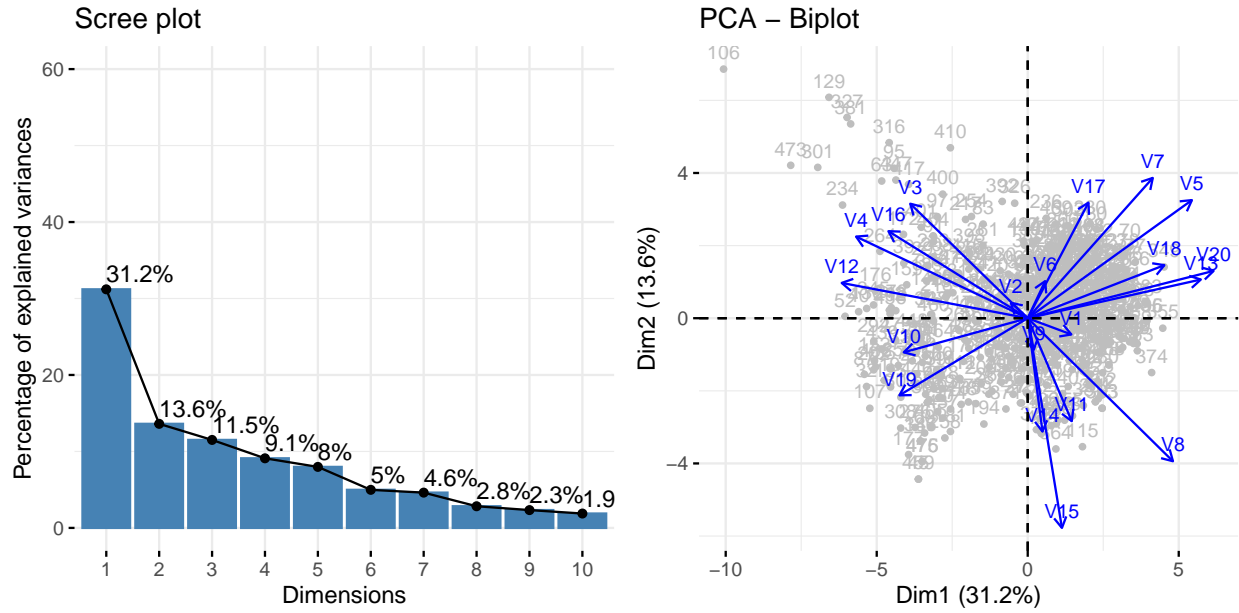
Figure 5: Scree and Biplot

Figure 6 below represents the length of the papers being analyzed. On average, the papers are 4000 words long after undergoing preprocessing and sparse term removal. That gives a total of 2.5 million words in total. 11 850 words are unique and occour a minimum of 10 times in the corpus.
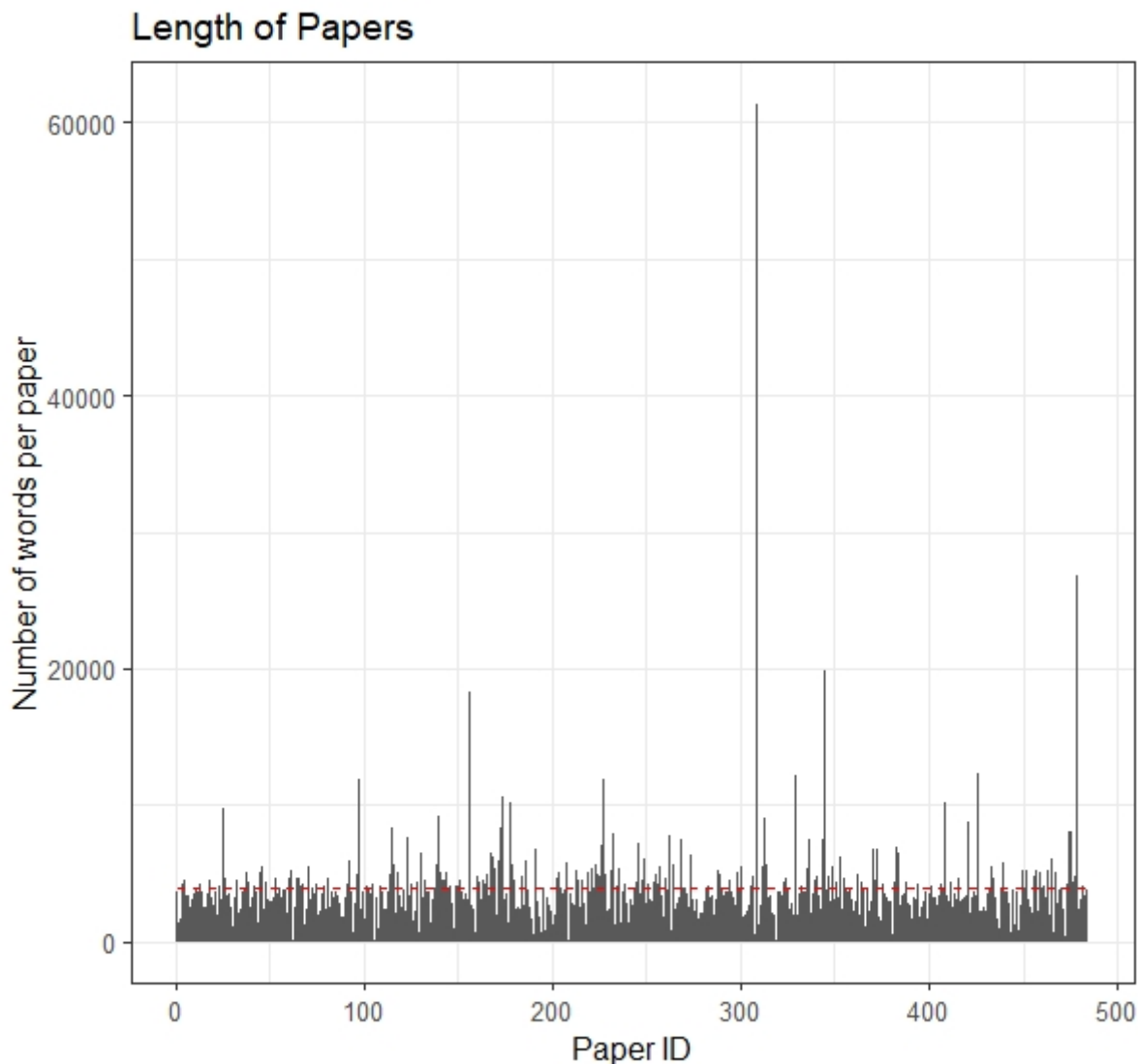
Figure 6: The lengths and mean lengths of Papers in Dataset

To confirm the hypothesis that literature on medical cannabis is increasing rapidly, Figure 7 below plots the number of papers sampled and their date of publication. The graph shows a clear uptrend, and there is expectation that this will grow exponentially in the future. There is a spike in papers published in 2015, which is known as one of the biggest years for growth of the global medical cannabis sphere. This resulted from the largest year of capital injection into the market as recreational and medical use in both the US and Canada became more commercialized. Following 2015 there has been a sustained increase in published literature and this will further be supported by the commercialization of the industry around the world.
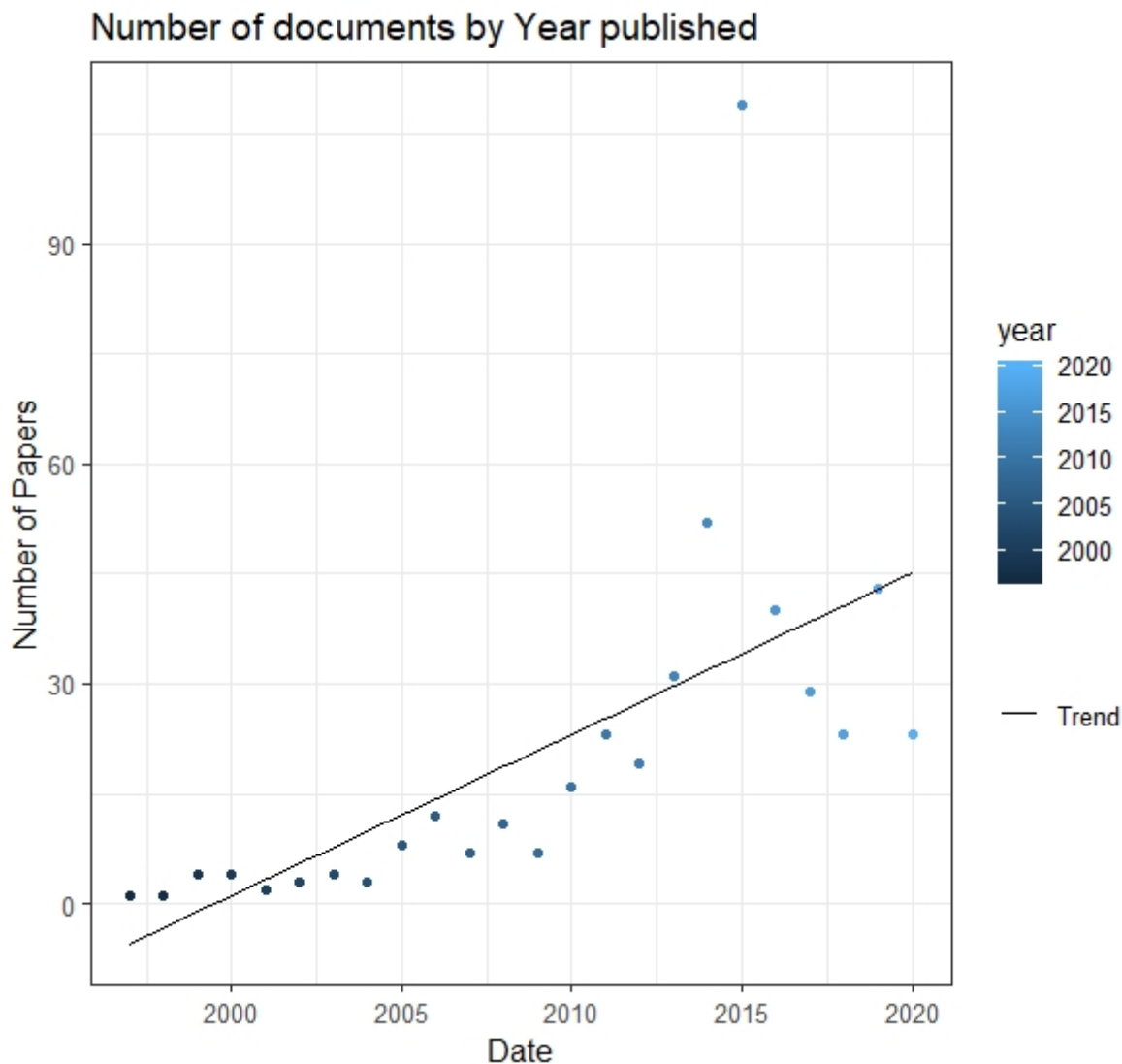
Figure 7: The numbr of papers published by year

Target Variables

As guided by the ongoing project at GH Medical – there are a range of target variables being worked into their 'Cannavigator' software.

As mentioned in the introduction to this chapter, the target variables are attributes which categorize 5 primary groups of biological and compoud specific terms showing strong correlation to diseases studied. These attributes are very important to the study, as they form the primary source of qualitative comparison between the correlations derrived from NLP models and the manually collected results.

These attributes are: *Phytocannabinoids*, *Enzymes*, *Endocannabinoids*, *Receptors* and *Terpenes* and they all describe the *Dieseases* they are correlated too.

We can in Table 5 below that there are 60 Diseases which have been documented, studied and have target attributes associated.

Table 5: List of Diseases used in analysis

addiction	cystitis	morphine interaction	fungi
adhd	depression	multiple sclerosis	non-alcoholic fatty liver disease
aids	diabetes	obesity	alcoholic steato-hepatitis
alzheimers	eczema	ocd	osteoarthritis
anorexia	epilepsy	pain	inflammation
anxiety	gastro	pancreatic cancer	atherosclerosis
arthritis	glioblastoma	parkinsons	colorectal cancer
autism	huntington's	prostate cancer	endometriosis
bladder cancer	hypoxic-ischemic encephalopathy	psoriasis	mosquito larvicidal
bone cancer	insomnia	psychosis and schizophrenia	mosquito repellent
breast cancer	leukemia	ptsd	osteoporosis
bulimia	lung cancer	stroke	mosquito oviposition deterrent
cervical cancer	malaria	tinnitus	gastric ulcer
copd	mdma intoxication	tourettes	melanoma
crohn's disease	migraine	bacteria	sedation

Table 6 below presents the list of Endocannabinoids studied and documented in Dataset 2.

Table 6: List of Endocannabinoids used in analysis

anandamide		oea	dhea
twoag	pea	cbg	epea

Table 7 below presents the list of Phytocannabinoids studied and documented in Dataset 2.

Table 7: List of Phytocannabinoids used in analysis

cbd	thcv	11-oh-9-thc	delta8thc
thc	cbc	thca	cbd
	deltdelta8thc	cbn	thc
cbg	cbcv	cbdv	

Table 8 below presents the list of Receptors studied and documented in Dataset 2.

Table 8: List of Receptors used in analysis

cbone	ppar	twor	trpaone
opioid	fivehtonea	trpvtwo	trpmeight
	gprfivefive	trpvthree	adenosine
cbtwo	trpvone	trpvfour	cbone

Table 9 below presents the list of Terpenes studied and documented in Dataset 2.

Table 9: List of Terpenes used in analysis

caryophyllene	linalool	pinene	menthol
	limonene	eugenol	citronellal
carvone	menthone	humulene	nerolidol
citral	ocimene	terpinene	myrcene
eucalyptol	pulegone	farnesene	pinene
isopulegol	phellandrene	cymene	caryophyllene

Table 10 below presents the list of Enzymes studied and documented in Dataset 2.

Table 10: List of Enzymes used in analysis

magl	dagl	faah	faag

Term Frequency Analysis

As basic EDA, I took a look at the top term occurrence as well as the term frequency *(TF)* inverse document frequency *(TfIDF)* as seen in Figure 8. This gives another indication exactly what the corpus contains.

In order to understand the most important terms used in the corpus, we carried a term frequency analysis and a term frequency inverse document frequency (TfIDf) analysis using the text mining (*tm*) function in R. *Figure 7* shows results of keywords attributes with the highest term frequency and TfIDf values. The terms "thc, cannabinoid, effects, cbd, cells, cannabis, and receptor are all mentioned over 10 000 times in the dataset. Many of the receptors, enzymes, and phytocannabinoids that are not the most spoken about, hold a high level of importance in the documents due to the high TfIDf value.

It is promising to see a range of key attributes (Diseases, Receptors, Enzymes, Phytocannabinoids, Terpenes) occurring in the rank of frequency occurrences. This confirms the analysis is on a very tight collection of documents on the subject matter. The terms "thc, cannabinoid, effects, cbd, cells, cannabis, and receptor are all mentioned over 10 000 times in the dataset. Many of the receptors, enzymes, and phytocannabinoids that aren't the most spoken about, hold a high level of importance in the documents due to the high *TfIDf* value.
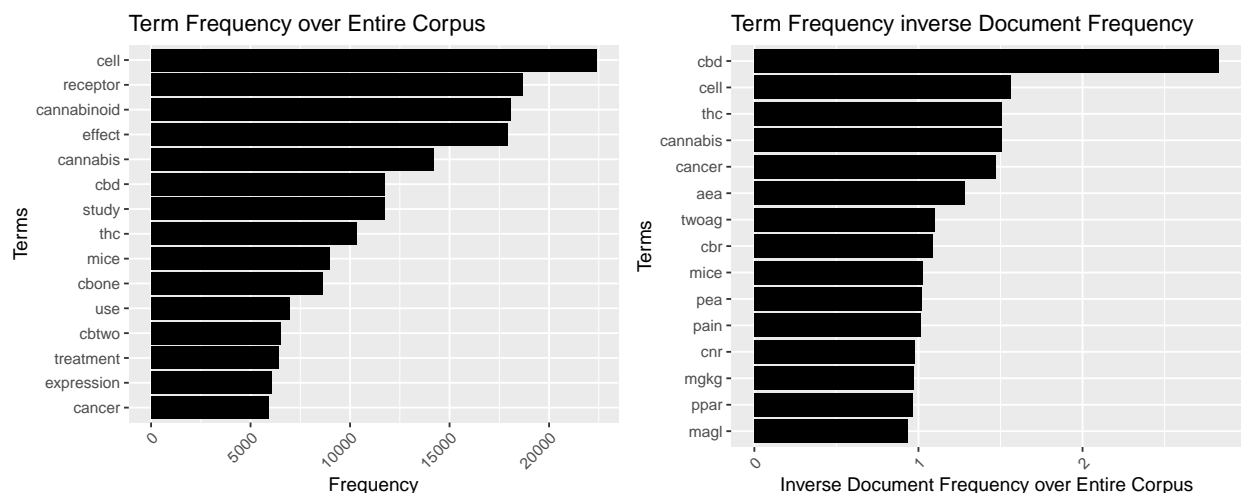
Figure 8: Comparison of the highest Term Frequency and TFIDF Terms

Figure 9 below shows the top sample of *TFIDF* values ranked for each target variables. This graphic displays the most important target variables seen through the dataset. They are as important as they may only occur in a smaller subset of documents, but in those documents the terms were highly frequent.This allows for terms which are only mentioned in a few pieces of literature and have a low relative TF to still be ranked highly if there is focus given in literature to these terms in a few papers.

In the dataset, the most relevent terms for each variable are as follows, in order of importance:

- **Diseases:** Pain, Epilepsy, Inflamation, ADHD, Autism and Anxiety.
- **Receptors:** PPAR, CB2, CB1, GRP55, TRPV1, Opiod
- **Endocannabinoids:** 2AG, PEA, OEA, Anandamide, DHEA
- **Enzymes:** MAGHL, FAAH, DAGL
- **Phytocannabinoids:** CBD, THC, THCV, CBG, THCA, CBN
- **Terpenes:** Eugenol, Caryophllene, Pinene, Limonene, Myrcene, Linalool
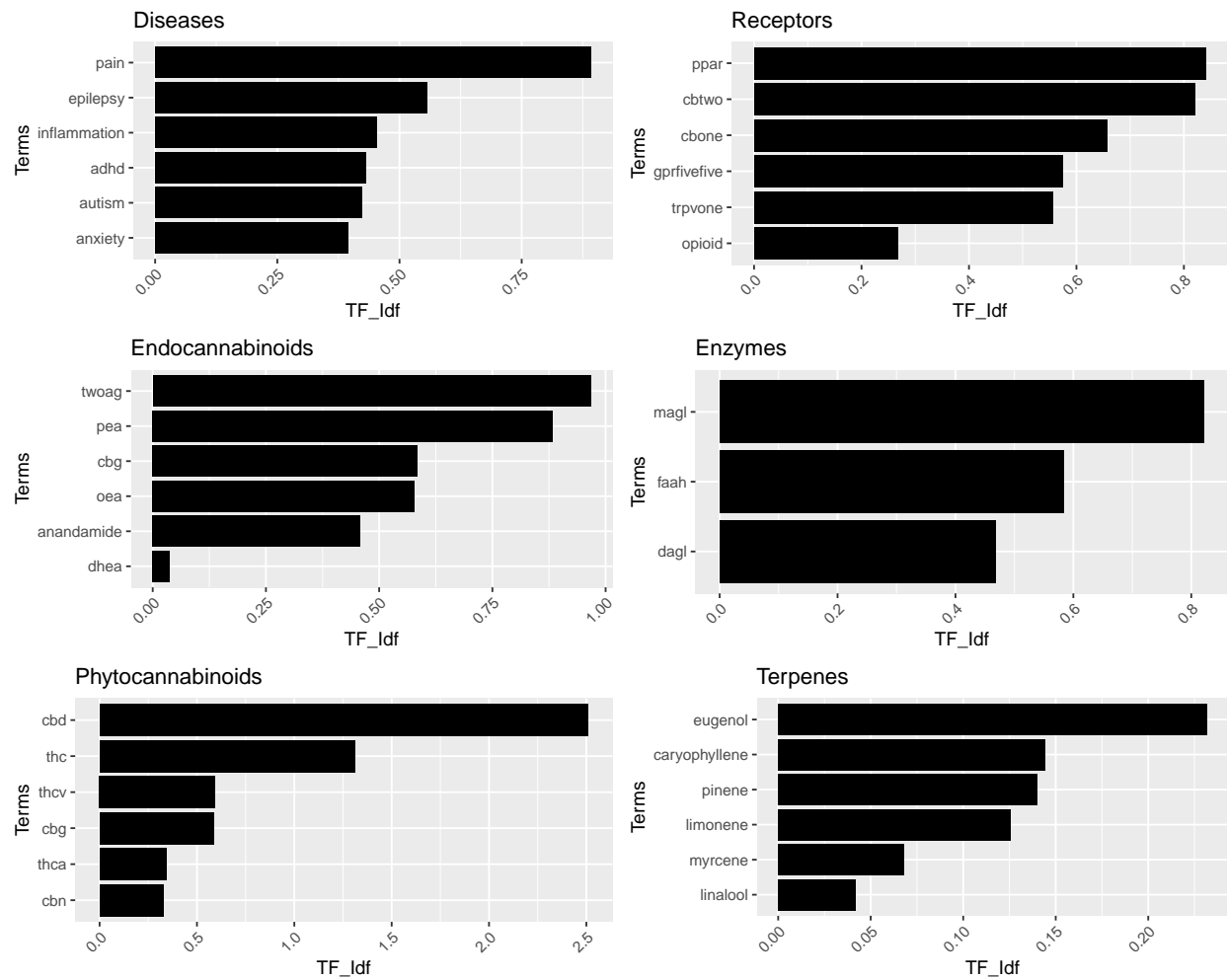
Figure 9: Exploring the highest ranked TfIDF variable feautures

4 CHAPTER 4 – METHODS

In this Chapter, we systematically elaborate on the methods used to answer the objectives of this paper. This methods section is structured to specifically answer each objective and sub-objective.

4.1 Objective 1

To develop an appropriate strategy using natural language processing techniques to accurately group the medical literature according to broad research topics, to analyse the interaction and correlation between cannabis compounds, human physiology and diseases, and to train a classifier to classify unseen documents.

In this objective there are three main issues to be addressed that are categorized as Objective 1's sub-objectives: (1) we seek a technique that we can use to accurately group the medical literature according to broad research topics; (2) we seek a way to analyse the interactions and correlations between cannabis compounds, human physiology and diseases; and (3) we seek to classify unseen documents into the broad research topics discovered.

The institution GH Medical was initially interested in seeing if their manual method of collecting information on the correlation between diseases and cannabis compounds could be replaced by a more efficient system. This can be described by sub-objective (2). Through the process of developing the techniques to tackle this sub-objective, it became apparent that there were other methods that lie within NLP that could assist in other areas of their research. These other methods are described by sub-objective (1) and (3). Collectively, these 3 sub-objectives would be able to provide a complete package of tools to increase the effectiveness and efficiency with which they collect information from literature.

As discussed in Chapter 3, Dataset 1 is the collection of literature from which GH Medical's manual analysis had previously been derived. Dataset 2 represents the aggregated results from their manual findings. Satisfying the reproducibility of Dataset 2 from the literature would adequately satisfy sub-objective (2). Sub-objective (1) and (3) are bonus features that provide an added element to the usefulness of this study.

Below, under each sub-objective we discuss the methods followed and give reasons for their use. We will describe the actual application of these models, and include explanations of model tuning and optimization.

4.1.1 Sub-objective 1 of objective 1

In order to develop a topic model that will help organize literature into discrete topics, and later be used to help train classification algorithms, three methods were followed and tested initially. These methods have been described thoroughly in Alghamdi (2015).

The methods we explored for accurately grouping medical literature into broad research topics included the following: Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). However, after preliminary studies, we settled for LDA as the base method to develop the topic models. We dropped LSA because

although it is easy to train and tune due to the very few hyperparameters, it is not easy to interpret, and does not scale. It is also difficult to determine the number of topics in comparison to the other methods. We also dropped PLSA for this use case because at the level of documents PLSA cannot do probabilistic models. We discovered PLSA is more suited to Image Retrieval and automatic question recommendation than document clustering and hence was not producing results that matched up to the effectiveness of LDA.

We settled for LDA because it out-performed all the other methods on almost all measures of performance. Our initial findings of the performance of LDA are also confirmed by findings from previous work (Sbalchiero, 2020). Stefano Sbalchiero's (2012) paper also concludes that LDA topic modelling as a generative probabilistic method proves to be a superior method for textual and document topic modelling.

As mentioned in Chapter 2 in the reviewed literature, LDA aims to create a probabilistic generative model for the text documents in the corpus, that represents the corpus as a function of hidden variables, that have parameters estimated using the particular corpus (Charu & Zhai, 2018). LDA requires a distribution priori and is therefore a Bayesian version of LSI, and is advantageous as it is resistant to overfitting. It is a dimension reduction approach with rich interpretive quality.

When using probabilistic generative models, the categorization of documents in clusters is less important to the latent clusters of topics and this way of clustering has branched into topic modelling specifically, and has become the most popular approach (Charu & Zhai, 2018).

If there are N documents in a corpus, the dataset of publications is assumed to be generated by k topics. This means that any document can belong to a collection of different topics, all defined by their probability of occurring, reflecting the nature of documents to contain a magnitude of topics and subjects (Charu & Zhai, 2018). For given document Di in a set of topics $Ti\ldots Tk$ the probability of document Di belonging to topic Ti is $P(\frac{Ti}{Di})$. In this case, topics correspond to clusters. $P(\frac{Ti}{Di})$ defines the probability of the ith document belonging to the ith cluster. In contrast, when defining clusters with nonprobability-based methods the cluster is definitive, with clean cut document divides.

LDA models K topics in D documents, where each topic is a distribution over W tokens. If Ω is the matrix of mixed weights for topics in each document, and Φ is the matrix of multinomial coefficients for each topic, then it is possible to generate a model to describe documents. LDA computes the matrix factorization of the feature matrix, in the form of a Term Frequency Inverse Document Frequency. This linear algebra is the product of matrices $\Phi(KxW)$ and $\Omega(DxK)$, that gives the original DTf by Document matrix.

This method allows for a number of topics to be specified by the user and the associated probability distribution of words that occur in each topic is generated. Based on the training data set and accuracy of classification the value of K that pre-determines the number of topics to distribute the words into is set and varied to obtain the correct number of topics. The output is a matrix of nxk dimensions, where n is the number of documents, and k the number of topics. The matric contains the topic probabilities for each document. The allocation of topics was distributed using the highest probability of topic for each document.

The number K of topics needs to be selected in advance, as well as the two important hyperparameters being alpha and beta. Alpha controls the per-document topic distribution and Beta controls the per-topic word distribution. The three controls (K,$alpha$ and $beta$) are varied in a grid search, and the perplexity and coherence between topics are calculated. The results of the perplexity and coherence between topics are used to choose the best model hyperparameters.

Because the value of K needs to be predetermined, it led to the need for an algorithm to help produce a reasonable starting K value. This is where clustering comes in. Clustering is used to try determine a suitable K value, or at least to determine a good initial guess for K.

Topic modelling does not yield complete understanding or generative meaning of the text but gives a well-rounded overview of themes and most probable occurrence that would not have been attained otherwise. DiMaggio et al X indicates the clear distinction that topic modelling provides utility as opposed to purely accuracy. Researchers are needed to interpret the topic model outputs and use it to further improve understanding of the data for further analysis.

Clustering

A principled way to determine the value of K is to use clustering. There are two main methods used in this case, Hierarchical and Kmeans. Given that we do not know the number of topics in a corpus apriori, it is best to use the hierarchical clustering methods first followed by the Kmeans. This is because the Hierarchical clustering method is completely unsupervised whereas the Kmeans requires a hyperparameter, K, equal to the number of topics to be defined a priori. The results of the Hierarchical clustering method can indicate the number of clusters that can then be used in Kmeans to obtain better clusters.

Hierarchical Clustering involves creating clusters that have a predetermined ordering from top to bottom, for example, all files and folders on the hard disk are ordered in a hierarchy. There are two types of hierarchical clustering methods: divisive and agglomerative. Divisive clustering starts with one large cluster and breaks it down recursively into smaller clusters until there is one cluster for each observation. Agglomerative starts with individual observations as single clusters, then by computing the similarity between each of the clusters, the two most similar clusters are merged. This is repeated recursively until all observations are in a single cluster. Ward's algorithm can indicate through the merging cost function the ideal number of clusters: it says, if the cost increases rapidly then it has gone too far and has overestimated the number of clusters. A rule of thumb is to keep reducing the number of clusters until this cost function deviates off its linear trend upwards.

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each observation using a distance function. This matrix displays the distance between each cluster. There are various clustering methods that all differ in how the distance between each cluster is measured.

Leveraging the search process in text clustering is particularly useful, and is made significantly easier with agglomerative hierarchical clustering. Pairwise similarities between groups of documents form the basis of hierarchical clustering, and the main differences between

methods falls on their method of computing these similarities. The most common similarity computations utilize best-case, average-case and worst-case similarities between documents from these groups.

- Single Linkage: When the similarity of two groups of documents is equaled to the similarity of the most similar pair of any two documents from these groups. This method can have a drawback, that is a phenomenon called 'chaining' when dissimilar documents are grouped into similar clusters.

- Group-Average Linkage: When the similarity of two different clusters is the overall average of the similarity between all pairs of documents that can occur between the two clusters.

- Complete Linkage: When the similarity between two unique clusters is the similarity of the least similar documents in two pairs of clusters. This method is set up in a way that helps to reduce 'chaining' as the placement of two dissimilar documents in the same cluster is avoided.

- Ward's Method: Ward's method states that the distance between two clusters is how much the sum of squares will increase when they are merged.

- K Means intends to partition n objects into k clusters where each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of the greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is known as a priori and must be computed from the data.

Kmeans

The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function. K-means is a super-efficient method. The only drawback is that it requires the number of clusters to be specified in advance, and the final results are sensitive to the initialization and often terminates at a local optimum. There is no global theoretical method to find the optimum number of clusters, however, iterating over different values for k is widely used. In general, a larger value for K usually decreases error but increases risk of over fitting. There are a number of heuristics that are used to estimate the number of clusters to use with Kmeans. These include: include the $\sqrt{(n/2)}$ rule, the elbow method, the average silhouette method, and the statistical based Gap Statistic.

Once the value of K is set, the algorithm selects observations at random and iteratively assigns each to a cluster to which it is most similar. The similarity is based on the mean value of the observations in that cluster. At the end of an iteration, the means of the clusters are then updated. The algorithm iterates until the means converge. As mentioned above, methods such as the 'elbow' method, silhouette method, and within-clusters sum of squares (WSS) can be used as methods to determine suggested clusterability. K-means method requires very few iterations to converge which is an advantage over the k-medoids method (D. Cutting, 1992).

4.1.2 Application of clustering and LDA to Dataset 1

Clustering

40

In order to determine the number of optimum clusters present from Dataset 1, hierarchical clustering and Kmeans clustering was performed in R. As discussed in Chapter 3, document term frequency (dtm) matrices were created from the preprocessed text dataset and these dtm matrices were used as inputs to the clustering process. A distance-based matrix was created using the 'dist' function in R utilizing the Euclidean pairwise distances between all features. This distance matrix is needed as an input to hierarchical clustering. The silhouette statistic in R was used over an iterated range of 1:50 cluster sizes to determine what seemed to be the most accurate number of clusters based on the silhouette value.

A loop was then created that varied the method of hierarchical clustering, namely for: Complete Linkage, Single Linkage, Wards Method, Average Linkage and Centroid Method. The cophenetic statistics was then calculated for each of the methods using R's 'cophenetic' function. These computations were then *correlated* to the distance matrix created at the beginning using R's 'cor' correlation function. The result with the best cophenetic correlation to the distance matrix provided the most accurate method of hierarchical clustering. These results were used prior to Kmeans clustering. A similar strategy was then applied to Kmeans clustering, using the document term frequency matrix as an input. An iterative loop was created, looping the Kmeans computation over a group of 1:25 clusters as it was clear from Hierarchical clustering that very little meaning was generated for anything for?? than 20 clusters. For each loop, a cost function is calculated. This cost function is the total within sum of squares (wss), and is generated inside the loop using R's 'kmeans' function. A plot of wss vs the number of clusters (K) is plotted. Similarly, a loop is generated but this time the silhouette value is generated using R's 'silhouette' function. A plot of silhouette value vs number of clusters (K) is plotted.

A vertical line is inserted on the graph where an 'elbow' is present in both the wss and silhouette function graphs, indicating an area of interesting change in the statistic.

A graph for each of these more interesting K values is plotted using R's ggplot function, demonstrating the grouping of documents based on the given K values.

Topic modelling using LDA

There were a number of functions available to do LDA topic modelling in R. The method used was from the text2vec package, which is superior to most. Text2vec implementation is based off WarpLDA which is a state-of-the-art sampling algorithm. It has a sampling complexity that means the run time is not dependent on the number of topics, and the current implementation is single-threaded and extremely fast compared to other packages that were experimented with. This package meant that the same range of K values iterated on for the clustering above could be used for LDA. This was not possible for other LDA packages as with an increased number of topics the processing time increased exponentially.

Text2Vec LDA package improves the log-likelihood with every iteration over the data which is unique to the package. This allows the user to set a convergence tolerance parameter for early stopping should the improvement not be significant enough.

There are several important hyperparameters that can be tuned in the model. These are:

- n_topics - Number of latent topics.

- doc_topic_prior - document-topic prior. Normally a number less than 1 (e.g. 0.1) to prefer sparse topic distributions (i.e. few topics per document).

- topic_word_prior - topic-word prior. Normally a number much less than 1 (e.g. 0.001), to strongly prefer sparse word distributions (i.e. few words per topic).

The perplexity and coherence between the topics were calculated over a cross validated set, that reduces the variability and ensures more reliable results and reduces the probability of overfitting the model. The aim is to produce a set of topics corresponding to low perplexity and high coherence between topics, while making sure the words generated for each chosen topic size makes sense practically.

The primary aim when carrying out this method of LDA is to reduce perplexity and coherence between number of topics while changing 'n_topics' and keeping the other hyperparameters constant. Once having found the range of topics producing the lowest perplexity and coherence values, then the hyperparameters 'doc_topic_prior' and 'topic_word_prior' are adjusted to try and produce a collection of generated words that make the most practical sense within the field of medical cannabis.

The lda$new function in R's text2vec package is used to generate a topic model from selected hyperparameters. This is done for each value of K, with a mental note on the values generated in the clustering section. Each of these models is generated using the dtm matrix as an input feature, with 1000 iterations and a convergence tolerance of 0.001.

In order to assess the words generated from these probability models, the function lda_model$get_top_words are used for each of the models. The models with the highest coherence and lowest perplexity are prioritized and compared to the results from the suggested number of clusters.

These values are assessed quantitatively and the generated subset of words for each topic assed qualitatively, with the hyperparameters adjusted accordingly to produce the final model presented in the results.

4.1.3   Sub-objective 2 of objective 1

To analyse interactions between cannabis compounds, human physiology and diseases we created global vector embeddings using GloVe to serve as the resource where similarities, co-occurrences, and correlations can easily be retrieved and used to draw relationships between target attributes and their variables. This is a relatively novel approach within the medical cannabis space and there were no prior examples in other literature following the same methodology. These Global Vector embeddings are a new way of storing semantic value between words from a corpus of literature, and it was designed specifically for that purpose. Applying it to medical cannabis literature was an attempt to recreate the manually aggregated data shown in Dataset 2 that was collected over a decade. If this method is successful in recreating these results, it yields the potential to rapidly improve methods of literature investigation that focuses on drawing correlation from acute relationships.

This method was applied to the problem by transforming the entire corpus Dataset 1 into a

multidimensional vector space, assuming that this would be a resource that closely represents correlations between like terms. This was validated in a different field of literature by Jeffrey Pennington (2014). Using this assumption, a method was developed around creating a database of all diseases investigated in Dataset 2, and the ranked correlations of all target attributes and their variables. It was assumed that this database, if ordered by rank, and filtered by removing unwanted terms, would produce a list of terms that highly correlate to the diseases in question and should represent the results found in Dataset 2.

This methodology serves as the backbone to the study, and the results are novel. This methodology if successful directly answers the objective of GH medical, that is to improve their method of manual information retrieval, both by accuracy and efficiency.

Working with word embedded Vectors

Before the development of algorithms such as Word2Vec and GloVe, word encoding could be associated with statistical-based models such as, Linear Discriminant Analysis, that uses singular value decomposition on a co-occurrence matrix. Since then, significant improvement to this method has been developed in the form of semantic feature models, that use large amounts of text to identify semantic relationships in a vector space (Colyer, 2016). Geometrically, word2vec is a 'shallow word' embedded model (Karani, 2018). To visualize this, you can interpret these vectors as having a weight in the multidimensional space, and each word surrounding one another in the vector space, or having a similar directional attribute, may be more similar and more semantically correlated than a word on the other end of the vector space with an orthogonal angle between them (M Moradi, 2020).

Word2vec uses a feed-forward neural network (NN), where there are sets of independent variables and target variables that are being predicted for. The word being fed in would be an independent variable and is classified as the target word, and the words that have some sort of co-occurrence to this target word is classified as a context word. The target word is fed into the neural network, through an embedding layer with random weights initialized, as well as through a softmax layer, with the aim of predicting a 'context word' or a word that is seen to appear alongside the target word numerous times in the dataset. Word2Vec produces good results and is ultimately a better feature than a term frequency matrix, but through the investigation of the GloVe algorithm, a better solution is presented.

GloVe does word embedding by either explicit or implicit matrix factorization to word co-occurrence matrices (Alzazah & Cheng, 2020). Unlike the Word2Vec predictive model, it is in fact a count-based unsupervised learning model. This model learns the vectors through dimensionality reduction of a co-occurrence counts matrix. GloVe has Global information, whereas Word2vec does not by default. The intuition of GloVe yield is a higher potential to hang onto meaning in the text through creating a global co-occurrence matrix that estimates the probability a given word will occur with other words, makes Glove often a better option as described by Colyer (2016).

Selivanov (2015) investigates the comparisons between word2Vec and GloVe global vectors, to determine which is the optimum method. The accuracy of results, execution time and RAM occupation were all tested, with GloVe showing considerable advantages on all of the test parameters. GloVe also has an increased number of parameter changes, that help optimise

the method, including early stopping, reusable term cocurence matricies, and incremental fitting. Based on this review of methods, the chosen method for creating word embedded vector models is the **GloVe** algorithm.

Jeffery Pennington et al (2014) describe that the design is done so that the differences in vectors capture as much meaning as possible, that can be demonstrated by the association of two words. This can be shown in the simple example demonstrated in the Stanford university official release of GloVe, that shows the interesting visualisations as seen in Figure 10 below.

The correlation between the vectors can be done in many ways, but as explained (M Moradi, 2020) in his paper about summarizing biomedical articles using the domain-specific word embeddings, the most effective method to calculating the correlation or 'similarity' in these terms is through the cosine distance between vectors. The similarity of a word, that is in the form of a multi-dimension embedded vector, can be calculated with the cosine similarity. This is computable on a large scale and in the case of GloVe is shown to accurately identify correlations in words and maintain a high level of semantic integrity.

Depending on the set context parameter (window), GloVe first constructs a [word x context] co-occurrence data. This matrix is incredibly large in comparison to a standard co-occurrence matrix and so it is factorized to a lower dimension [word x feature] matrix, where each word's feature consists of a vector. This can be done by minimizing a 'reconstruction loss function' In GloVe's instance the matrix is preprocessed and by normalizing and log-smoothing them (Jeffrey Pennington, 2014).

Glove essentially is a log-bilinear model with a weighted least-squares objective (Jeffrey Pennington, 2014). The weighted least-squares model is used to help deal with the co-occurrences that happens very rarely. These create noise, and carry less information than the more frequent co-occurrences.

The figures below demonstrate the linear substructures captured in GloVe embedding.

Figure 10: GloVe word embedding visualized in vector space (left: man-woman; right: comparative-superlative (Jeffrey Pennington, 2014)

The figure above illustrates a sample of vectors and their correlations. It demonstrates the semantic correlation between monarchical terms and their gender. The correlation is so strong that even word-vector addition and subtraction can result in logical language outputs. Initially, this is hard to wrap one's head around, as it would make no fundamental sense to be able to add and subtract words with one another. For example, one wouldn't assume the Sum of word vectors 'Man, Woman and Queen would give the word vector value for 'King', but yet is does when vectorized with GloVe. This was grounds to develop a method around the hypothesis that the vector manipulation of medical terms could solve for target variables associated with these medical terms and, at the very least, show a high correlation between terms with strong contextual relevance.

These types of relationships are the exact relationships desired for output between the variation of target variables provided by GH Medical as this will allow for the correlation of various target terms to be computed efficiently over a large range.

The correlation or 'similarity' between word vectors is most effectively calculated through the cosine distance between the vectors as opposed to other methods such as the Euclidean distance that doesn't as accurately capture the 'essence' or the semantic similarity to words around it.

The formula for the cosine similarity of a vector is as follows:

$$
cosSimilarity = cos(\Theta) = \frac{(A.B)}{(||(|A|)|||B||)} = \frac{\sum\limits_{i=1}^{n} AiBi}{\sqrt{\sum\limits_{i=1}^{n} Ai^2}\sqrt{\sum\limits_{i=1}^{n} Bi^2}}
$$

45

The cosine similarity ranges from -1 to 1. With -1 being opposite similarity and 1 being complete collinearity. The package text2vec that is used to compute the similarities uses a normalized adjustment to this method where the $similarities = 1 - (-cosim)$ yielding a positive number between 0 and 2, the *lower* the number (closer to 0) the more correlated the vectors.

4.1.4 Application of word Embedded vectors to address sub-objective 2 of objective 1

Creating and analysing the embedded word vectors

Using a vector space of medical terms and deriving meaning through correlation is the most exciting part of the methodology, and in our opinion provides the largest scope for future work and discovery.

In order to develop of GloVe embedded model, there are a number of steps required to follow to get the correct inputs. The library 'text2vec' is loaded and comes with the GloVe preset functions. Firstly, the tidy corpus that is produces in the initial preprocessing of the corpus from Chapter 3 is used, and stripped into individual words. This database is then stemmed, and tokenized. The 'create vocabulary' function is used to create a text2vec vocabulary object. The vocabulary can then be pruned to exclude terms that occur less than a certain number of times. This is one parameter used in the tuning. We cannot create a meaningful word vector from words that are hardly seen, and thus the minimum term count is set to 10. The filtered vocabulary is then sent through the 'vocab_vectorizer' function, where each word is embedded into a unique vector. This vector feature is then used to create a term co-occurrence matrix. The 'create_tcm' function creates the term cooccurrence matrix. Using this tokenised matrix, the word vectors, and a skip gram parameter value parameter is set. Skip gram value lets the algorithm know the window size around each term to include in each word's specific co-occurrences. Text2vec uses a parallel stochastic gradient descent algorithm to factorize the term cooccurrence matrix (Selivanov, 2018).

Using the 'GlobalVectors$new' argument we can retrieve the vectors. The 'rank' is a set parameter, that determines the dimensionality of each word vector. In order to preserve computational space, the dimensionality of each word vector is limited to 50 dimensions. Any number greater than this showed no signs of improving the accuracy of correlations. The result is that every word making up the corpus is now in a 50-dimensional vector.

This is the foundation for the discovery of the correlations between certain terms, namely the target attributes and their variables.

Analyzing the correlations between cannabis compounds, human physiology and diseases

In order to analyze the correlations between the diseases specified in Dataset 2, and the target attributes, a database of word correlations needs to be created.

At first, Dataset 2 is wrangled into its separate parts, that being '*Diseases*', '*Phytocannabinoids*', '*Terpenes*', '*Enzymes*', '*Receptors*' and '*Endocannabinoids*' and the values for each of

these attributes is stored in a list.

Once this is done, we use R's 'sim2' function that allows for the computation of vector similarities. The x variable fed into sim2 is the entire word vector object calculated above from GloVe, and the 'y' variable is the word that the similarities to all word vectors is desired. By simply arranging this by magnitude, the highest correlated words in the corpus to specified variable 'y' is displayed. Seeing as we are interested in the correlations of target attributes to diseases, a for loop is generated that all instances of the target attributes and their similarities to the disease corpus word vectors are computed. The results are arranged by magnitude and stored in a database. This database now holds all desired similarities between diseases and target attributes based on their correlation in the global vector space.

For 60 diseases listed in Dataset 2, with 6 target attributes, and over 10 variables within each attribute we can create over 360 unique tables representing different types of correlations. Immediately the scale of the findings is apparent – it is noted at this point that it will not be possible to manually interpret all of the results qualitatively and that a sampling method will be required to justify the effectiveness and validity of correlations.

Having a database of correlations between all target attribute variables and diseases, meant an attempt at validating this method could be attempted. The validation of this method is described in Objective 2, and it compares the results from the correlation database created to that of Dataset 2. This method is described as validation as it proves whether the model reproduces the manually derived results.

4.1.5 Sub-objective 3 of objective 1

A classification algorithm needs to be able to predict to which category a new observation belongs. In this case, a new observation would be a newly published medical text journal, that has not been used in the dataset that trained the classification model. To classify documents, we need to train a supervised learning model based on predetermined outputs. GH Medical has, before this, not categorized any of their literature from Dataset 1 into topics. Sub-objective 1 looked at methods of grouping literature into broad research topics in order to solve this problem. While these generated topics are not validated by the institution themselves yet, the models have been tuned to produce the most accurate research topics based on thorough understanding of the field. These topics are used as an initial document classification for the purpose of this study. In future, these topics can be approved or rejected through manual classification by GH Medical, to improve the classifier built.

In order to build a classifier, we need a feature space to be inputted into the models, that corresponds with the pre-determined output. We have discussed the predetermined output to be that of the topics generated in sub-objective 1. The feature input space however left room for experimentation. Word attributes are high in dimension, very sparse, and have low word frequencies relative to the corpus of words they occur in. As mentioned in the review, the applied classification algorithms need to work effectively with text data and account for its characteristics and be able to leverage the non-negative sparse features of text.

It is typical to use a dtm matrix as a feature input, that represents a document by the words in the document and the number of times these words occur. It is even more common in this case to use the Term Frequency Inverse Document Frequency ($TfIdf$) matrix as an input feature, due to reasons discussed in Chapter 2.

However, due to the investigation into GloVe's global vector, and the documented accuracy presented by this new solution, a new type of feature space was created to compare alongside the more traditional Tf and $TfIdf$ input features. The GloVe embeddings created in sub-objective 2 were used to describe each document, similarly to the way a $TfIdf$ matrix would, except instead of using frequency of words it uses the vector of words to represent the document. This is expected to produce a higher classification accuracy and is the justification for trying this method.

Building a document classifier

A number of classification algorithms were looked at, to allow for experimentation in the methodology. These included: Support Vector Machines (SVM), Generalized Linear Models (GLM) with regularization and Gradient Boosted Machines (GBM). These were the top recommendations from literature surrounding document classification from text data. The literature indicated that Support Vector Machines might be the best technique for this application because of its ability to handle highly sparse and disordered text characteristics.

Support Vector Machines are based on the maximal margin classifier, derived by their support vector classifiers. Support Vector Machines attempt to partition the plane of a data space using either linear or non-linear separations between classes of data in that space. A successful model will select precise boundaries between classes in data, that satisfy classification (C.

C. Aggarwal, 2004). The maximal margin classifier can be thought to be a plane in an n-dimensional space that separates observations into their classes by the boundary the hyperplane creates. In a 2-dimensional space, the hyperplane can be thought of as a line. In a 3-dimensional space the hyperplane can be thought to be a 2- dimensional sheet. For any multidimensional (n>3) spaces it becomes hard to visualize, but the hyperplane takes on a dimension of n-1. The space in between separate classes could theoretically be filled with infinite hyperplanes, however. Maximizing the space between the hyperplane and each dataset is desirable, as this results in the most observations falling within their true class. The hyperplane fitted in the middle of the maximal margin classifier is the maximum-margin hyperplane, and all datapoints falling along this line are classified as support vectors.

The hyperplane created in the dataspace is a key aspect of the SVM and is described below. Take a nxp matrix X consisting of n training observations in the p-dimensional space. This accurately describes the structure of feature space of the text data being analyzed in this report.

$$xn = (xn1, xn2, \ldots, xnp)^T$$

And these observations fall into two classes, namely:

$$(y1, y2, \ldots, yn) - 1, 1$$

where 1 represents one class, and -1 the other. A separating hyperplane is an approach used to classify the test observation based on training data. This would theoretically yield a hyperplane with the property:

$$(B0 + B1xi1 + B2xi2 + \ldots + Bpxip) > 0 \, if \, yi = 1$$

$$and (B0 + B1xi1 + B2xi2 + \ldots + Bpxip) < 0 \, if \, yi = -1$$

The hyperplane is therefore defined by:

$$B0 + B1xi1 + B2xi2 + \ldots + Bpxip = 0$$

If the hyperplane generates two parallel planes that separate the classes of data, the region which is bound by the two planes is called the 'margin', with a 'maximum-margin hyperplane' lying halfway between them. When optimizing the SVM, it is desirable to maximize the distance between the parallel hyperplanes, and thus find the most accurate maximum-margin hyperplane, that is a solution to the optimization problem, and utilizes methods such as Lagrange Multipliers (Er, 2018).

Figure 11 below depicts a simple example of a binary classification problem making use of a maximum margin hyperplane to separate and classify observations.
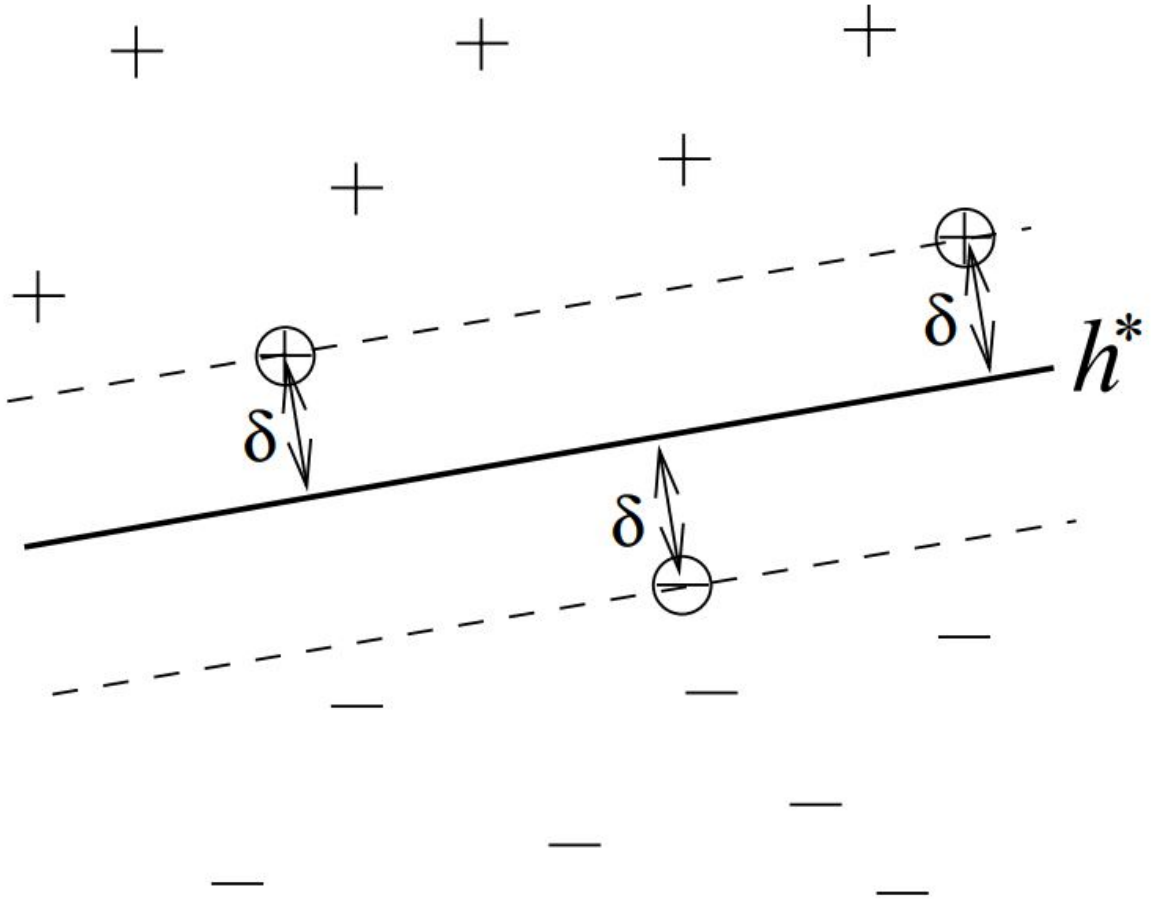
Figure 11: Support Vector Binary classification in two dimesnions (Joachims, 2001)

The hyperplane separates observations while maximizing margin δ. Points on the maximum margin are support vectors and have been illustrated with circles. In an attempt to reduce overfitting, one applies functions that allows some of the observations to cross their margin. There are also cost functions introduced that allows for increased and deceased strictness of a margin, allowing observations to fall on the incorrect side in an attempt to better classify the dataset as a whole, and better predict unseen data. Another characteristic of SVMs that makes them often robust to outliers and more resistant to unimportant data is the fact that only data lying on and within the support vectors affect the performance of the classifier.

4.1.6 Application of the classifiers to address sub-objective 3 of objective 1

The first step to building the classifiers was to define the feature space inputs. The above-mentioned $TfIdf$ feature was created previously and discussed in Chapter 3.

The vector space representation of each document still needed to be created. For this, we used the 'Sofrmaxreg' package in R, that has a built-in function allowing each document vector

representation to be fed in as in input, producing a **mean** vector output of each document. This reduces the dimensionality of each document into a single vector. Due to the rank and high dimensionality of each vector, it preserves the important information within it while significantly reducing the computational space required to process it. This was executed using the 'wordEmbed' function from the 'softmaxreg' package, and setting the variable 'meanVec' to equal TRUE and the entire corpus vector space as the input. The entire corpus of documents is now vectorized and ready to be used as a feature input to the classifiers. Three classifiers were built on a training set of the corpus, and then tested on a training set. Each classifier used the $TfIdf$ and document vector as a feature input, and the prediction accuracy results were compared.

For SVM, the 'e1071' package was used. The hyperparameters were tuned in order to produce a model with the highest prediction accuracy. The cost function helps desensitize the model to overfitting, while gamma helps reduce noise over the support vector boundary. Both the cost and gamma parameters are the most sensitive to tuning. A grid search was performed for both input features, as well as for two kernel types: 'linear' and 'radial', and the models and associated hyperparameters with the best prediction accuracy were chosen to represent the final model.

For Generalised linear model, the 'Glmnet' package in R is used. Glmnet fits a generalized linear model, making use of a penalized maximum likelihood. This regularization is calculated for the lasso penalty at a set of values on the regularization parameter lambda. This method of predictive models is used because of its fast algorithm that is able to take advantage of the sparse input matrix, thus making it highly efficient for text analysis (Trevor Hastie, 2014). A Lasso regularization parameter is worked in the glmnet linear model that is also classified as a shrinkage model. This method improves the least-squares estimation, done by introducing constraints on the coefficients, known as a penalty term. This ensures that the variables contributing significantly remain in the model and those that do not are excluded. These parameters were tuned through a grid search and the parameters that yielded the best prediction accuracy were fixed in the final model.

The 'H2o package' has been used to fit Gradient Boosting Machine (boosted trees) model. The h2o.grid function has been used to adjust the model hyperparameters to identify optimal operating settings for the model, and the best prediction accuracy over the test set determined the parameters to fix in the final model.

The results to these models are presented in the next chapter.

4.2   Objective 2

4.2.1   Compare the correlation results from the global word vector embeddings to the results in Dataset 2 to validate this strategy

In the method surrounding Objective 1 sub-objective 2, a global vector word embedding model was built using GloVe in order to analyse the interactions between cannabis compounds, human physiology and diseases using a database of correlations that could easily be retrieved. This was described as a novel approach and previous examples of this application in this field

could not be found.

In order to validate this strategy, the results stored in Dataset 2 would need to be recreated using this model. If these results were recreated, it would mean that the model effectively achieved the objective and would stand ground to be used in the future to replace the need for manual analysis and aggregation.

A few assumptions were drawn from Dataset 2 that aided the attempt to validate the implementation of this method. These assumptions also helped provide insight into the discovery of new relationships that had previously not been documented.

- (1) It was assumed that if the results from the vector word embeddings reproduced the findings in Dataset 2, the model was successful in drawing the correct correlations between terms.

- (2) It was assumed that if these results are recreated, it would not only provide a method for replacing an old technique, but it would generate previously unknown quantifiable correlations. In Dataset 2, we know only if variables are correlated to specific diseases, without any indication on **how** correlated. The cosine similarity function generates a quantifiable correlation. This knowledge regarding the strength of correlation between terms would add an additional novel element to the research, one previously unattainable through manual analysis.

- (3) It was assumed that there was limited information provided by the manual analysis of diseases and variables because some sections of Dataset 2 were far sparser than others. This indicated that the literature was either incomplete in areas of medical discovery, or that the manual collection of information had not been effective in capturing these relationships. Due to the algorithm built in Objective 1 sub-objective 2, a quantifiable correlation could be assigned to every disease and every target variable. This meant that the Dataset 2 would now be populated completely.

In order to satisfy assumption 1, a qualitative approach is required where the results from the generated correlation database are compared to the results in Dataset 2. If assumption 1 is satisfied, then the model is effectively holding the semantic relationships between diseases and variables associated with these diseases. This would be a huge and novel achievement that gives grounds to accept assumption (2) and (3). If this is the case, there is an opportunity to further this research and collaborate with medical professionals to validate these assumptions.

4.2.2 Qualitatively comparing the results generated to those in Dataset 2

As discussed previously, there are 60 Diseases that hold 5 types of target attributes. Each target attribute is up to 10 variables. From each of the 60 diseases, 5 tables can be reproduced, each displaying the terms correlated to each disease. It was not possible to qualitatively asses 300 tables, and so a sampling method was implemented. From these samples, results were compared to the equivalent set of results in Dataset 2.

The method for sampling is simple. A few diseases are chosen at random, for example, Epilepsy, Alzheimer's, Dementia and Anorexia. Each of these diseases have known correla-

tions to the target attributes 'Phytocannabinoids', 'Terpenes', 'Enzymes', 'Receptors' and 'Endocannabinoids'. Using the above-mentioned correlation database that was generated from the model, the sampled diseases were filtered from the database.

The 5 tables for each disease sampled was then compared to the results in Dataset 2. This was done for 9 different diseases, representing a 15% total sample size yielding 45 individual tables. In future work, it is recommended that all 300 tables which can be generated from the data are used to validate the method as an exhaustive search for any inaccuracies. These tables were assessed qualitatively by comparing the terms present in each table as well as the rank of their correlation. If Dataset 2 only accounted for a few terms, it was a prerequisite that these terms would have to be quantifiably highly correlated in the generated correlation database.

A sample of this sample-set is presented in the results section and then discussed in the discussion.

5   CHAPTER 5 - RESULTS

5.1   Document Clustering and Topic Modelling

Objective 1 - sub-objective 1

Dissimilarity Matrix

The dissimilarity matrix below shows the similarity levels between observations, and this helps us visualize the tendency for clusters to form. Blue areas indicate high dissimilarity between observations.
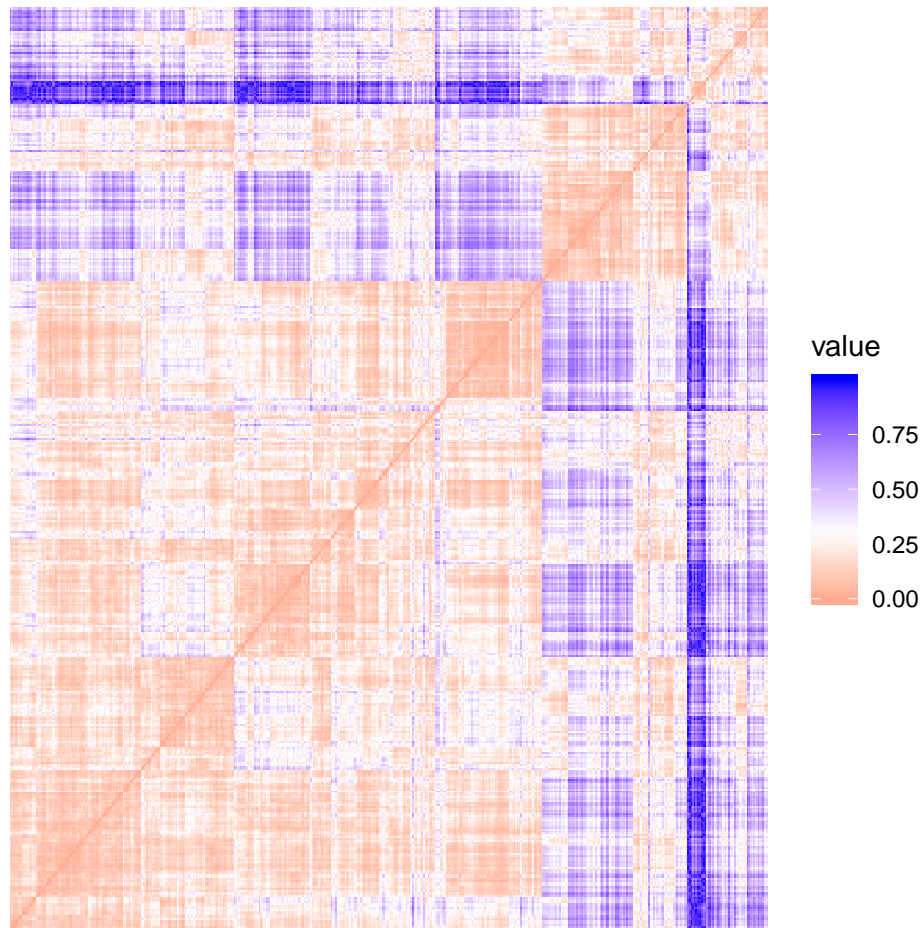
Figure 12: Dissimilarity matrix

Hierachical clustering

Table 11 below shows the Cophenetic Correlation values for various hierachical clustering methods. Wards method shows the highest correlation between clusters.

Table 11: Cophenetic Correlation values for various hierachical clustering methods

Complete Linkage	0.6718010
Single Linkage	0.5733303
Wards	0.7481251
Average Linkage	0.6259721
Centroid	0.5954328

K Means Clustering

Figure 13 shows the total within sum of squares and silhouette method applied to the itterative kmeans approach, interpreted to have 3 areas of greatest change to the WSS and Sil values. These are at 3, 5 and 14 clusters respectively.
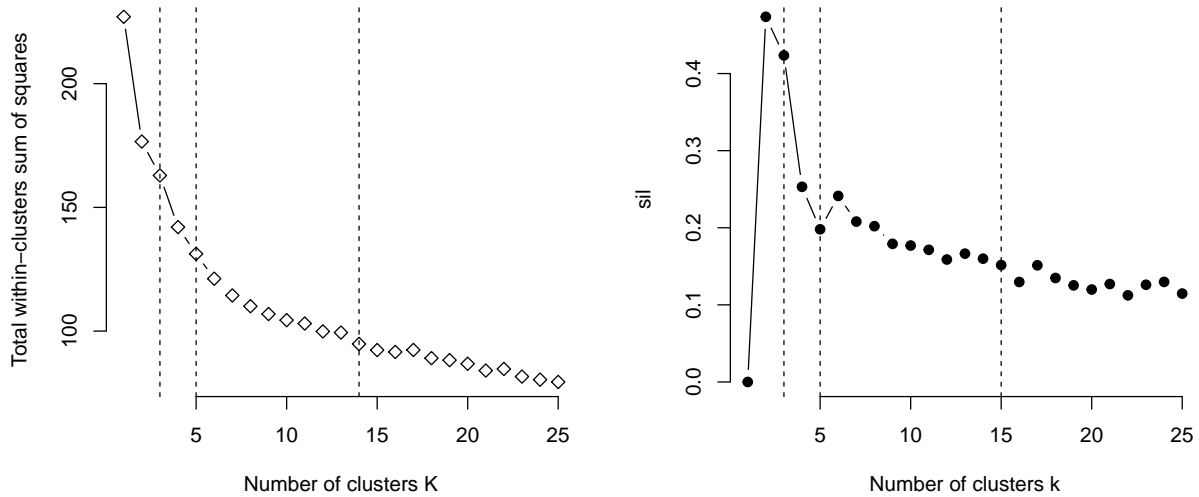
Figure 13: Total within-sum of squares and Silhouette statistics

When allowing the documents to fall naturally into these 3 groups of clusters, the distribution between clusters is distributed as seen in Table 12, 13 and 14. When documents were split into cluster sizes outside of this range there was a clear disparity in the distribution and there would often be very unbalanced grouping.

Table 12: Distirbution of Documents per cluster

Cluster	Number of Documents
1	327
2	121
3	36

Table 13: Distirbution of Documents per cluster

Cluster	Number of Documents
1	68
2	23
3	138
4	98
5	157

Table 14: Distirbution of Documents per cluster

Cluster	Number of Documents
1	8
2	51
3	8
4	49
5	35
6	49
7	58
8	11
9	47
10	28
11	22
12	44
13	54
14	20

Figure 14 below shows the clustering tendancy for the observations over 3, 5 and 14 clusters. It is apparent that for 3, and 5 clusters there are clear cut divides between observation, whereas with 14 clusters there is more overlap. The next section on Topic modelling looks at whether groups of topics of these sizes make sense, and whether they would be suitable broad and more refined research topics.
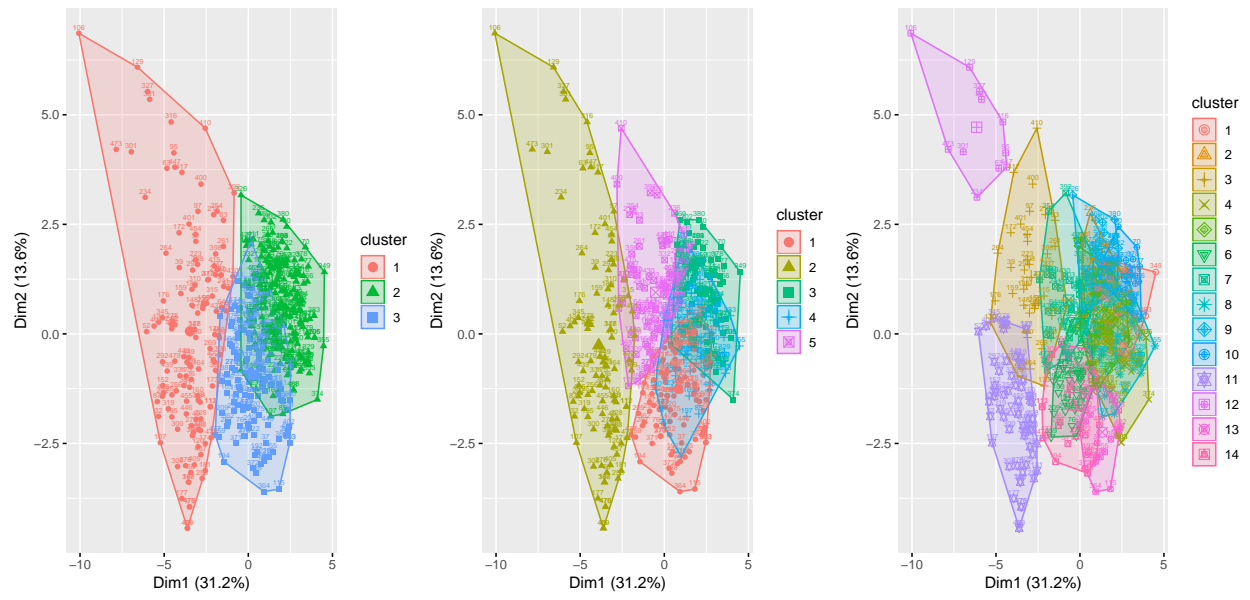
Figure 14: Optimal k values for k-means using WSS and silhouette Methods

Topic Modelling

Using the results from the cluster analysis, LDA was used to develop topic models.

These topic models were iterated from 2 to 20 topics. This range was chosen because the cluster analysis indicated 14 clusters being the highest number of clusters before a smooth reduction in improvement metric. LDA could have been done on these 3 chosen values, however due to the speed of text2vec's algorithm, the LDA models were also iterated and the results between the cluster analysis and topic modelling compared. The coherence between topics as well as the topic perplexity on predicted outcomes were tested. The results showed similarly to the cluster analysis, showing that the groups 3, 5 and 11 topics increased topic coherence and reduced perplexity.

An interactive application was produced with the help of LDAViz package in R to help visualise the results. Screen shots of this application can be seen below these results. This would be used by researchers after loading and analysing text data. LDAViz was also used while generating these results to adjust the relevence metric in order to reduce perplexity and increase coherence for each specified K value.

Table 15 below is a model with K =3, which we call the condensed topic model. It is called condensed as it shows the 3 primary topics over the dataset and would be categorised as the most broad topics in the dataset. We also sugegsted labels for these topics. These labels were determined based on the terms generated from the topic model.

The 3 topcis formed from LDA with K = 3:

- topic01 = Cells / Cancer
- topic02 = Endocannabinoid System / The Brain
- topic03 = Use case / Studies

Table 15: LDA Topic Model using 3 topics and a relevence metric of 0.4 and their sugested labels

topic01	topic02	topic03
cell	receptor	cannabis
cancer	brain	use
expression	cbone	thc
human	endocannabinoid	study
cbtwo	animal	drug
tumor	rats	patients
cbd	mice	cbd
growth	neurons	risk

Table 16 below expands on the size and number of topics by increasing K to equal 5. K = 5 was chosen based on the cluster analysis and reduced topic perplexity through LDA. Similarly to the above,labels were determined based on the terms generated from the topic

model. This model is less broad than the model where K = 3 and includes more topics. This can be observed clearly by the groups of terms generated by the topic model with K = 5.

- topic01 = Cells / Cancer
- topic02 = Use Risks
- topic03 = The Brain/ Neurons
- topic04 = Endocannabinoid System
- topic05 = Animal tests / behavioral results

Table 16: LDA Topic Model using 5 topics and a relevence metric of 0.3 and their sugested labels

topic01	topic02	topic03	topic04	topic05
cell	cannabis	neurons	disease	rats
cancer	use	jwh	faah	administration
tumor	risk	gprfivefive	fatty	mgkg
growth	alcohol	synapses	magl	test
apoptosis	users	sections	pea	animal
breast	psychosis	neuron	endocannabinoid	behavioral
proliferation	sleep	channels	lipid	dthc
death	medical	release	ecs	morphine

Finally, allowing K = 11 as the final topic model. K = 11 was recorded as it was the only K value greater than K = 5 that minimised perlexity and maximised coherenece. This K value represents a less broad group of topics, naturally. This could be observed clearly by the groups of terms generated by the topic model with K = 11. Like the previous models, labels were assigned for each model based on the terms generated.

- topic01 = Psychiatry / Cognitive
- topic02 = Cannabinoids
- topic03 = Macrophages
- topic04 = Plant / Extraction
- topic05 = Cancer
- topic06 = Animal Tests
- topic07 = Using Cannabis
- topic08 = Epilepsy / CBD
- topic09 = Endocannabinoid System / Receptors
- topic10 = Hippocamnpus / Neurons
- topic11= Gut/Metabolic

Table 17: LDA Topic Model using 5 topics and a relevence metric of 0.3 and their sugested labels

topic01	topic02	topic03	topic04	topic05	topic06
schizophrenia	plasma	disease	van	cell	test
psychiatry	cannabinoid	cbtwo	samples	cancer	animal
psychosis	concentrations	immune	method	tumor	mgkg
symptoms	dthc	ppar	plants	apoptosis	rats
disorder	thc	cnr	hemp	growth	vehicle
risk	binding	macrophages	standard	breast	injection
age	membranes	inflammatory	extraction	proliferation	group
cognitive	concentration	inflammation	mass	tumors	morphine

topic07	topic08	topic09	topic10	topic11
cannabis	cbd	endocannabinoid	neurons	gut
medical	effect	twoag	cbr	intestinal
sleep	stroke	aea	neurosci	acids
placebo	preclinical	faah	synaptic	dietary
smoking	epilepsy	anandamide	neuronal	diet
use	exercise	magl	gprfivefive	essential
participants	hta	pea	hippocampal	insulin
fibromyalgia	antidepressant	trpvone	synapses	compounds

Visualising the topic models

The below visualisations (Figure 15) illustrate an interactive model graphically represented by the LDAViz package in R. This allows the user to interactively browse among topics for key words, change the relevence metric, select key terms and determine the topics they occur in with the relative probailities of them occuring in those topics denoted by the size of the circle.

The figure below represents the base visualisation for 11 topics. It shows that almost all topics have their own geometric location on the principal components. Varying numbers of topics over the size of 5 were tested for, with 11 topics giving thr highest coherence and reduced perplexity. This visualisation was done to help verify the selected number of topics which should be used to describe the dataset.
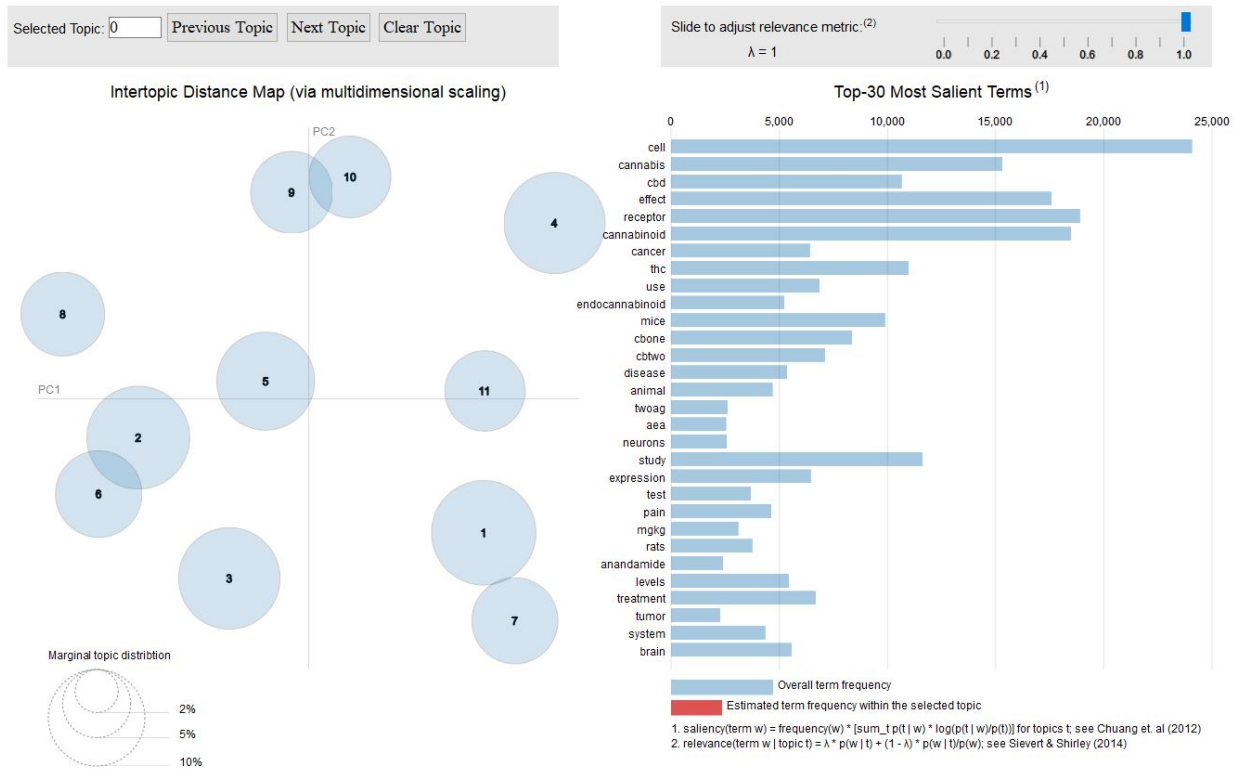
Figure 15: Topic Models with 11 Topics

Figure 16 demonstrates an example of what is displayed when a topic is clicked on. This is the equivalent of filtering out that topic. On the right, the words that make up that topic highlight in red, with the size of the red bar denoting the propability that word occurs in that topic and the grey portion denotes the probability it occurs in any other topics. The relevence metric scale 'tightens' the fit. If it is all the way to 0.1 almost only words which appear in the chosen topic will be displayed. for 1 topics, a relevence metric of 0.3 gave the most interpretable and clear cut topic words.
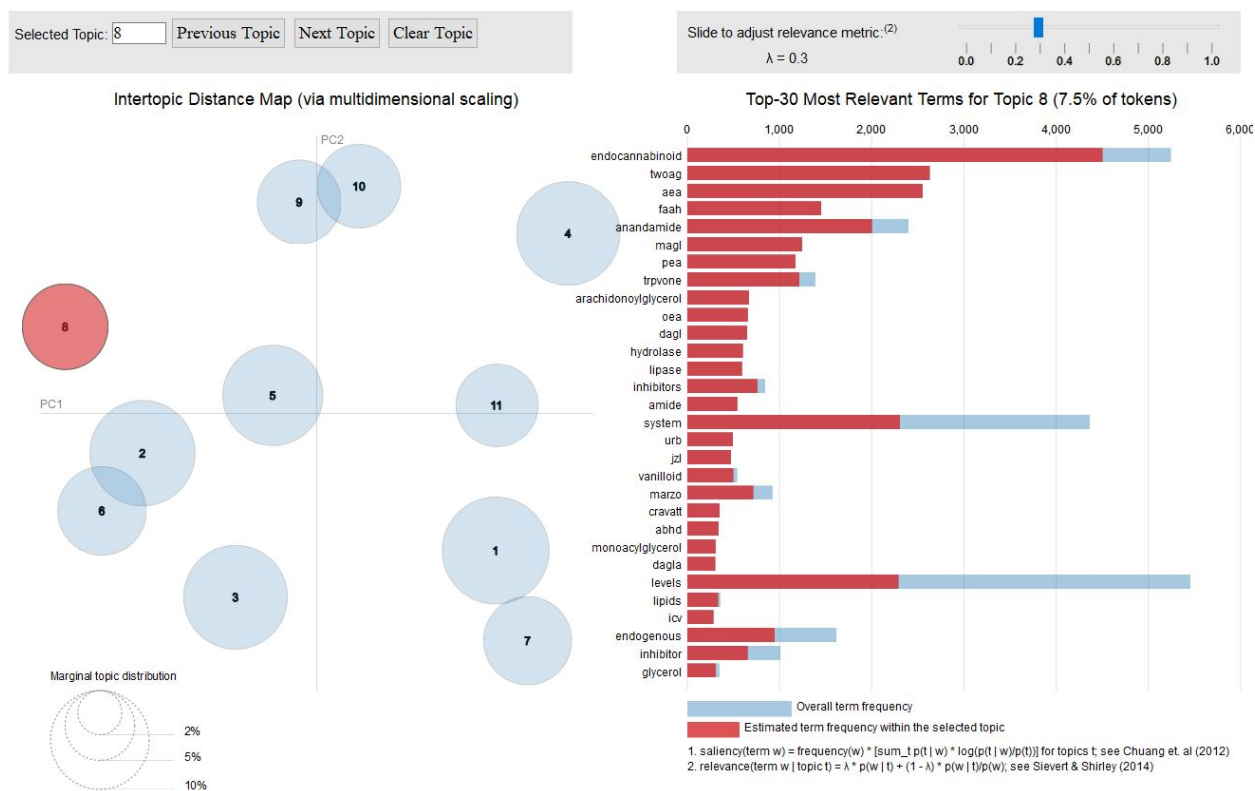
Figure 16: Topic Models with 11 Topics

Figure 17 demonstrates when a word is selected instead of a topic. In this visualisation the word and *Receptor* '**2AG**' is selected. In this instance, the topic which is most probable to contain this word is blown up, with the size of the circle representing the probability that this word only occurs in this topic. **2AG** is a receptor, and it only occurs in topic 9. This tells us a lot about the word **2AG**, as well as telling us a lot about **topic 3**, as well as the **receptors** in general. These reults are used to verify the suggested labels described above for each topic and each varying value K.

It is noted that the topic numbers tabuated above do not correspond to the numbers on the visualisation.

Figure 17: Topic Models with 11 Topics

5.2 Working with word embedded Vectors

Objective 1 - sub-objective 2

The results below are a sample taken from the results described in the methods. 15% of Diseases listed in Database 2 were randomly selected, and tables like the ones shown below were generated for all 15 diseases. The tables shown in this section are a sample of those, giving enough information for the reader. The full sample of the 15 diseases were used to help answer objective 2.

As described in the methods, the lower the value in the adjusted cosine similarity metric, the more correlated the results are. The value allows for a quantitative indication of the correlation.

Table 18 ranks the Phytocannabinoid word vectors most similar to the diseases *Epilepsy* and *Alzheimers*. It shows that Phytocannabinoids CBD, THC, CBC, CBG and THCV are most correlated to Epilepsy whereas the phytocannabinoids CBD, THC, THCV, CBG and THCA are most correlated to Alzheimers.

Table 18: Phytocannabinoid word vectors with highest correlation to the diseases Epilepsy and Alzheimers

Epilepsy	Phytocannabinoids	Alzheimers	Phytocannabinoids
0.7110849	cbd	0.9636233	cbd
0.8667064	thc	1.0696240	thc
1.0265821	cbc	1.1767633	thcv
1.0337075	cbg	1.2643571	cbg
1.0421107	thcv	1.2756995	thca
1.0668913	cbdv	1.3125347	cbdv
1.1027803	cbn	1.3477905	cbc
1.1459775	thca	1.3986223	cbn

Table 19 below looks at the Terpenes. Eugenol, farnesene, caryophyllene and limonene are most correlated to Epilepsy whereas the Terpenes neorlidol, terpinene, farnesene and phellandrene are most related to Alzheimers.

Table 19: Terpene word vectors with highest correlation to the diseases Epilepsy and Alzheimers

Epilepsy	Terpenes	Alzheimers	Terpenes
0.8368667	eugenol	1.051402	nerolidol
0.8423068	farnesene	1.062829	terpinene
0.9350894	caryophyllene	1.088243	farnesene
0.9635637	limonene	1.121631	phellandrene
0.9997576	humulene	1.143409	eugenol
1.0184635	nerolidol	1.191808	pinene
1.0279166	menthol	1.197512	myrcene
1.0524911	pinene	1.205254	caryophyllene

Table 20 does similarly with the Endocannabinoids.

Table 20: Endocannbinoid word vectors with highest correlation to the disease Epilepsy and Alzheimers

Epilepsy	Endocannabinoids	Alzheimers	Endocannabinoids
1.016681	oea	0.8406007	anandamide
1.028039	pea	0.9154874	twoag
1.033708	cbg	0.9415671	pea
1.033951	anandamide	0.9956226	oea
1.051837	twoag	1.2643571	cbg
1.358763	dhea	1.3991464	dhea

Table 21 does similarly with the Receptors.

Table 21: Receptor word vectors with highest correlation to the disease Epilepsy and Alzheimers

Epilepsy	Receptors	Alzheimers	Receptors
0.8572603	ppar	0.7147653	cbtwo
0.8639435	cbone	0.8276139	gprfivefive
0.8858769	gprfivefive	0.8344535	cbone
0.9084535	trpmeight	0.8719171	ppar
0.9099384	cbtwo	0.9376996	opioid
0.9207071	opioid	0.9490466	trpvone
0.9332626	trpvone	1.0692970	trpvfour
1.0023968	trpvfour	1.1236282	trpmeight

Table 22 does something different, which is not comparable to the Dataset 2 provided. It compares the correlation between *Receptors* and *Phytocannabinoids* as well as the corrleation between *Receptors* and *Terpenes*. These are novel results. We can see that Receptor TRPV1 is highly correlated to completely different Phytocannabinoids than that of the CB1 and CB2 receptors which are known to be stimulated by thc and cbd. In this case, the model suggests that cannabinoid cbg is most correlated to the TRPV1 receptor. As en example, the terpenes correalted to this TRV1 receptor are also listed as linalool, humulene and farnesene.

Table 22: Phytocannabinoid and Terpene word vectors with highest correlation to the CB2, CB1 and TRPV1 Receptors in the brain

CB2	Phytocannabinoids	CB1	Phytocannabinoids
0.5343017	cbd	0.5083623	cbd
0.6290925	thc	0.5759289	thc
0.7590253	thcv	0.7020453	thcv
0.7674450	cbg	0.7535954	cbg
0.9004777	cbc	0.8722343	cbc
1.0026241	cbdv	0.9012390	cbn
1.0075076	cbn	0.9126533	thca
1.0226329	thca	0.9634869	cbdv

TRPV1	Phytocannabinoids	TRPV1	Terpenes
0.6177680	cbg	0.9033020	linalool
0.6510847	cbc	0.9170339	humulene
0.6631128	cbd	0.9526343	farnesene
0.7404480	thcv	0.9633812	caryophyllene
0.8284807	cbdv	1.0095352	pinene
0.8475733	thc	1.0174053	nerolidol
0.8573505	cbn	1.0294236	limonene
0.9789560	thca	1.0446433	phellandrene

5.3 Building a document classifier

Objective 1 - sub-objective 3

The table below shows the accuracy using two types of features. The TfIDF takes 20 times longer to run than the condensed document vector features made up from the corpus word vectors. This assisted in the use decision to use document vectors to train the various classification algorithms.

In terms of accuracy, the *TFIDF* features underperformed compared to the document vectors, which contained an order of magnitude less information. It is also apparent that the Support Vector Machine classifier using the document vector feature outperfromed all other classifiers with 95% accuracy on a test set. The SVM classifier also showed the largest difference in performance between features.

Classification Results are shown for K = 11 topics, using the predetermined labels generated by LDA in the section above.

Table 23: Using both document vectors and TfIdf as inputs for classification comparison on 11 topics

	Classification Accuracy (%)
GLmnet(Lasso) Doc Vec	89.04110
Glmnet(Lasso) TfIdf	82.87671
Gradient Bosoted Machines	89.72603
Gradient Boosted Machines	90.41096
Support Vector Machine Doc Vec	95.24793
Support Vector Machine TfiDf	63.01653

The below show's the confusion matrix and accuracy of the best classification algortiihm trained on 70% of the dataset in vector form and tested on 30% of the dataset.

This algorithm was a Support Vector Machine model over the maximum 11 document topics.

5.4 Objective 2

Comparing the correlation results from the global word vector embeddings to the results in Dataset 2

The table below presents a sample from Dataset 2, containing the diseases **Epilepsy** and **Alzheimers** along with the four target attributes **Receptors**, **Endocannabinoids**, **Terpenes** and **Phytocannabinoids**.

This is just two columns of the 45 columns analyzed in validating the embedded vector model. These two columns correspond with the quantitative correlation data extracted in the above results section. The significance of these results are discussed further in the next chapter.

Table 24: A sample of Dataset 2 - The Manually aggregated findings to compare with the vector embedding generated results

Target Attributes	Epilepsy	Alzheimers
Receptors	CBONE	CBONE
	CBTWO	CBTWO
	TRPVONE	PPAR
		PPAR
Endocannabinoids	2AG	2AG
	Anandamide	PEA
Terpenes	Caryophyllene	Caryophyllene
	Eugenol	
Phytocannabinoids	CBD	CBD
	THC	THC
	THCA	
	THCV	
	CBN	
	CBDV	
	Delta8THC	

6   CHAPTER 6 - DISCUSSION

6.1   Summary

The medical cannabis sphere had suffered years of neglect from a post war-on-drugs movement around the world. With the rapid increase in the amount of research being done on cannabis sativa - the plant and its composites have become more widely accepted and studied, so has the amount of written literature consistently increased. This increasing progressive movement of medical discovery has been alongside a time of unapparelled technological advances – especially in machine learning and natural language processing.

Making use of these powerful methodologies within the field of Natural Language processing for the digestion and analysis of the new data is hypothesized to be a key factor in making the widespread understanding and knowledge of this medicine. If this technological power has the potential to improve the standard of living, it should be explored to its full potential. The research in this paper suggests that these theories have grounds to develop these methods and continue this research.

Within the field of Natural Language processing and its application to medical text data discussed in the literature review, there is evidence to show that these methods support

existing theories and applications. The research conducted is novel to the specific analysis using these methodologies in the field of medical cannabis and so the comparison to other examples of this specific application is limited, and thus build on the repertoire of practical applications to these techniques available.

The practical application to the results and methods generated could be useful to the process of conducting Cochrane systematic reviews which is an established process for reviewing high-quality medical research studies to provide accurate advice to medical professionals on the use of new treatments.

6.2   Introduction

When interpreting results, it is always valuable to first assess whether the results have met one's expectations prior to the research, and evaluate whether they have supported the hypotheses. It is also important to contextualize these results within the previous research which has been explored.

Based on this, it is important to be reminded that the purpose of this research was to investigate the application of natural language processing techniques on scientific medical cannabis publications to help solve and assist in a real-world application. This specific application which prompted the research was to directly assist the medical researchers at GH Medical which have dedicated years to understanding the role of cannabinoids in human physiology. These academics realized the pertinent point, which is discussed in the literature on cannabis and its pharmacology, regarding the deep complexity of plant-based medicines and their many natural constituents which vary greatly between sub-species and even growing conditions. The sheer number of variables involved in the therapeutic effect one gets is proving to be the hardest hurdle to overcome with regards to conclusive research. The fact that this is within the field of human medicine also brings ethical concerns into the equation when developing new drugs, as these drugs need to be tested in controlled and reproducible environments such as clinical trials. Based on the number of variables and the growing quantity literature following research, it implies that manual analysis is not a practical solution in the future. It is therefore high on the agenda of medical professionals to solve this multivariable problem and be able to deduce which compounds and which profile of compounds is most likely to have therapeutic effects on the human body. To put this into perspective, using the 6 target attributes, which combined are made up of 40 variables, we can conservatively yield 406 unique profiles of ranked compound concentrations. I say this is conservative as these 6 target variables are only what has been investigated in this paper. And the 40 variables among them is also what has been looked at in this paper. We know there are far more have not been addressed by this data. Considering the concentrations stay roughly constant for each variable in a concentration profile, making this an even more conservative estimation, this yields 4.1 Billion chemical profile variations.

Epidiolex is the first registered pharmaceutical drug containing cannabinoids which was released by UK Pharmaceutical Company GW Pharma. Epidiolex is being used to treat two types of epilepsy. The registration of Epidiolex signified the great shift in the movement to conclude research and release new disease targeted drug formulations using cannabis by

running compound profiles through controlled clinical studies, and eventually documenting these methods in Pharmacopoeias as standardized formulations and methods which can eventually be accepted by local pharmaceutical regulations globally. Epidiolex only contains one cannabinoid, which is cannabidiol. As we understand the chemical effect of multi-compound profile compounds there will be more registered medicines utilizing this knowledge.

The research conducted in this thesis aimed to aid in the process of these discoveries and build on the conclusive scientifically proven formulations. In the discussion below, I chronologically interpret the results which set out to answer the objectives discussed at the beginning of this paper.

6.3 Categorizing literature into broad research topics using Document Clustering and Topic Models

While reviewing literature, it was apparent that the use of Global Vectors in botanical medicine was a fairly novel idea, while the use of standardized text analysis methods was very common. It was apparent that there was more application that could be worked into trying to reproduce the manually derived results from GH Medical, and that there could be extra application which would aim to fulfill another purpose to help researchers in future. This added application would be the collection and classification of the type and class of a document. The logical inclusion of this application allowed for the first objective to be supplemented, making up subobjective 1. The purpose of this was with a bigger picture in mind, looking to the future of improved deciphering of research. Besides collecting correlations between compounds, it would apparently be useful to have a way of splitting large document datasets into refined topics, so that a domain of research could easily be targeted and analysis rather carried out on a smaller subset of domain specific documents.

For example, a breast cancer researcher might want to look at the topic models to confine their search to papers of relevance to cannabinoids and breast cancer, including papers not specifically focused on breast cancer but dealing with receptors and other physiological processes that share commonalities with breast cancer treatment. For this specific application it would be useful to break a dataset of documents into a subset on breast cancer, and then use the Global Vectors analysis on this subset alone to produce the most relevant and productive research. By way of contrast, a pharmacologist might want to dig deeper into the possible correlations between specific compounds through an entire corpus of literature, and in doing so would use the correlation of vectors in a model built over the entire dataset.

For this reason, considerable effort was put into developing techniques to cluster and model topics based on the probability of key terms occurring within these topics, which used various metrics to optimize this process to produce the most certain and distinguishable results.

As seen in **Section 5.1**, three main topic bands were chosen, representing 3, 5 and 11 topics respectively. The combination of Kmeans and LDA unsupervised clustering techniques were used to determine this optimum number of topics. This result produced a great collection of general topics within the medical cannabis field. As expected, the greater the number of discrete topics, the more refined they became. This occurs because the topics are being

generated from the same corpus, and as the number of topics increases, the algorithm is dividing the documents further.

This conveniently results in a broad application and option for researchers, depending how specific and how refined the want to categorize their literature. Labels for each topic are derived manually based on the collection of most probable occurring terms. The topic models were iteratively run to minimize the perplexity, which is a statistical measure of how well the probability model predicts a sample. These methods are used to validate the results generated in the paper and give the best estimation.

For example, table 14 indicates three broad research topics. (1) Cancer and Cells, (2) The Endocannabinoid System and the brain and (3) Use Cases and studies. Whereas in table 16, where the topic model has divided the corpus into 11 topics, we see much more specific topics such as (1) Psychiatry and Cognitive, (8) Epilepsy and CBD, and (11) Gut and Metabolic.

The qualitative analysis on the spectrum of topics justifies the quantitative metrics produced by the models. These topic bands were chosen based on the reduced perplexity values between topics of these sizes while simultaneously increasing coherence and the qualitative analysis on the results is a solid confirmation of this success. This becomes more apparent in the larger topic divisions such as the model with 11 topics. I.e., we know that Psychiatry and Cognitive behaviour are well defined together, we know Epilepsy and CBD are well defined together (as mentioned in the above paragraph relating to Epidiolex epileptic drug which is manufactured with CBD), and we know the Gut and metabolism go together too. This is true for all topics generated and it is promising to see the models accurately recommend these terms as having high probability of occurring in those topics.

Practically, if a researcher opted to look into one of the 11 topics, it would narrow the document cluster down from 500 to between 10 and 50 papers. This is a 90-95% reduction and would allow for the manual sampling of data into a more targeted and relevant subset based on the desired topic of interest. *In doing so, this has answered subobjective 1 of objective 1.*

6.4 Analysing interactions and correlations between cannabis compounds, human physiology and diseases using Global Vectors

The main component of the first objective, which is listed as sub-objective 2 of objective 1, was to identify whether NLP could analyse and reproduce the manually aggregated target variable results from years of research by GH Medical. In order to do this, a new method of analysis relative to medical cannabis text mining was investigated, using correlation between target variables via global vector word embeddings.

After thorough review of the literature, it was apparent that the task of retrieving specific correlations between diseases, phytomedicinal compounds and human physiology would not be easily done with broad and standard NLP techniques such as PCA, clustering and topic modelling. Even though these methods effectively grouped similar and like terms, topics, and even documents, they were unable to extract the semantic meaning and relationship

between certain target variables. This called for an investigation using a completely separate method, which was to vectorize the corpus through complex word embedding techniques such as GloVe. This assessed whether vectorization over the entire corpus, using specialized dimensionality reduction of a co-occurrence counts matrix, would be able to yield actual scientific correlation and meaning from the corpus.

As described in the methods, this embedding was explored for Stanford developed GloVe. This method empirically embeds every word in the corpus into a high dimension geometric vector space based on the context of words in each document, their relevance within the corpus as a whole, and the co-occurrence of these words to other terms. Because these methods are able to hold onto contextual semantic meaning through the mapping of terms onto the vector space, it seemed to be a logical starting point to see whether medical cannabis terms and the correlations between them and diseases could be retrieved.

As described in the methods, a database was created of all correlation values between diseases and target attributes and their variables. In Table: 17, 18, 19, 20 and 21 we have presented a range of examples sampled from this database. These tables Represent the correlation of variables within target attributes to specific diseases. For example, Table 17 describes the correlation between all the available **Phytocannabinoids** to the disease **Epilepsy**. The correlations are ranked highest to lowest, using the text2vec adjusted cosine similarity score, most correlated being the value closer to zero, and least correlated being the value closer to 2. In this table, we can interpret the phytocannabinoids CBD, THC, CBC and CGB to be most correlated to Epilepsy, whereas the Phytocannabinoid THCV is more correlated to the disease Alzheimer's.

Similarly, with Table 18, except here the Terpene attribute and all the variables listed withing terpenes are correlated to the diseases **Epilepsy** and **Alzheimer's**. The disparity between the results is very clear in this example. Eugenol, Farnesene, Caryophyllene and Limonene are represented as the most correlated to Epilepsy, whereas terpenes Nerolidol, terpinene, Farnesene and Phellandrene are most correlated to Alzheimer's.

This can also be observed for table 19 and 20 where Receptor and Endocannabinoid attributes are correlated to these diseases.

Table 21 is unique, and can not be compared to Dataset 2, and this generated table is the beginning of the *discovery* phase of the thesis, and holds the potential for future work. This table instead of correlating attributes to a Disease, it correlates two attributes together. This table specifically correlates Phytocannabinoids to the different Receptors in the brain namely CB1, CB2, and TRPV1. In the final column lists the correlations between Receptor TRPV1 and the Terpenes.

These are novel results, and are not comparable to any provided data as it simply does not exist. The prospect for discovery here is huge, and the scale at which this data can be generated for is mentioned above. Basically, for every variable within an attribute, a table of correlations can be generated for every other variable and attribute. This can scale to thousands of tables, all providing useful insight into the correlation between these compounds, human physiology and disease. The volume of the results that can be generated testifies the possibility of hidden discovery by sheer probability, and this is grounds to propose further

studies and research.

It is noted that the validity of these results would need medical and clinical studies to confirm or reject this hypothesis, however, based on the accuracy of the correlations to the known data as represented in Table 17:20, there is grounds to believe the other correlations generated have a level of accuracy assigned.

6.5 Classifying unseen documents

As mentioned in the methods section, classifying unseen papers into the broad research topics derived from Objective 1 subobjective 1 became an added element of this study, which would have a potential use case when wanting to speed up the process of splitting unseen research papers into these categories.

These topics generated in subobjective one was further used to train classification algorithms so that future unseen literature could be classified immediately into topics without the replication of analysis. It must be noted that these training labels were derived from our analysis, and thus have not had a third party check the validity of these labels.

The supervised learning algorithms used were Generalised Linear models, Support Vector Machines and Gradient Boosted Machines. These methods were all able to predict new outcomes with a level of accuracy, SVM being the most successful. SVM was able to predict on 11 topics using the Document Vector feature input to 95% accuracy and was the winning model. It is noted that it is in the best interest of the model's accuracy to increase the size of the corpus dataset to increase datapoints used to train the model.

6.6 Validating correlation results from Objective 1 subobjective 2 and Dataset 2

This section looks at determining whether the Global Vector solution to drawing meaningful and accurate correlation data is viable going forward. The validation of these results uses a combination of quantitative principles and metrics discussed through the methodology review, and qualitative analysis following logic and assistance from professionals of in the field.

If all results were to be displayed, it would consist of a list 60 objects, each object comprising of 6 tables, one for each target variable. Instead of presenting all of these, a sample of this was analysed and a further sample was recorded in this paper. This subsample is presented in Table 23 of section 5.4, and is a direct extraction from Dataset 2, where the Diseases **Epilepsy** and **Alzheimer's** are shown with their know target attribute correlations.

What this table means to the researchers that developed it, is that for each disease, there are known correlations to certain Phytocannabinoids, Terpenes, Endocannabinoids and Receptors. These correlations however, are not ranked in their level of correlation, and they have no quantitative data to support the strength of correlation.

For this subsample presented, the results generated from the global vector model and presented in table 17:20 are compared directly with table 23. For the Phytocannabinoids, Table 23

has listed THC and CBD as phytocannabinoids most correlated to both Alzheimer's and Epilepsy. In the results generated by the model in Table 17, these two have been quantifiably the highest correlation. The other mentioned phytocannabinoids listed by table 23 are also present in table 17, but what is interesting here is that the results generated by the model have been able to rank these correlations by strength.

Table 18 shows the correlation between Terpenes and these two diseases. Epilepsy is known to be correlated to Caryophyllene and Eugenol in Dataset 2, and in the results generated by the model these terpenes are listed in the top 3 most correlated, and have also been assigned a strength of correlation. What jumps out here, is that the model has suggested that the terpene Farnesene is also correlated to Epilepsy very strongly, however this is missing from the manual results by GH Medical. Due to these correlations being generated by the corpus itself, it would suggest that there is a missing bit of information in the manual results, starting to show where the holes may lie in aggregating this type of data manually. We went and manually researched whether Farnesene is in fact correlated to Epilepsy, and there was strong evidence to suggest this from peer reviewed journal Booth (2019). This validation is solid proof that the global vector model is picking up on correlations that the researchers at GH Medical may have missed.

This is important to the study, because it also suggests that if a value is present in the model results that does not occur in Dataset 2, it does not mean the global vector model has made a mistake, but it suggests that there may be a gap in the manual findings.

This can be seen further when looking at the Receptors in the brain correlated to Epilepsy. Dataset 2 suggests CB1, CB2, and PPAR for Alzheimer's. Table 20 shows the model generated results which confirms these three receptors in the top 4 most correlated to Alzheimer's, validating these results. It also suggests that GRP55 is also a receptor highly correlated to Alzheimer's, a result which is not present in the manual findings.

These validations are scattered throughout the entire dataset of the 300 tables, as well as the sample of 45 tables. For each of these validations, there are suggestions of high correlations that aren't present in the manual findings.

The combination of **quantifiable correlation** between variables as well as the discovery of **previously undocumented correlations** provides a valuable and exciting prospect for the development, improvement and application of these NLP techniques in medical cannabis research

6.7 Limitations and recommendations for further research

All research has limitations. In this section the limitations to this study are explored and, where possible, recommendations are given to address these in future work. The limitations to the methods used, the data which I had at my disposal, and the computational resources are briefly discussed.

Word vector space models and word embeddings in general do come with their limitations. Most significantly, words with multiple meanings can be blended to a single meaning. For example, the word "club" can have numerous meanings in many different contexts - a baseball

club, a club sandwich, a clubhouse, a night club. There is a need to accommodate multiple meanings per word into numerous vectors, which is termed multi-sense embedding.

Another limitation, is the accuracy of the correlations partly depends on the quality of the research papers from which the data is sourced. That is beyond the scope of these NLP techniques to evaluate. Source data should ideally be derived from reputable, peer-reviewed studies, and sometimes be confined to "gold standard" clinical trials, depending on the purpose of the assessment.

In this application of word embeddings and the target variables being specifically derived medical terms, there was less of a risk of the collation of meanings and disambiguation in meaning. However, it would be necessary to train models based on 'Multiple-Sense' embedding architecture to assess whether this directly improves these results and whether there was a proportion of error in the predictions based on the semantic collation of terms used in many contexts. However, it is hypothesized that the collation of meanings does not have as much of an influence as it would normally hold if the corpus of text was made up of a wider and more generalized range of topics.

With regards to a more practical approach to developing this field of research there needs to be a high participation of data analysis and medical research centres such as that of GH medical or other pharmaceutical companies. In order to investigate the limitations of the proposed high correlation between physiology and phytomedicinal compounds it is recommended that partnerships between research institutions and pharmaceutical companies are developed with specific mandates to investigate these correlations. If there are noticeable indications that the data is suggesting plausible treatments and correlations between known active compounds and physiological effects it will give grounds to the further perusing of this field of collaborative research.

With regards to the clustering and modelling of discrete topics in literature it is very limited by the unlabeled data. In future, these results could be further validated by having experts manually review and assign their own labels to the data through a collaborative approach to modifying the algorithm. The combinations of manually assigned labels with the help of initial topic modelling can help to produce sound and concrete labels for medical documents, which can then be used to train classification algorithms with a higher certainty.

With regards to the data available to complete the analysis of this paper, and the notion of 'underfitting', it is logical to point out that a limitation of this research is the size of the data available. This applies to both the number of documents used to build the models, as well as the dataset used to compare results against. There are 28 000 papers listed in the National Library of Medicine which reference cannabinoids and cannabis. It is recommended that a web scrape of this database is performed which isolates and downloads all of these papers to bolster the models built. It is also recommended that resources are allocated into developing the summary provided by GH Medical where the results are compared against. An increase in data available to analyse, and an improvement in the data which is used to compare these results, will ultimately lead to an improvement of analysis.

A further limitation of the analysis is that of computing power. A quad core processer operating with 16gb of RAM was still a imitation in processing large matrices upwards of

7Gb. This ultimately limited the processing and led to the decreased matrix sparsity in order to quickly iterate through various permutations of findings. It is unknown whether these extremely sparse terms may have had an effect on the results when permuting through all variables of interest. The processing power also limited the vector space representations to 50 dimensions deep. This effect of the results is also unknown. It is recommended that future work consists of writing and running iterative results which produce all combinations of correlations of interest through a super computer which could handle the increased vector space dimensionality, an increased feature space to take into account sparsity, and to do this on a more complete full dataset of 28 000 papers. The accurate results from the results produced in this thesis gives grounds to pursue this recommendation as these powerful computers are easily accessible through research institutions.

A final limitation of this study identified is in fact the field in which the research is applied. Plant derived medicine in general has been growing at a rapid rate, and in certain sectors equivalent to that of the growth in medical cannabis. South Africa specifically is home to 10% of higher flowering plant species globally (R Street, 2012). This methodology of analysis is not limited to cannabis only, and provides huge potential to be applied to a range of all types of medicinal compounds in plants and their effect on humans. If the algorithms can be proven by further research and clinical trials with medical cannabis there is a huge potential for these methods to be applied broadly across many divisions of phytomedicinal research.

It was therefore not only important to develop methods for dealing with the available data – but it is recommended to develop methods that can handle the forecasted increased rate of change in quantitative data which will become available in the future, and draw conclusions from this data which is set to positively improve the contribution to science in the fields of medical cannabis and phytomedicinal compounds in general.

Validating the ability to reproduce results aggregated manually through years of research via correlation of terms in a vector space is the most exciting part of the application in my opinion, and holds the most potential for future work and development in the field. As mentioned in the introduction, it is the notion that we are all playing "catch up" of information, research and discovery in the medical cannabis field that is driving the progress and understanding of medical cannabis. The fact that these techniques are able to reproduce manually derived observations which has taken years of research suggests that this could be a route to increasing the rate of discovery.

Doctors and medical practitioners could use algorithms such as these to steer the boat in the right direction with regards to resources and time allocation. As mentioned earlier the combinations of chemical profiles and potential combinations is exponential, and we need to turn to data and tailored algorithms to reduce the chance of hit and miss assumptions, by investigating the suggested correlations presented by the data. This will ultimately lead to improvement in discovery, and would increase the rate of inception of clinical trials. The more studies that are carried out through clinal trials, the higher the success rate of modern medicine development.

7 CHAPTER 7 - CONCLUSIONS AND FUTURE WORK

7.1 Summary

In this paper we examined a corpus of 500 medical research reports, journals and papers which closely look at all areas of the medical cannabis ecosystem, which is particularly focused on the medical use of cannabis and its constituent cannabinoids and terpenes, as well as the connection to the important receptors in the brain and the interaction with the endocannabinoid system.

This analysis was performed with the purpose reproducing findings from years of manual research and result-aggregation performed by Dutch company GHmedical, as well as with the purpose of further developing the research and knowledge surrounding the medicinal understanding of cannabis as a means to treat diseases. Methods of grouping literature into broad research topics as well as classifying unseen documents into these topics was also investigated. In doing so, the analysis tries to increase the efficiency of medical discoveries being drawn through the collective aggregation of global research as a whole.

This research is important as cannabis has become more formally accepted internationally as a powerful medicinal plant, and is becoming globally accepted in treatment methods due to the increase in more conclusive clinical studies. This rapid and sudden increase in the number of studies and amount of quantitative research being performed by the sector has resulted in the abundance of literature, making it hard for researchers and medical professionals to keep up to date with current and new findings.

We aimed to develop an appropriate strategy using natural language processing techniques to accurately group the medical literature according to broad research topics, to analyse the interaction and correlation between cannabis compounds, human physiology and diseases, and to train a classifier to classify unseen documents. This would prove the use of statistic and machine learning models effective and useful to researchers, academics and medical physicians looking to deepen their knowledge and understanding in this field by increasing the rate and accuracy of their research.

7.2 Conclusions

Based on quantitative and qualitative analysis, the results indicate that the developed working code and methods were able to effectively and accurately demonstrate correlation and reproducibility of results which had been previously manually collected and aggregated by GH Medical over a period of many years. This was shown by the close similarity and correlation of ranked target variables to diseases. The unsupervised methods were able to effectively cluster and model topic distributions between the data to group documents by topic, while the supervised learning methods were able to accurately train models based on these labels. In doing so, these methods will allow researchers to fine tune and optimize their time by directing them towards the correct research based on their targets. This analysis supports the theory that the natural language processing techniques explored and decided

upon after thorough literature review, are able to solve for a practical real-world problem, that was not in use at the time of undergoing the research.

Through the in-depth analysis of the data, we began to understand how extensively one could apply the results because of the interplay between the two methods. After a greater understanding of the strengths and weaknesses of the methods we proposed relevant and possible areas of future work. This is a highly significant result that shows promise to considerably contribute and positively impact the scientific development within this field by rapidly improving the rate and effectiveness at which research can be analysed and reducing the amount of redundant information collected through manual discovery.

The results were validated against known correlations, and therefore, the top ranked unlisted medical correlations were determined to have a high probability of being novel, as presented through the vector space. This was therefore able to fulfill the second objective and aid in the development of recommendations for the further application of these techniques practically.

We noted that this thesis cannot produce conclusive medical evidence as ultimately clinical trials would need to be performed in order to confirm the suggested medical correlation results, which is outside of the scope of this paper. However, based on the methodology proving conclusive for known relationships, it is highly suggestive of the currently unverifiable relationships being indicated. This also shows the importance to further develop this research due to the magnitude of potential if these results are taken through to medical trials and yield conclusive findings.

It can be concluded that the developed working code and methods were able to effectively and accurate reproduce the results of key target variables and key terms of interest, as well as their correlations between one another. These methods were also able to effectively cluster and model topic distributions between the data to effectively group documents by topic, and in doing so is able to help researchers fine tune and optimize their time by directing them towards the correct research based on their targets.

This is a highly significant result that shows promise to considerably contribute and positively impact the scientific development within this field by rapidly improving the rate at which research can be analysed and reducing the amount of redundant information collected through manual discovery.

Bibliography

Aggarwal, C.C. 2015. "Data Mining - the Textbook."

Aggarwal, C.C., and S. C. 2004. "On Using Partial Supervision for Text Categorization." *IEEE Transactions on Knowledge and Data Engineering.*

Alzazah, F.S., and X. Cheng. 2020. *Recent Advances in Stock Market Prediction Using Text Mining: A Survey.* London, UK: IntechOpen.

Ben-Shabat, S., and F. E. 1998. "An Entourage Effect: Inactive Endogenous Fatty Acid Glycerol Esters Enhance 2-Arachidonoyl-Glycerol Cannabinoid Activity." In *Eur J Pharmacol*, 23–31.

Bennett, C. 2010. "Early/Ancient History." In *The Pot Book a Complete Guide to Cannabis.* Rochester, Vermont: Park Street Press.

Bridgeman, M.B. 2017. *Medicinal Cannabis: History, Pharmacology, and Implications for the Acute Care Setting.* P&T.

Charu, A.C., and C. Zhai. 2018. *A Survey of Text Clustering Algorithms.* Yorktown Heights, NY: Springer NY.

Claus Asmussen, C.M. 2019. "Smart Literature Review a Practical Topic Modelling Approach to Exploratory Literature Review." *Journal of Big Data.*

Colyer, A. 2016. "GloVe: Global Vectors for Word Representation."

Cothenet, C. 2019. *Short Technical Information About Word2vec, Glove and Fasttext.* Towards Data Science.

Cutting, D., and D. K. 1992. "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections."

Er, S. 2018. "Support Vector Machines."

Heeroma, D.J. 2016. *Cannabis Medically Revisited.* GH Medical.

Jain, A., and R. C. 1998. "Algorithms for Clustering Data." Englewood Cliffs, NJ.

Jain, A.K., and R.C. Dubes. 1988. "Algorithms for Clustering Data."

Jardine, N., and C. J. 1971. "The Use of Hierarchical Clustering in Information Retrieval."

Jeffrey Pennington, R.S. 2014. *GloVe: Global Vectors for Word Representation.* Computer Science Department, Stanford University.

Joachims, T. 2001. *A Statistical Learning Model of Text Classification for Support Vector Machines.* Germany: Sankt Augustin.

Kaufman, L., and P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley Interscience.

Kothari, M. 2017. "Feature Extraction Techniques – Nlp." https://www.geeksforgeeks.org/

feature-extraction-techniques-nlp/.

Luque, C., and J. M. 2018. "An Advanced Review on Text Mining in Medicine."

Mara Chinea-Rios, G.S.-T. 2015. *Sentence Clustering Using Continuous Vector Space Representation.* Pattern Recognition; Human Language Technologies Center.

Mary Barna Bridgeman, D.A. 2017. "Medicinal Cannabis: History, Pharmacology, and Implications for the Acute Care Setting."

McPartland, J.M., and G. G. 2014. "Care and Feeding of the Endocannabinoid System: A Systematic Review of Potential Clinical Interventions That Upregulate the Endocannabinoid System."

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space."

Romanov, A., K. Lomotin, and E. Kozlova. 2019. "Application of Natural Language Processing Algorithms to the Task of Automatic Classifcation of Russian Scientific Text." *Data Science Journal.*

Selivanov, D. 2018. "Topic Modelling." http://text2vec.org/topic_modeling.html#latent_dirichlet_allocation.

Slavazza, P. 2019. "What Is the Best Method for Automatic Text Classification?" https://towardsdatascience.com/https-medium-com-piercarlo-slavazza-what-is-the-best-method-for-automatic-text-

Uysal, A.K. 2013. *The Impact of Preprocessing on Text Classification.* Eskisehir, Turkiye: Department of Computer Engineering, Anadolu University.

Wilson, A.T. 2010. *Term Weighting Schemes for Latent Dirichlet Allocation.* California: Association for Computational Linguistics.