The University Of Sheffield.

Topic modelling and multiclass text classification methods for the online posts about people with diabetes

A study submitted in partial fulfilment
of the requirements for the degree of
MSc Data Science

at

THE UNIVERSITY OF SHEFFIELD

by

Nurbanu Aksoy
180223132

Word-length: *11.445*

August 2019

Abstract

Background: Online health communities have become more common and there has been a massive growth in the amount of information generated by such communities in a short period of time, and this information continues to accumulate day by day. Hence, these communities have become a potentially important resource not only for health studies, but also because they can provide a massive amount of data for text mining studies. Correspondingly, it is also crucial to understand the language of the members of these communities in order to process the data accurately and effectively.

Aims: The overall purpose of this study is to identify themes in online posts about diabetes patients and to determine the category of new posts by using supervised machine learning classification algorithms. The study also aims to demonstrate some of the alternative methods that can be employed to use the data accumulated via the Internet in academic studies.

Methods: The Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) methods were used to identify themes. The Support Vector Machine (SVM), Logistic Regression (LR), and multinomial Naïve Bayes (NB) supervised machine learning algorithms were applied to solve the multiclass classification problem.

Results: The LSA and LDA models were generally successful at extracting meaningful topics from the data set and they yielded beneficial information about the discussions about diabetes taking place in online forums. However, when using the LSA model, interpretation was easier and semantic integrity was higher. Multinomial NB, LR and SVM achieved good accuracy scores of 74%, 78%, and 80%, respectively. However, the recall, accuracy, and f-score results of the SVM were better than those produced by LR and multinomial NB.

Keywords: Machine Learning, Multiclass Text Classification, Topic Modelling, Latent Semantic Analysis, Latent Dirichlet Allocation, Multinomial Naïve Bayes, Support Vector Machine, Logistic Regression

Acknowledgement

I would first like to thank my supervisor Dr Laura Sbaffi for all her support and guidance.
I would also like to thank Arjun Panesar for providing the data set for this dissertation.

I would also like to give my sincerest thanks to my father Professor Dr Ömer Demir and my mother Associate Professor Zekiye Demir for supporting, encouraging and inspiring me throughout my study, as they have during my entire life.

In addition, I would like to thank my former colleagues Kadir Okumuş and Abdussamet Dumankaya for challenging and enriching my ideas.

Finally, yet importantly, I must express my profound gratitude to my beloved husband Fatih Aksoy for tolerating the long-distance relationship and supporting me at every stage of my life.

This dissertation is fully supported by the Ministry of Education, Turkey.

Table of Contents

Chapter 1: INTRODUCTION

1.1 Introduction

Technological change has caused many transformations in almost all aspects of the life of human beings throughout history. In the time before the World Wide Web, accessing health information was not always easy, because the various sources were scattered in different locations and permission was often required to access the information (Fiksdal et al., 2014). The fundamentals of seeking and using health information have been changed by the development of modern communication technologies. The spread of the Internet has become wider due to improvements in the related technology and this has led to a rapid increase in the utilisation of information obtained via the Internet. Web technologies and countless online platforms are now enabling people to access information easily and rapidly (Chen, Y.Y., Liang, & Tsai, 2018).

People have begun to prefer searching on the Internet for answers to their problems, including crucial issues such as finding out the symptoms, diagnosis, and treatment methods for various conditions and diseases, instead of getting an appointment with a doctor or consulting medical staff (Mead et al., 2003; Sillence et al., 2007). On the back of being a resource of general medical information, somehow the Internet has become an authority that individuals use to find information on specific diseases such as diabetes (Ravert et al., 2004). People with diabetes now access information over the Internet and also share their experiences and connect with others to find support and advice (Taylor et al., 2015). As a result of these developments, many diabetes-oriented online forums and communities have emerged to provide diabetes patients with platforms to manage their issues systematically (Hilliard et al., 2015; van der Eijk, 2013). Hilliard et al. (2015) also pointed out that people with diabetes feel blame, stigma and other negative feelings in respect of such subjects as consuming insulin and managing diabetes (Folias et al., 2014; Wolf & Liu, 2014). Therefore, patients and caregivers have begun to connect with each other to talk about topics related to living with diabetes in online societies across all branches of social media (Taylor et al., 2015). In 2014, websites devoted to diabetes such as TuDiabetes.org and EsTuDiabetes.org reached more than 20,000 members, and thousands of new users log into popular forums every month (Hernandez, 2014). Therefore, various previous studies (e.g., Bernardi & Wu, 2017; Tang et al., 2012) have argued that how diabetes patients use these online communities. Bernardi and Wu (2017), who used data from www.diabetes.co.uk in their

research, found that the impact of these forums on the life of people with diabetes is favourable, and that the information that is acquired from this platform makes people feel more empowered.

1.2 Research Background

Online health communities have become more common and there has been a massive growth in the amount of information generated by such communities in a short period of time, and this information continues to accumulate day by day. It is an undeniable fact that these online communities, or forums, contain valuable information about patients and their lives, and a lot of unexplored data is embedded in these posts. Hence, these forums have become a potentially important resource not only for health studies, but also because they can provide a massive amount of data for text mining studies (Chen et al., 2015; Christensen et al., 2017; Ibrahim et al., 2017; Li & Wu, 2010;). Correspondingly, it is also crucial to understand the language of the members of these communities in order to process the data accurately and effectively. In contrast to medical professionals, most people are inclined to use daily terms to explain their symptoms because they are not familiar with the medical jargon and terminology that is widely shared among professionals (Hasan et al., 2017). Therefore, the processing and analysis of everyday language is essential for obtaining information. In order to understand the language of users, it is necessary to convert the human language into a language that machines can understand, because computer technologies (both software and hardware) are needed to process, analyse, manipulate, and store the large amount of data available online.

1.3 Research Aims and Objectives

The overall purpose of this study is to identify themes in the online posts about diabetes patients and to determine the category of new posts by using supervised machine learning classification algorithms. This is accomplished by applying text mining techniques and natural language programming (NLP) to the data. This study also aims to demonstrate some of the alternative methods that can be applied to use the data accumulated on the Internet,in academic studies. Furthermore, the results will contribute to research on the comparative advantages and disadvantages of some data mining models and methods.

The main objectives of this research are to (1) identfy the underlying trends in the discussions, (2) apply text mining approaches to unstructured daily language and (3) identify any complications and factors that may affect the results of the text mining process.

1.4 Conclusion

In order to achieve the above aims and objectives, this research specifically focuses on one of the discussion forums in the online diabetes community www.diabetes.co.uk. Data from this community is subjected to topic modelling and text classification algorithms. The Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) methods are used to identify themes, whereas the Support Vector Machine (SVM), Logistic Regression (LR) and multinomial naïve Bayes (NB) supervised machine learning algorithms are applied to solve the multiclass classification problem. Detailed information about the algorithms is provided in Chapter 3, 'Methodology'. An evaluation of the results and a between models comparison is presented Chapter 4, 'Results'. The results are discussed and compared with those in the literature in Chapter 5, 'Discussion'. Finally, limitations, challenges and suggestions are presented in Chapter 6, 'Conclusion and Future Work'.

Chapter 2: LITERATURE REVIEW

2.1 Literature Search Strategy

In order to capture the extensive research that has been conducted on and around the topic of interest to this study, several databases were searched, including Google Scholar, Starplus – Library Catalogue, and IEEE Xplore. Moreover, many academic social networking sites, such as Academia and ResearchGate, were used to find similar studies. In addition, the reference lists of relevant articles were also examined in order to expand the researcher's knowledge of the field of study. The literature was searched in two different languages and articles published in the English and Turkish languages were included in the review. Generally, a combination of keywords was used together, including 'online health communities', 'data mining', 'topic modelling', 'text classification', 'comparison of machine learning algorithms' and so on.

2.2 Data Mining and Online Health Communities

Studies in the various health fields receive a lot of attention at any time because of the impact that they have on society. Due to the digitalisation of data and the escalation in Internet usage, several studies have concentrated on examining online health information, not least because patients have started to focus more than ever before on the information they can obtain from online communities (Nath et al., 2016; Wang et al., 2017). These studies, which have been conducted since the 1990s (Brennan & Ripich, 1994; Cummings et al., 2002; Finn, 1999) have pointed out that, by means of online health forums, people share views and advice related to their own diseases, and also support others in dealing with their problems. As a consequence of this, virtual social circles have emerged, the members of which continuously exchange their own experiences and information (Willis & Royne, 2017).

Moreover, Kummervold et al. (2002) investigated the use of four main online health forums in Norway, and according to the responses to their survey, most participants concurred that these platforms provide factual information and social support. Furthermore, the participants mentioned that disclosing their private and sensitive problems over the Internet is quite a lot easier than talking about them face-to-face with another person, a finding that is supported by Berger et al. (2005). This implies that the data from online forums is more valuable than that held in administrative health records in this respect.

In the wake of this phenomenon, the number of active members has grown rapidly and this has ensured that people continue to rely on the information provided by online forums and communities (Nambisan, 2011; Neal et al. 2007). White and Dorman (2001) indicated that, apart from helping patients to receive information more efficiently and faster, these communities have contributed to the health literature. The authors also supported previous studies and specified that a considerable number of people are grateful for the presence and informational support provided by such communities, and that the advantages of these communities have been acknowledged by society. In addition, they stated that, as the interest of people in these platforms has been escalating it is inevitable that the amount of online data will also increase. Moreover, the data that has accumulated over time has started to contain meaningful information for the health field (Wang et al., 2017). Therefore, studies have been conducted to analyse, examine and extract the information that can be found in online health forums (Gao et al., 2017).

Many studies in this area have particularly tackled the reliability of online health information (Lederman et al., 2014; Wang & Liu, 2007). However, one of the key challenges that such studies have encountered is finding a way to examine the inputs of users that is both efficient and accurate. It is here that data mining comes into its own, because traditional, mostly manual, systems do not have sufficient infrastructure to evaluate and analyse a large amount of data. Therefore, it was inevitable that data mining technologies would be applied to this kind of analytical problem, as the methods and techniques provided by data mining can carry out operations, which would usually take a long time with traditional approaches, in just minutes or seconds (Koh & Tan, 2011). Furthermore, data mining can also draw on different technologies to find and extract meta-data, and the novel information thus discovered can be used by different branches of healthcare organisations (Kaur & Wasan, 2006). Accordingly, many studies (e.g., Corley et al., 2010; Hasan et al., 2017; Warrer et al., 2012) have been conducted to examine health-related posts by using data mining techniques.

In order to achieve the aims of the current study, a system that extracts information from the whole text is needed to examine and analyse the posts. Here, the key point that needs to be borne in mind is not to confuse ‘information extraction’ and ‘information retrieval’ (Corcoglioniti et al., 2016). Information retrieval has many definitions. However, in the field of computer science, the purpose of information retrieval is to obtain the needed information from a source that usually does not have a regular structure (Manning et al., 2010). Hence the

focus of information retrieval is on facilitating and simplifying access to information rather than on analysing or mining data for information (Allahyari et al., 2017). On the other hand, information extraction is one of the text mining approaches that is used to minimise human intervention, which is achieved by generating software that processes large amounts of unstructured or semi-structured data automatically to extract information (Allahyari et al., 2017; Sarawagi, 2008). Although the extracted information is usually in the form of documents or text, the sources of data can vary. For instance, databases, documents on websites, sharing on social media channels and posts on online communities may be the subject of text mining operations (Allahyari et al., 2017).

In essence, text mining is the process of obtaining non-discovered, potentially useful, structural and regular data from unstructured and irregular electronic text stacks. The hidden patterns, relationships, hypotheses, and trends in textual sources are determined by the analysis of the gathered data (Allahyari et al., 2017). Although text mining is considered to fall under the data mining 'umbrella', it is different from ordinary data mining. In text mining, patterns are usually extracted from natural language texts rather than event-based databases. The term 'natural language' can be defined as how humans communicate in their daily life (Manning et al., 2017). However, human language changes and evolves over time, unlike the machine languages (programming languages) used in devices. Also, grammatical rules vary from language to language and even within a language over time. Therefore, it is difficult to examine data containing human language (Bird et al., 2009).

In its widest definition, Natural Language Processing (NLP) is the process that a computer-employs to detect the text and sound waves in natural languages, such as Turkish and English, to be analysed and transferred to the digital environment by software (Chowdhury, 2003). In other words, NLP aims to understand, resolve and manipulate the regular structure of natural languages. Thus, text mining algorithms are usually run alongside with NLP because NLP plays an important role in the processing of raw data (Allahyari et al., 2017). The use of NLP is especially important for the analysis of the data collected from online health forums, as members tend to express problems in their own words and mostly use uncommon abbreviations. Therefore, before topic modelling and text classification algorithms can be applied to a data set derived from an online discussion forum, it is a prerequisite that the daily language used in the forum is examined and processed by using a NLP application.

2.3 Topic Modelling

Topic modelling is an unsupervised text mining approach that uses statistical methods to analyse unstructured large clusters of text and automatically extract the major themes and topics (Williams & Betak, 2018). Topic models can be referred to as key concepts extracted from document collections and topics can be inferred by employing basic clustering methods and machine learning algorithms (Yi & Allan, 2009). The extracted topics are represented as a group of terms, or in other words, as a bunch of words. Topic extraction is a crucial step in summarising the majority of text documents; moreover, it provides an infrastructure to reveal hidden patterns/semantics in the data (Li et al., 2016; Shi et al., 2019).

Due to the growing population of online communities and social networks, the extraction of appropriate key points from original text and summarisation has become a field of interest for many researchers. Much attention has been accorded to the usage of topic modelling in studies encompassing many different research areas (Liu et al., 2016), because topic models can provide effective representation of all types of content (Yi & Allan, 2009). Not only long texts require the extraction of pivotal concepts; even short texts need to be semantically summarised in some cases (Li et al., 2016), particularly online data collected from social networks and communities, as this type of data is usually in the form of brief texts or dialogues. However, it is challenging the apply a text summarisation or clustering approach at the short-text level (Abdalgader, 2017), because the algorithms used in these approaches do not work perfectly on this type of data. In those cases, the topic modelling approach is much more efficient in terms of getting reliable results, and provides suitable methods to analyse a large amount of unclassified data (Alghamdi & Alfalqi, 2015).

There are numerous algorithms that can be used for topic modelling, and among the effective models that have been proposed, both LDA and LSA have been found to provide appropriate and accurate results (Cvitanic et al., 2016; Phan et al., 2008;). David Blei, Andrew Ng and Michael I. Jordan introduced the original LDA model in 2003 (Blei et al., 2003). The LDA model is a Bayesian statistical/computational model that uses an unsupervised algorithm efficiently to identify latent semantics in documents (Dyer et al., 2017; Guo et al., 2017). The LDA model is a productive model that enables the observed groups to be explained by unobserved groups in order to describe why some parts of the data are similar.

Therefore, studies, which include learning algorithms, are used LDA model in many various domains such as clustering, classification, etc. (Momtazi, 2018).

Jiang et al. (2017) used the LDA algorithm to understand how the approaches to the analysis of climate change vary among publications. In simple terms, they utilised LDA to generate topics that consisted of keywords that represented their targets. On the other hand, Christou (2016) applied a LDA model to capture the features of paragraph vectors in order to demonstrate the importance of using feature extraction in recommendation systems, More recently, Ekinci and Omurca (2017) used the LDA model to identify and extract aspects and features from Turkish-language hotel reviews. Meanwhile, Momtazi (2018) proposed a new method based on LDA for classifying questions in community-based question and answer forums. As it can be seen, LDA has been widely used in many research studies, and this seems to be because it has been found to be successful in matching the terms of a given subject (Blei et al., 2003). Moreover, as the LDA model is not a supervised model, it does not require prior information about topics and can therefore provide well-defined outputs for unseen data (Blei et al., 2003; Chen, 2017). On the other hand, Wang and Bei (2013) and Chen (2017) have indicated that a drawback of this model is that although it can be applied effectively to large-scale data, it does not work efficiently on small amounts of data.

Latent Semantic Analysis (Deerwester et al., 1990) is also a computational technique that creates a semantic representation of words in a corpus. It is based on the idea that similar words occur in similar contexts (Kwantes et al., 2016). In LSA, the similarity metric is a semantic resemblance or interrelation (Süzek, 2017). The LSA algorithm can be implemented in a variety of approaches. For instance, Badry et al. (2013) applied LSA in text summarisation, and they also compared the performance of all the existing LSA algorithms. More recently, Vrana et al. (2018) applied LSA methods to analyse the correlation between doctor–patient expressions and to identify similarities between them by examining their interactions. By the same token, in an earlier study, Babcock et al. (2014) examined partners' conversations and used LSA to evaluate the similarity between their statements. Meanwhile, Süzek (2017) used an LSA algorithm as a key-extraction method to define the semantically significant subject in documents.

2.4 Text Classification

Due to the increasing amount of online text data, access to information has become easier; however, it has brought with it a new problem of finding useful information from among a tremendous volume of data. To find the desired information, requires that the data is collated under certain headings. This requirement necessitates the application of text classification. Text classification is a process of assigning a text to one of a set of predefined classes based on its content (Allahyari et al., 2017). The aim is to predict the class of a document based on training documents. Before the rapid diffusion of digitalisation, text classification was carried out manually by humans, and this process was extremely time-consuming.

The introduction of computers into the text classification process has resulted in this process becoming a significant research topic for a variety of fields such as machine learning, data mining, database management, and information retrieval. In addition, text classification is now being used in various applications (Aggarwal & Zhai, 2012). For instance, Luss and d'Aspremont (2015) applied text classification algorithms to forecast intraday price movements of financial assets by using news articles. On the other hand, Dang and Lawrence (2014) utilised text classification methods to develop a sequence-based allergen prediction tool. Furthermore, Aphinyanaphongs et al. (2016) used text classification techniques to automatically identify Twitter posts about electronic cigarettes,.

In common, text classification models apply machine learning algorithms to predict the class of a new input. In the algorithm selection phase, the number of tags in the target class should be taken into account. If there are more than two labels in the prediction, text classification models should run an algorithm that is suitable for the problem domain. Most of the popular machine learning models make use of NB, LR and SVM because these algorithms can accommodate many approaches (Allahyari et al., 2017; An, Sun & Wang, 2017; Brindha et al., 2016; Mohammad et al., 2016).

The NB classification algorithm is used to determine the class of data presented to the system by a series of calculations according to probability principles based on Bayes probability theorem with independence assumptions (Raschka, 2014). Some exemplary uses of a NB classifier include optimising treatment decisions (Kazmierska & Malicki, 2008), sentiment

analysis of movie and hotel reviews (Dey et al., 2016), classifying music emotions based on lyrics (An et al., 2017) and diagnosing Alzheimer's disease (Shree & Sheshadri, 2018).

On the other hand, LR is often used in quantitative research to analyse the association between dependent and independent variables through the prediction of probabilities (Sainani, 2014; Sedgwick; 2013). In NLP, LR is one of the successful machine learning algorithms used in text classification (Martin & Jurafsky, 2018). Logistic regression categorises an observed state into binary or multiclass. When the target variables are more than two, a multinomial LR algorithm is employed to find the probability of each class (Martin & Jurafsky, 2018; McCarthy et al., 2019). The application of LR in text mining and NLP is widespread. For instance, Mukku et al. (2016) used a LR algorithm for sentiment classification of Telugu text, while Pranckevičius and Marcinkevičius (2016) utilised a LR approach for multiclass (also known as multilabel) text classification. Moreover, Alsmadi and Hoon (2018) applied a LR classification algorithm to short-length data collected from Twitter corpuses.

Lastly, a SVM is a non-parametric method based on computational (statistical) learning theory and structural risk minimisation (Cortes & Vapnik, 1995). It can also be defined as a supervised machine learning method for classification, regression, and the determination of outliers in data sets where the patterns between variables are not known (Guenther & Schonlau, 2016; Meyer & Wien, 2015). The method was initially designed for the classification of binary linear data; later, it was improved for the classification of multiclass and non-linear data (Üstüner, 2013). A SVM has been used in different kinds of approach such as the pattern recognition of human actions (Laptev & Caputo, 2004), classification of images (Üstüner, 2013), graph classification (Rousseau et al., 2015), face recognition (Kremic & Subasi, 2016), regression problems (Guenther & Schonlau, 2016) and the prediction of document or text categories (Mohammad et al., 2016; Vadivel et al., 2018).

Chapter 3: METHODOLOGY

3.1 Data Set

The primary data used for this study was provided by the owner of www.diabetes.co.uk, which is a global community of people affected by diabetes. On this website, almost every subject related to diabetes patients is handled by placing it into one or more categories. There is also a forum section that allows users to link to each other interactively. The discussion topics are classified in the forum and each title contains related subtopics as well. For instance, the posts under the heading 'Living with Diabetes' are covered by the eight major forums on the website. This main heading or headline contains three subtitles. The first, 'Jobs and Employment' is used to discuss problems related to diabetes and work, issues with employers and discussions about employment. The second subtitle, 'Benefits', is used for any queries or advice about benefits, tax credits, value-added-tax exemption, and diabetes-related legislation. The third and final subtitle, 'Driving and DVLA, is used to discuss driving and the Driver and Vehicle Licensing Agency (DVLA) guidelines regarding diabetes. All the threads from these three subtitles were provided by the owner of the site in the form of csv file and there are 9008 posts in all.

3.2 Programming Language

The Python programming language was used for the implementation of the proposed method. Python, which was developed by Guido van Rossum in 1991, is a versatile, interpretive and object-oriented programming language, which emerged with the main philosophy of code readability. Python has become increasingly popular, because its learning process is easier than that of other programming languages, it does not require a separate compiler and it provides clean code syntax (Pedregosa et al., 2011; van Rossum & de Boer, 1991). In addition, Python provides useful libraries for the implementation of NLP and text mining algorithms, and therefore, it is commonly used and widely preferred (Nagpal & Gabrani, 2019).

3.3 Data Set Preparation

First, the **read_csv** method of the **pandas** library was used to import the data set. Pandas is an open-source library that facilitates data analysis and data preprocessing, and it uses Python as a programming language. Pandas enables the fast and efficient use of data frames.

It also allows to read, examine, write and save csv files easily. In addition, pandas is a library that is optimised for speed. In cases where the data set is not very large, pandas is one of the most preferred libraries.

After reading the data from the csv file and loading it into the data frame, the data set had 9008 rows and 17 columns that contained detailed information about the posts such as date, number of likes, last edit time etc. The columns that were not used in the text mining process in this study were dropped from the data frame, leaving only the columns "node_id" and "message" in the data set. Although node_id was not used for topic modelling, that column was required for labelling each message by title when applying the text classification algorithms. According to the given node_ids, the Title column was added to the classification data set. Tables 3.1 and 3.2 below provide examples of the contents of the resultant data sets.

Table 3.1

Topic Modelling Data Set

	message
0	Hi\r\n My son is 17 now type one and is finding it hard to find work.he wanted t…
1	Hi James\r\n\r\nI suggest you get in touch with [url=http://www.dialuk.info/]htt…
2	Hi, \r\nI work as a paramedic currently and have just been diagnosed as type 1 d…

Table 3.2

Text Classification Data Set

	message	**node_id**	**Title**
0	Hi\r\n My son is 17 now type one and is findin...	8	Jobs and Employment
1	Hi James\r\n\r\nI suggest you get in touch wit...	8	Jobs and Employment
2	Hi, \r\nI work as a paramedic currently and ha...	8	Jobs and Employment

3.4 Data Preprocessing

One of the most important steps in tackling text mining problems is preprocessing; if the data is not well prepared, the results will be biased and/or inaccurate. Furthermore, the preprocessing step differs according to the data set.

3.4.1 Data Cleaning

Data should be as clean and as simple as possible. In performing the data cleaning process for the data used in this study it was taken into account that the inputs were written by people in natural language. Therefore, it was essential that the data was properly cleaned for predictable errors such as irregular sentence structures, misspellings, and punctuation mistakes.

First, because all the data in the message column was in the form of text, regular expression language was used to edit and manipulate the posts to prepare them for analysis. Regular expression language is a computer language used to edit texts or to obtain subtexts from within the text that conform to certain rules. It is applied to a string or collection of characters (Wang et al., 2019). As an outcome, substrings are formed and/or modified new texts containing some of the original text are obtained. In other words, the use of regular expression allows the identification of a particular string by the shortest path and the searching of specific information within the text (Duru et al., 2018). In Python, the **re** package was used to implement the regular expression process. Initially, one of the advanced regular expression functions, i.e., the substitute function, was used to perform the various operations on the text.

All words must be separated by spaces in order to be able to extract each word precisely from the sentences. Therefore, spaces were added after all the full stop (.) characters. In addition, because the data was obtained from an online discussion platform, the posts also contained replies, for example, **([b] Re: ... [/b])**, quotes **([quote="xxx"] ... [/quote])**, user names (**[USER=xxx] … [/USER])** and different characters. So, before the removal of all the punctuation, the strings between two statements were deleted.

Then, all the characters, except for those that were alphabetical, such as invalid characters, punctuation, digits, new lines (/r/n etc.), were removed from the strings by using the **replace()** function. The next step involved the removal of short words.

That is to say, words that had a length of two or fewer characters were deleted from the strings. This step is performed to try to reduce the presence of words that do not reveal any information, such as “ok”, “hi”, “ve”, “s”. After that, all the stop words were removed. This further text modification step to remove stop words such as “the”, “at”, “and” must be undertaken in order to maintain the accuracy of the models when they are applied to the text. The English-language stop words defined in the Natural Language Toolkit corpus were used for this process.

The last step involved a process called stemming which normalises the terms in the text. Basically, this process is used to convert words into simple forms or roots by removing the plural suffixes from nouns or the affixes from verbs (Colton, 2016). The most common rule-based stemming algorithm for the English language was developed by Porter (1980) and is still extensively used. However, when the Porter and other stemming algorithms were implemented on the data set, the returned results were not the same as those in the dictionary. For instance, “licence” was converted to “licenc” and “people” had become “peopl”. As such, these stemmed words would affect the accuracy of the models and consistency of the data, therefore, the stemming process was performed manually by using extended regular expressions (regex) where necessary.

3.4.2 Exploratory Analysis

An exploratory phase was conducted to gain a comprehensive understanding of the data set. This phase reveals whether additional preprocessing is required before the model algorithms are applied to the text. In addition, its use can lead to some insights while at the same time ensuring that the data set is sufficiently prepared. During this phase and thereafter, the **numpy** library was used frequently to help to define the functions. numpy is one of the beneficial libraries that allows scientific calculations to be performed quickly and it can also handle multidimensional arrays,. Moreover, numpy has been used with the pandas library to carry out text analysis. In order to better understand the outcome of the exploratory analysis, the results were visualised by **matplotlib**, which is one of Python's most basic libraries. It provides many visualisation options such as line plots, scatter plots, bar plots, subplots, and histograms.

The first step in the exploratory analysis was to check the distribution of topics. The result is illustrated in Figure 3.1.

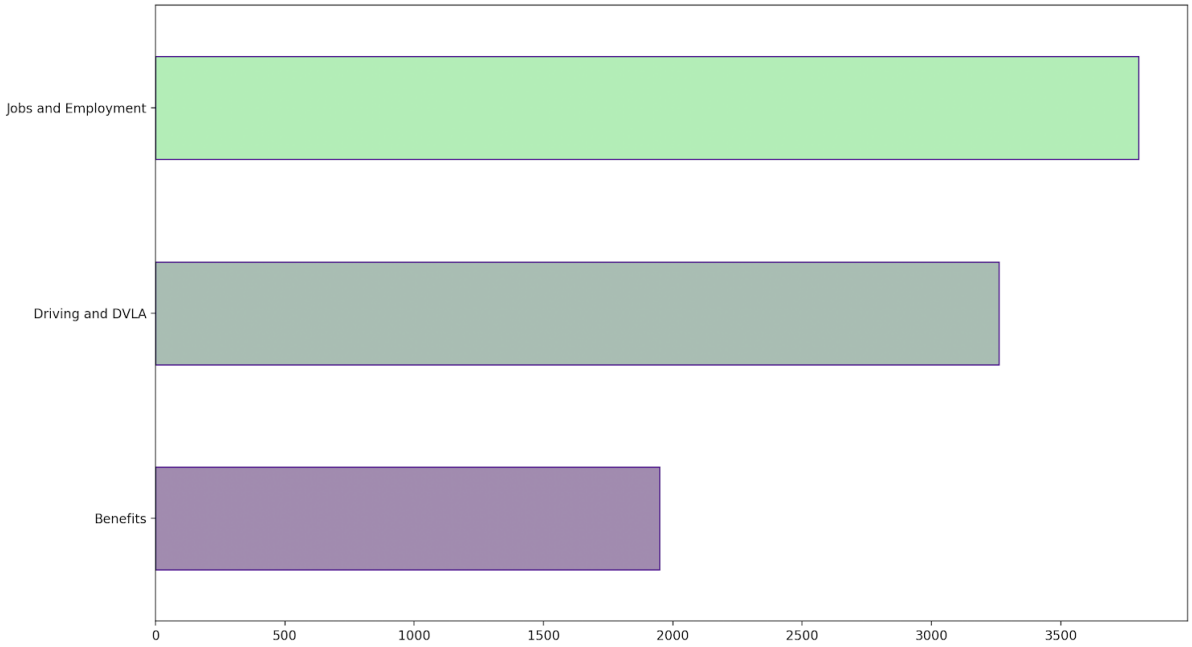

Figure 3.1

Topic distribution

As can be seen from the topic distribution bar chart in Figure 3.1, the number of posts is imbalanced. Users' posts are more biased towards the Jobs and Employment and Driving and DVLA. In a data set where the distribution is imbalanced, the results of the models are usually biased towards the majority group. If the number of the minority group is overmuch in the distribution, they are treated as outliers and ignored, or aggregated into one group usually called *Others*. However, in this study, no group is considered an outsider, so there is no need to make a change to the data set regarding balance state.

The next step in the exploratory analysis was to identify the most frequently repeated/used words in order to determine how well the data set had been cleaned. The results are illustrated in Figures 3.2 and 3.3 and discussed below.

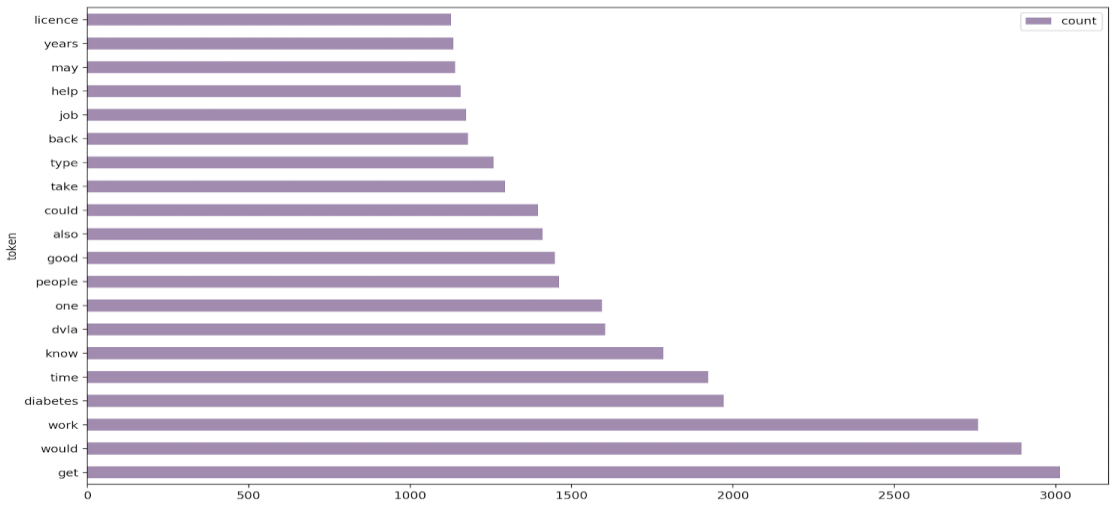

Figure 3.2

20 most frequent words in the data set

As shown in Figure 3.2, the words “get”, “would”, “know”, “also”, and “may” are the most commonly used in the sentences. However, some of these words do not carry any meaning by themselves (similar to stop words) and some are used for the ordering of the commands. On the other hand, because this data set is taken from a discussion forum, the repetition of some words is due to the nature of the language used in this context. For instance, the replies to questions start with “I think” in most cases. Therefore, in order to extract relevant insights, these kinds of words were deleted from the data set.

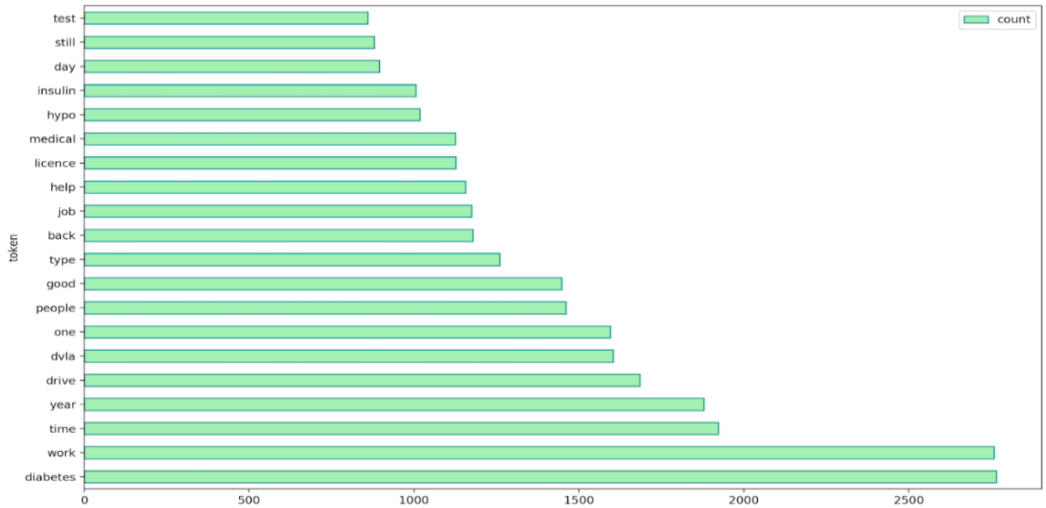

Figure 3.3

20 most frequent words in the data set after cleaning

Figure 3.3 shows that the most frequently used words are “diabetes” and “work”. The high usage of the word “work” can be explained by the fact that the Jobs and Employment topic contains more data than the other two topics. Moreover, the word “work” has many non-occupational uses. In addition, “type”, “one”, “hypo” and “insulin” are among the most commonly used words in the data set. Hypo (stands for hypoglycaemia), which is defined as a blood glucose level of 50 mg/dl or less and it is often regarded as indicating the need for diabetes treatment (International Hypoglycaemia Study Group, 2015). This condition generally occurs in people taking insulin or oral antidiabetic medication (Khunti et al., 2015). Also, the usage of insulin is more common in type 1 diabetes patients. Thus, the insights can be extracted even with the interpretation of the figure. Similarly, some groups of words, presented in the bar chart, give hints about the topics people talk about. For example, “drive”, “dvla”, “licence” and “test” show that people are talking about issues regarding driving. On the other hand, “work”, “job”, “time” and “year” may also be clustered under a topic related to working life.

Last but not least, the exploratory analysis involved the creation of a word cloud. A word cloud is formed on the basis of word frequencies and has a slightly more understandable, elegant presentation and thus aids visualisation of the data set. In Python, Andreas Mueller’s **wordcloud** library was used to produce the word cloud, which is depicted n Figure 3.4.

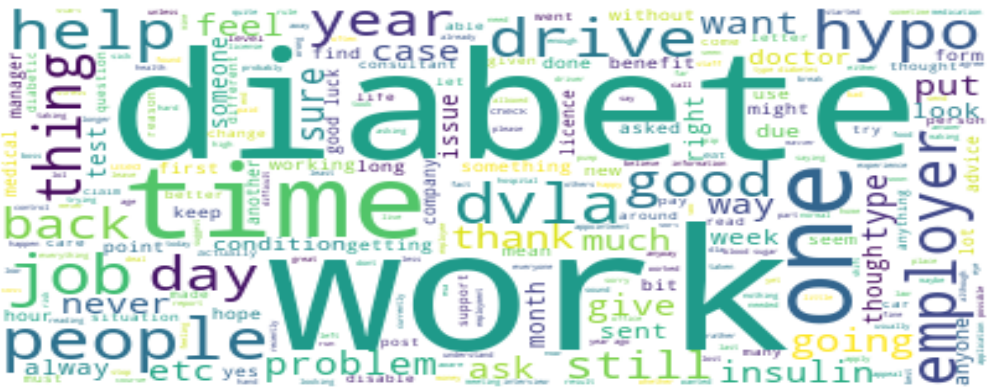

Figure 3.4

Word cloud

When visualisations and the data set are explored together, general information can be obtained about what the data set is about and which topics are the most discussed. Consequently, the data set is more interpretable and this indicates that the data is fairly clean and adequately prepared for the implementation of the models.

3.5 Topic Modelling Algorithms

In the current study, LSA and LDA were used for topic modelling. They are both prominent methods that can be used in the clustering and similarity analysis of texts according to their topics (Yıldıztepe & Uzun, 2018). Due to the limitation of space, the mathematical and statistical details/backgrounds of the methods are not discussed here. Instead, only the logical working principles are described in order to indicate why these methods were chosen.

Latent Semantic Analysis is a method used to analyse the patterns in a text document and is based on extracting concepts from the similarities of the terms in each sentence in the document. It creates a low-dimension vector representation of the document and helps to find similarities by calculating the distance between the vectors (Anandarajan, Hill & Nolan, 2019). It also allows the performing of operations such as determining whether a text group belongs to a known group, classification and topic modelling (Alghamdi & Alfalqi, 2015).

Latent Dirichlet Allocation makes it possible to define the hidden, primary topic of the document. The fundamental idea behind LDA is that documents are represented as mixtures of hidden topics identified by words (Jelodar et al., 2019). Accordingly, Dirichlet distribution is used to determine the likelihood that texts and words belong to the topic (Levandoski & Lobo, 2018). Three assumptions are taken into account when applying the LDA method. First, if the topic of a text is known, similar words can be found in the other texts that belong to the same topic. Second, if there is no information about the topics of the texts, they are equally likely to belong to the topics. Lastly, if there is no information about which words are more likely to be found in the topics, the words can be found with equal probability in the topics (Yıldıztepe & Uzun, 2018).

Basically, the two models are based on the idea that each document is a combination of topics and that each topic consists of a group of words (Alghamdi & Alfalqi, 2015). Therefore, the outputs of both models are collections of words divided into different groups.

In order to implement both methods, a feature extraction technique needs to be performed because neither method works directly on the data. First, the data has to be converted into a vector representation. Basically, this representation is used by the methods to identify the topics. *TfidfVectorizer* and *CounterVectorizer* in the sklearn.feature_extraction.text module are common classes used to produce term frequencies.

In LSA and LDA methods, these two approaches were tested. *CounterVectorizer* was used for the final version of models, because it was concluded that the results of *CounterVectorizer* were more interpretable and accurate.

For **CounterVectorizer ( )**, two parameters were adjusted: min_df = 2 and max_df = 0.8. **min_df** was used for the exclusion of rare words. In contrast, **max_df** (known as corpus-specific stop words), was used to do not include frequently used words. Then, the vector output was used to create the document term matrix. The document term matrix was used by both models.

In order to build the LDA model, the **LatentDirichletAllocation** class was imported from the *sklearn.decomposition* library. The number of the topics parameter was tuned as three, as the data was originally received from three different titles. The acquired model was fitted with a document term matrix. Subsequently, auxiliary functions were implemented to print topics and the top 10 most-used words within each topic.

In the case of the LSA model, the **TruncatedSVD** class from the sklearn.decomposition library was used to reduce dimensionality. TruncatedSVD () provides a reduction of the linear dimension through truncated singular value decomposition (SVD). The SVD model was also fitted with a document term matrix and some additional key functions with LDA model were used to get the top 10 most-used words within each topic.

3.6 Text Classification Algorithms

There are multiple different uses of text classification, including spam filtering, sentiment analysis, emergency response system, commercial world, etc. Classification algorithms differ for each use case according to the data set. As some algorithms have been developed specifically to predict binary classes, they cannot handle multiclass problems properly. Therefore, a different approach needs to be adopted for these cases. Multiclass text classification is a supervised machine learning technique that is used when the target variables are more than two classes. The technique assumes that each variable in the data set can only belong to one label.

In this study, each topic name was introduced to the system as a label. This meant that a three-label text classification system was required. Accordingly, a multinomial NB classifier, multinomial LR, and SVM were used to build models. The underlying logic of all three

models was based on two basic processes. First, the system has to be trained by the data set, and second, when a document needs to be classified, the trained model decides on the class by making a comparison with the training documents.

Therefore, first, the data was divided into two parts – a test set and a training set – in order to predict and classify the incoming data. Hence 67% of the data set was randomly assigned as the training set and the remaining 33% was used as the test set. This process was conducted by using the **train_test_split** module imported from *sklearn.model_selection*. The next process was feature engineering, which is one of the essential steps for systems based on machine learning. As indicated in the Section 3.5, 'Topic Modelling Algorithms', when the primary data is textual, there are several approaches that can be adopted to represent the data in numerical form. In order to meet system requirements, suitable feature extraction modules were used: *CountVectorizer ()* provided a term frequency matrix and *TfidfTransformer ()* was used to normalise this matrix. After this process was completed, the axes of space consisted of words specifying the classes and all the words had been converted into numerical data. Next, each of the algorithms was fed with the training data, which included features and labels, to build three, three-label classification models.

Naïve Bayes is one of the widely used classifiers to solve NLP-based multiclass classification problems. Naïve Bayes uses probability theory and Bayes theorem to predict the class of data. It calculates the probability of all classes for each sample and returns the statistically highest probability as a prediction. In contrast to other models, each word is assumed to be independent of the sentence, and the probabilities are calculated for each word severally. This is one of the major advantages of this classification algorithm over others. Moreover, it is easy to construct a system with a NB classifier and the operations are conducted swiftly because it has a simple structure and it is easy to adapt it to other components in the system.

In order to build and train the multinomial NB model, first, a pipeline was created by using the Pipeline class from sklearn.pipeline. Then, the MultinomialNB class from sklearn.naive_bayes was employed and used with CountVectorizer() and TfidfTransformer() to establish a pipeline for the training set. The NB algorithm was trained using this pipeline by means of the fit method.

The second classification model was developed using LR which one of the supervised classification algorithms that can be used to divide data into subcategories. A LR classifier uses the calculated logits to predict the target class. Moreover, LR classification algorithms perform differently according to the number of target classes. If the target variables are binary, such as female/male, positive/negative, etc., a binary LR classifier is employed, whereas when the target classes are multiple, a multinomial LR classifier is used for classification.

The implementation of multinomial LR in Python was conducted by using the **LogisticRegression** class from the *sklearn.linear_model.* Then, a pipeline was created by using CountVectorizer() and TfidfTransformer(). In order to notify the system that LR will be applied to classify for multiple target variables, some parameters have to be specified. Therefore, the *multi_class* parameter was altered from the default to '*ovr'*, while the *class_weight* that was selected was '*balanced'*, and the *solver* was changed from 'liblinear' to '*lbfgs'*.

The last model that was constructed for the classification task was a SVM model. The SVM is a method that uses the optimal algorithm to determine the boundary between classes in multidimensional space. It uses the decision boundary (also known as the hyperplane) to separate different classes from each other. The decision function uses a number of training points to find the maximum marginal hyperplane which can ideally classify the data set. In a SVM model, many different kernel functions can be used for the decision function, therefore memory is used efficiently.

Support vector machines can be divided into two groups according to whether the data set is able to separate linearly or not. Considering the data set of this study, a linear SVM was utilised to build the multinomial SVM classification model. In order to create a pipeline, LinearSVM was imported from *sklearn.svm* and combined with CountVectorizer() and TfidfTransformer(), similar to the other two models. The parameters were tuned by using options *multi_class='ovr'*, *loss='hinge'*, *class_weight='balanced'*.

3.7 Ethical Implications

The ethical consideration that needed to be taken into account in this study was that the data that was used belongs to humans. The health data of humans is referred to as sensitive data in the literature. However, all of the data used in this study has already been published on a public platform. It is therefore available to whoever wants to access it and hence there is no detected risk of breach of privacy. In addition, the data was not extracted directly from the website without permission. Nevertheless, to minimise ethical concerns, all the data underwent anonymisation, which was done by deleting the users' demographic characteristics.

Chapter 4: EVALUATION

4.1 Performance Measurement

Performance can be described as the state where models perform in accordance with predetermined standards in order to achieve the desired results. In this regard, performance evaluation is a process in which performance is monitored, evaluated and recorded. The performance of a model may depend on the learning algorithm used, as well as the distribution of the data set or size of the training and test sets. Therefore, performance evaluation methods demonstrate the accuracy of the models when run on the existing data set. Furthermore, there are different types of metrics that can be employed to measure performance according to the algorithm structure and model used.

The performance measurement of topic modelling algorithms is an ongoing challenge in text mining research. Although some techniques have been suggested in the relevant literature, a common and effective method for evaluating topics precisely, rather than appraising them through human intervention, has as yet not been found. Hence, some measurements based on manual evaluation and interpretation were considered in order to determine whether the LDA and LSA models were accurate or not. Initially, the parameter for the number of topics was fed with different variables into the models in order to discover whether the models would be able to divide the data set into reasonable topics when the number of topics was unknown. This process was also conducted to examine whether the proper number of topics was selected. Moreover, the density of the words in every topic was investigated, or to put it another way, the words under each topic were examined to ascertain whether they were consistently scattered. In addition, the content of each topic was evaluated to find out whether semantic integrity was present or not. If the results of these investigations are easy to interpret and meaningful, the model can be considered successful. Lastly, the topics provided by the models and those from the www.diabetes.co.uk website were examined and a comparison was made.

On the other hand, there are several accepted performance metrics, such as the confusion matrix, accuracy, precision, recall, and f-score, that can be used to evaluate the performance of supervised machine learning algorithms. The **confusion matrix** is the method that is most commonly used in machine learning to evaluate the performance of models derived from

predetermined data sets. This matrix shows the extent to which the test set accurately classifies the positive and negative samples of the model.

The other most popular and basic metric that is used to measure model performance is **accuracy**. The ratio of the number of correctly classified samples to the total number of samples gives the accuracy of the model. The error rate is the complement of this value to 1. **Precision**, on the other hand, shows how accurately the classes are predicted. It is calculated as the ratio of correctly classified positive samples (TP) to the total number of positive samples (TP + FP), where TP denotes a true positive and FP denotes false positive result. **Recall** indicates how well positive classes are predicted and it is defined as the ratio of the number of correctly classified positive samples (TP) to the total number of positive samples (TP + FN), where FN denotes a false negative result. Furthermore, evaluating the precision and recall metrics together gives more reliable results, and this evaluation is represented by the f-measure is defined. The **f-score** (also known as the f-measure/f1-score) is the harmonic mean of precision and recall.

In addition to using the above metrics, the time elapsed during compilation and runtime was calculated for each model to demonstrate the operating speed of the models. The implementation of the metrics was conducted by utilising the *sklearn.metrics* module and using accuracy_score, confusion_matrix, precision_score, recall_score, classification_report and f1_score. In addition, the confusion_matrix was visualised by using *matplotlib* and *seaborn* modules to gain a better understanding of the results.

After the above analyses had been performed, the accuracy scores were evaluated by cross-validation to test the effectiveness of the models. The cross-validation technique is used to minimise the errors that occur during the distribution phase of the training and testing process. This technique is used to validate the stability of the models, and it does this by randomly resampling the training and test data set for a predetermined number of times. In this study, resampling was done 10 times, which is in line with the literature (Tripathy, Agrawal, & Rath, 2015). Then, the values for the minimum, maximum, mean and standard deviation that were obtained from cross-validation were examined and the performance of the model was interpreted. The cross-validation technique was implemented by using **cross_val_score** from *sklearn.model_selection*, while the *mean(), min(), max() and std()* functions of the *numpy* library were used for the calculation.

4.2 Results

4.2.1 Latent Semantic Analysis and Latent Dirichlet Allocation

The topics generated by the LSA and LDA models are provided in Table 4.1 and Table 4.2, respectively. The context of the topics that was identified by each model and the semantic relationship of the topics with diabetes were examined and the results are presented and discussed below.

In regard to the output of the LSA model, it was found to be reasonably consistent with the content of the data set. Therefore, it can be concluded that the LSA model is suitable for generating topics automatically for the context of interest to this study. As can be seen from Table 4.1, each topic consists of words related to a particular diabetes issue. One of the reasons for the precise and plausible distinction between topics is that the number of topics was already known due to the data having been obtained from three different known headings. However, the fact that the classification of the results was almost the same as the headings in the website proves the success of the model. In addition, when 5 and 4 were assigned to the number of topics, using the trial-and-error method, it was observed that similar topics were identified by the LSA model.

Table 4.1
Topics Identified by the Latent Semantic Analysis Model

Topic 1	work	time	diabetes	year	job	good	people	type	help	day
Topic 2	drive	dvla	licence	year	test	medical	hypo	form	months	car
Topic 3	people	diabetes	help	disable	dla	benefit	care	good	claim	pip

Table 4.1 shows that the first topic generated by the LSA model consists primarily of words that can be related to working life, namely, *work, time, job, year, day, etc.* The second topic identified by the LSA model clearly addresses driving-related issues, as almost every word (*drive, licence, test, medical, form, dvla, car*) in the topic is relevant to this domain, except for hypo (hypoglycaemia). Since the rest of the words clearly come together under one topic,

it may be deduced that hypoglycaemia is one of the challenges that diabetics encounter during driving or when applying for a driving licence. Finally, Topic 3 contains six words that bring out the subject of the discussion, namely, *disable, diabetes, dla (Disability Living Allowance), help, benefit, pip (Personal Independence Payment)*. From an examination of the meaning of these words, it can be concluded that the people in the forum were discussing a number of benefits that diabetics can access.

The topics generated by the LDA model are presented in Table 4.2. As can be seen, the LDA model also extracted three topics from the data set, such as the LSA model the number of topics was defined by the user. The overview of the topics by the LDA model also seems to have been successful. However, the topic contents are not as consistent as those of the LSA model.

Table 4.2
Topics Identified by the Latent Dirichlet Allocation Method

Topic 1	dvla	licence	drive	year	medical	test	sent	form	time	letter
Topic 2	people	help	work	year	diabetes	good	benefit	dla	time	thanks
Topic 3	work	diabetes	time	drive	job	year	hypo	type	employer	good

From Table 4.2, it can be seen that Topic 1 is about driving, as this can be inferred from the words, *dvla*, *drive*, *test*, *licence* and *form*. Topic 2 includes words related to benefits. However, these words are not as distinctive as those produced by the LSA model. Only the words *benefit*, *dla*, and *help* imply that benefits is the subject of the discussion. The last topic, Topic 3, appears to consist of some expressions about working life, namely, *work*, *job*, *year*, *time* and *employer*. Nevertheless, these words lack semantic integrity with the others that the LDA model identified as belonging to Topic 3. For instance, it cannot be inferred that hypo (glycaemia) is related to workplace or occupation, even though it is present in the topic. Thus, although not clearly distinguishable, when the words produced by the model were examined, some insights and themes could still be extracted.

In light of the above results, it can be said that both models were generally successful at extracting meaningful topics from the data set and they yielded beneficial information about the diabetes discussion on the forums. However, the results produced by the LSA model, were easier to interpret and semantic integrity was higher. Furthermore, there was a considerable overlap between the headings used by the diabetes forum and the topics extracted by the LSA model.

4.2.2 Multinomial Naïve Bayes, Logistic Regression and Support Vector Machine

The classification metrics for each of the models are presented in Table 4.3. The SVM model achieved the highest mean accuracy (80%), followed by the LR (78%) and the multinomial NB (74%) models. The SVM and LR mean F1 scores were very close to each other (0.79 vs 0.77). However, the F-score of multinomial NB was quite a bit lower (0.67). On the other hand, multinomial NB achieved the best result for precision (0.82), compared with SVM (0.80) and LR (0.77). The average recall scores of SVM, multinomial NB and LR were 0.79, 0.77 and 0.67, respectively.

Finally, Table 4.3 shows that the LR model is almost four times slower (11.38 seconds), compared to SVM (3.80 seconds) and multinomial NB (3.24 seconds). Thus, overall the accuracy measurements and statistics demonstrates that, in this experiment, the SVM model is a more successful than the LR and multinomial NB models.

Table 4.3
Accuracy Statistics and Metrics for the Classification Models

	Accuracy (%)	**Error (%)**	**Recall**	**Precision**	**F-score**	**Elapsed time (seconds)**
Multinomial Naïve Bayes	74%	26%	0.67	0.82	0.67	3.24
Logistic Regression	78%	22%	0.77	0.77	0.77	11.38
Support Vector Machine	80%	20%	0.79	0.80	0.79	3.80

Table 4.4 shows the results of the cross-validation analysis that was conducted to evaluate the stability of the models. According to the cross-validation of 10 different test sets, SVM obtained the best accuracy interval (78.14% to 83.91%), followed by LR (75.83% to 81.3%).

However, even the best-acquired accuracy score achieved by the multinomial NB model (75.2%) is lower than the minimum accuracy score of the SVM and the LR models.

Furthermore, the cross-validation results also showed that the SVM model was able to classify the data set more accurately compared to LR and multinomial NB. However, it should be noted that the LR model was also able to provide reasonable and applicable results.

Table 4.4
Cross-Validation Scores of the Classification Models

	Min. Accuracy (%)	**Max. Accuracy (%)**	**Average Accuracy (%)**	**Std. Dev.**
Multinomial Naïve Bayes	70.03%	75.2%	73.71%	0.01550247
Logistic Regression	75.83%	81.3%	78.8%	0.01381879
Support Vector Machine	78.14%	83.91%	81.04%	0.01504449

When the confusion matrices of each model were examined (please refer to Appendix A1, Figures A1.5 to A1.7), it was found that the multinomial NB model was able to achieve a remarkably high level of prediction accuracy (93.08%) for the "Driving and DVLA" class. The SVM model (84.01%) also achieved a high prediction accuracy that was noticeably higher than that of the LR model (79.40%). On the other hand, the prediction accuracy for the "Benefits" class was very close for all models; they were all above 80.25%. However, the accuracy of the SVM model (80.93%) was slightly better than that of the other two models. Finally, the performance of SVM and LR was very similar in respect of estimating the "Jobs and Employment" class: SVM had a 72.30% prediction accuracy while LR had 72.14%. On the other hand, multinomial NB had an unexpectedly low prediction accuracy (24.7%). Referring back to Figure 3.1, the "Jobs and Employment" class in the data set contained the largest number of samples. However, in the test set, it had the lowest number of samples. This may explain the poor multinomial NB result, which may have been affected by the data distribution in the test and training sets.

Chapter 5: DISCUSSION

In this chapter, the concepts and approaches for the text mining of online data and particularly those related to the aspects of topic modelling and text classification are explained. The current study applied text mining algorithms to unstructured data. This data consisted of 9008 posts on a diabetes discussion forum that were provided to the researcher by the www.diabetes.co.uk. This study also addressed the use of data from online health forums as the primary source of data in academic research, and the challenges that this poses. Furthermore, this study also contributed in a small way to the debate on the usage of online health communities by researchers and the general public.

Online health forums contain a huge amount of data and thus are a potentially valuable resource. Therefore, it is important to consider and evaluate the information produced by these communities, because many people act in compliance with the advice that they receive from online health websites (Mead et al., 2003). However, inaccurate instruction and knowledge may cause people to get the wrong treatment, misuse drugs, etc. (Sillence et al., 2007).

An examination of the data set obtained for this study revealed that 952 of 9008 posts contained the expression "I think". Furthermore, there were 2529 uses of the following subjective expressions: *I think, I really/actually think, my perspective, I believe, I do believe, in my opinion, I guess, I figure, it seems to me, my view, my mind, as I see it, it seems, if you ask me, as I see, for me, I bet, I suggest, I would suggest, I can suggest, I assume, as far as I know, I find, I found, it sounds to me.* However, these were not the only subjective expressions used in the data set. Hence the total usage of subjective expressions was actually higher. Nevertheless, it can clearly be concluded that the majority of the posts were based on personal experiences. This finding is in line with Taylor et al. (2015), and Willis and Royne (2017), who stated that people with diabetes use these platforms for getting advice, sharing their own knowledge and finding support.

Although the current study confirmed that people exchange information and often convey their own opinions in these platforms, the quality and reliability of this information is still a matter of debate. Therefore, the validity of the insights extracted from such data is also debatable and this poses an obvious problem for studies using this type of data. Lederman et

al. (2014) stated that to prevent this problem, experimental and non-medical information could be supported by scientific facts, thus the contents of the statements and information online would become more reliable, and the consequent increased trust of the members in the information provided would enhance these platforms in terms of their veracity and validity.

Evidently, these statements must be identified by extracting data from people's conversations. Hence, the textual data must be processed and prepared properly. In any text mining approach, an essential part of that approach is preparing the data for analyses. Although some fundamental preprocessing steps are commonly conducted, such as the removal of stop words, stemming, etc., for most text mining studies (Leskovec et al., 2014), additional operations are needed according to the data set because the standard NLP rules face the challenge of dealing with data extracted from online platforms (Farzindar & Inkpen, 2015).

Therefore, in order to comprehend the structure of the data set obtained for this study in detail, first, the original posts were examined to detect abnormalities that were particular to the used data set, and the necessary processes were carried out to overcome these abnormalities. This included, for instance, the removal of repeated quotes and replies, replacing abbreviations with original words (if necessary), adding a full stop (".") at the end of sentences, correction of non-standard spelling etc. These operations were conducted with great care, because all the algorithms used the same processed data for building the models. Also, in contrast to the relevant literature (Chen et al., 2018; Robinson et al., 2016; Vijayarani et al., 2015), the stemming process was conducted manually, instead of using a provided library, in order to prevent potential errors such as inaccurate removal of suffixes. In addition to introducing errors, the results showed that automatic stemming decreased the accuracy of the models.

Moreover, as explained in Section 3.4.II, 'Exploratory Analysis', words and orders/commands frequently used in everyday language were also detected and deleted from the data set to make the results more consequential. According to the outputs, this process was repeated several times (11 times) until the data set was sufficiently cleaned and the models provided better results. This process was one of the most time-consuming and challenging parts of the study because it required a lot of attention and human interpretation at every step.

Even though the data set was cleaned as much as possible, some problems could not be obviated. As the data set was extracted from an online discussion forum, several issues occurred due to the nature of the language used in such forums. As an example, as pronouns are used instead of nouns in conversations within a thread, it is difficult to determine which names are being mentioned without manual analysis or human intervention. To be more specific, the reply to the post saying *"I have type 1 diabetes and my son has the same."* is *"I am sorry for you and how old is he?"* but it only makes sense when the whole conversation is considered.

However, every post is a different input for the models. Therefore "he" and "you" do not refer to the user and her son in every case. Problems of interpretation arise not only in the case of pronouns and nouns, but also with respect to references that are also frequently used for referring to objects. In this regard, many studies (Kennington & Schlangen, 2015; Purwarianti et al., 2016; Sharaf et al., 2018; Williams & Scheutz, 2015) have proposed models for reference resolution which involve replacing these words with their references to solve this problem.

The current study explored the potential and effectiveness of using the LSA and LDA algorithms to extract the themes of the discussions on an online forum. Based on the results of various evaluations, it was concluded that although both models gave acceptable outputs, the LSA model produced more functional and accurate results. However, even if the models succeeded in extracting the topics that people mentioned to a certain extent, the detailed information about the content of the topics could not be ascertained. For example, although both models clearly identified one of the topics as being related to driving, the exact details of the driving issues were not detected. However, the performance of the LSA model in regards to the "Benefits" topic should be considered separately from this generalisation, because it was able to produce more comprehensive, elaborate and informative outputs for this topic.

After the main analyses had been completed, one of the text summarisation algorithms was also implemented to summarise the inputs. However, as each entry was too short to summarise, with some entries consisted of only three words, no meaningful and significant results could be obtained. Therefore, this step in the analysis process was ignored and is not reported here.

Many previous studies have compared LDA and LSA. However, there is no consensus as to whether one model is better than the other (Cvitanic et al., 2016; Kang et al., 2019; Lee, Song & Kim, 2010). Moreover, Bergamaschi, Po and Sorrentino (2014) used the LSA and LDA models for a recommendation system and they also found that the output of LDA is not effective for suggestions, whereas LSA achieves good results.

The fundamental challenge and likewise limitation of both algorithms is that the results need to be read and deduced by human judgement, manually. The minimisation of human intervention through automatic analysis is considered one of the main advantages of text mining, for instance, using topic modelling approaches rather than reading documents manually to extract themes (Hagen, 2018). Nevertheless, manual intervention needs to be part of the overall process in order to determine the number of topics, assign the relevant label to each topic, gain insights from the results and find associations between documents and topics (DeGroof et al., 2018; Qi et al., 2019; Wood et al., 2017). In this respect, human interference was needed not only for the LSA and LDA algorithms, as mentioned above, but also for the data cleaning phase. Even the evaluation of these models is based mostly on manual judgement and interpretation, therefore the results tended to be subject to human bias.

The current study also assessed the accuracy of three different text classification models – multinomial NB, LR, SVM – on the multiclass data set selected for this study. Section 4.2, 'Results', described how several metrics were used to evaluate the performance and effectiveness of these models. In this study, the SVM model provided more accurate results compared with LR and multinomial NB. This finding is in line with that of Lucini et al. (2017), who used eight different text mining approaches, including the three algorithms used in this study, to analyse 16,703 textual medical records, and found that linear SVM provided the best f-score (77.59%) and precision (75.62%) followed by LR (76.95%, 74.02%) and multinomial NB (76.55%, 69.44%). However, the authors also found that multinomial NB yielded the highest recall score (85.30%) compared to linear SVM and LR. Moreover, Mohammad et al. (2016) also used NB and SVM approaches to classify textual data and they deduced that SVM is slightly better than NB. Similarly, Brindha et al. (2016) conducted a comprehensive survey of text classification algorithms and their effectiveness on different kinds and sizes of data sets. The authors increased the number of samples from 200 to 5000 for two different data sets and observed that SVM was able to perform better than LR and NB

on both data sets. They also found that NB was able to achieve higher f-score and recall results compared to LR.

Even if previous studies are valuable in terms of supporting or evaluating other research, their findings are not sufficient to clearly identify which machine learning algorithm to use with which system (Olson et al., 2016). Currently, there is no best or worst machine learning algorithm, every one of them has its own strengths and drawbacks. Apart from the models themselves, the data sets also affect prediction and performance, even when using the same samples. Yet, the cleaner the data set the more accurate the results. Therefore, it is crucial to benchmark algorithms by taking into account the desired criteria and system requirements.

In this regard, a number of issues were identified that affected the results produced by the models that were evaluated in this study. Figure 3.1 showed that the amount of data under the "Jobs and Employment" heading was, on average, twice that under the "Benefits" heading. Consequently, the imbalanced data set caused bias in the result and also affected the distribution of the training and test sets. As indicated in Section 4.2, 'Results', the extremely low prediction rate of multinomial NB for the "Jobs and Employment" class indicated that the distribution of data set had a significant effect on model accuracy. Furthermore, some ambiguous or vague expressions in the data set made the classification process difficult. Such expressions included, for instance, sentences that did not contain any specific arguments about the belonging class, and inputs that were in the form of short replies, such as "Thank you!", "You are welcome" and "I am sorry for you". As the data set for this study was retrieved from a discussion forum, such expressions were frequently encountered. The proportion of missing or insufficient statements was also reflected in the prediction scores. In addition, as a result of applying the trial-and-error method, it was observed that the accuracy of the algorithms increased by adjusting the parameters specified in Section 4.2, 'Results'. Olson et al. (2017) also indicated that parameter tuning improves the proportion of correct predictions on a large scale in plenty of machine learning problems.

Chapter 6: CONCLUSION AND FUTURE WORK

This study described the application of two machine learning approaches, supervised and unsupervised, on data received from www.diabetes.co.uk. The analysis involved a data cleaning phase and an exploratory analysis, which was followed by the implementation of the models. The LSA and LDA unsupervised machine learning algorithms were applied to identify themes in the online discussions. It was concluded from the results of the LSA and LDA models that the data set was based on individuals' experiences related to diabetes and three main topics, regarding jobs and employment, driving issues and benefits and help to support those with diabetes were extracted from the discussions. According to performance metrics presented in Section 4.1, the LSA model produced a better result than the LDA model. On the other hand, multinomial NB, LR and SVM, which are supervised machine learning algorithms, were implemented to perform multiclass text classification. The results showed that SVM performed better than LR and multinomial NB in respect of recall, accuracy, and f-score.

The current study also provided detailed information about all the text mining processes undertaken in order to help researchers who may wish to conduct a similar study. In addition, the challenges and limitations encountered by this study were presented in this chapter in order to ease and assist future studies.

Furthermore, several recommendations can be made to extend the study reported in this dissertation. First, the data frames were sufficient to save the data set that was used in this study. For a larger data set, it is recommended that a database system is used to manage the data as this would be an easier and faster way to access and process the data. Second, more feature extraction and/or combined features could be used in the preprocessing phase. In addition, more parameters can be adjusted to increase the accuracy of the models. Moreover, to eliminate the short-length message problem, and references issues, every message under a thread could be merged to create one long post. Through the application of this method, reference resolution could be applied and text summarisation algorithms could be used to get more detailed information from online discussions. Lastly, more discussion topics from diabetes.co.uk could be examined in order to gain a more in-depth understanding of the issues facing those affected by diabetes.

References

Abdalgader, K. (2017). Clustering Short Text using a Centroid-Based Lexical Clustering Algorithm. *IAENG International Journal of Computer Science*, *44*(4)

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer, Boston, MA.

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *6*(1).

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Alsmadi, I., & Hoon, G. K. (2018). Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications*, 1-13.

An, Y., Sun, S., & Wang, S. (2017, May). Naive Bayes classifiers for music emotion classification based on lyrics. In 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 635-638). IEEE.

Anandarajan, M., Hill, C., & Nolan, T. (2019). Semantic Space Representation and Latent Semantic Analysis. In Practical Text Analytics (pp. 77-91). Springer, Cham.

Aphinyanaphongs, Y., Lulejian, A., Brown, D. P., Bonneau, R., & Krebs, P. (2016). Text classification for automatic detection of e-cigarette use and use for smoking cessation from twitter: a feasibility pilot. In Biocomputing 2016: Proceedings of the Pacific Symposium (pp. 480-491).

Babcock, M. J., Ta, V. P., & Ickes, W. (2014). Latent semantic similarity and language style matching in initial dyadic interactions. Journal of Language and Social Psychology, 33(1), 78-88.

Badry, R. M., Eldin, A. S., & Elzanfally, D. S. (2013). Text summarization within the latent semantic analysis framework: comparative study. International Journal of Computer Applications, 81(11), 40-45.

Bergamaschi, S., Po, L., & Sorrentino, S. (2014). Comparing Topic Models for a Movie Recommendation System. In *WEBIST (2)* (pp. 172-183).

Bernardi, R. & Wu, P. (2017). The Role of Online Health Communities in Patient Empowerment An Empirical Study of Knowledge Creation and Sharing on diabetes.co.uk. Available at https://pure.royalholloway.ac.uk/portal/en/projects/the-role-of-online-health-communities-in-patient-empowerment-an-empirical-study-of-

knowledge-creation-and-sharing-on-diabetescouk(8b0651a4-721d-42f0-bae5-7087307d2b99).html

Brennan, P. F., & Ripich, S. (1994). Use of a home-care computer network by persons with AIDS. International journal of technology assessment in health care, 10(2), 258-272.

Brindha, S., Prabha, K., & Sukumaran, S. (2016, January). A survey on classification techniques for text mining. In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1-5). IEEE.

Chen, A. T., Zhu, S. H., & Conway, M. (2015). What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. *Journal of medical Internet research*, *17*(9), e220.

Chen, L. C. (2017). An effective LDA-based time topic model to improve blog search performance. *Information Processing & Management*, *53*(6), 1299-1319.

Chen, P. H., Zafar, H., Galperin-Aizenberg, M., & Cook, T. (2018). Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *Journal of digital imaging*, *31*(2), 178-184.

Chen, Y. Y., Li, C. M., Liang, J. C., & Tsai, C. C. (2018). Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. *Journal of medical Internet research*, *20*(2).

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creativity and Innovation Management*, *26*(1), 17-30

Christou, D. (2016). Feature extraction using Latent Dirichlet Allocation and Neural Networks: A case study on movie synopses. *arXiv preprint arXiv:1604.01272*.

Colton, D. (2016). Text Classification Using Python. *Text Mining and Visualization: Case Studies Using Open-Source Tools*, *40*, 199.

Corcoglioniti, F., Dragoni, M., Rospocher, M., & Aprosio, A. P. (2016, May). Knowledge extraction for information retrieval. In *European Semantic Web Conference* (pp. 317-333). Springer, Cham.

Corley, C., Cook, D., Mikler, A., & Singh, K. (2010). Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, *7*(2), 596-615

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Cummings, J. N., Butler, B., & Kraut, R. (2002). The quality of online social relationships. *Communications of the ACM*, *45*(7), 103-108.

Cvitanic, T., Lee, B., Song, H. I., Fu, K., & Rosen, D. (2016, January). Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents. In *International Conference on Case-Based Reasoning*.

Dang, H. X., & Lawrence, C. B. (2014). Allerdictor: fast allergen prediction using text classification techniques. Bioinformatics, 30(8), 1120-1128.

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inform Sci 1990; 41: 391-407.

DeGroof, R., Xu, H., Zhang, J., & Liu, R. (2018). Mining Significant Terminologies in Online Social Media Using Parallelized LDA for the Promotion of Cultural Products. In *Proceedings of the 14th International Conference on Data Science (ICDATA'18)* (pp. 3-9).

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. arXiv preprint arXiv:1610.09982.

Duru, I., Diri, B., Özçevik, M. E., Ataseven, K., Dogan, G., & White, S. (2018, October). Analysis of English Language Groups with Regular Expressions. In 2018 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-5). IEEE.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, *64*(2-3), 221-245

E Hilliard, M., M Sparling, K., Hitchcock, J., K Oser, T., & K Hood, K. (2015). The emerging diabetes online community. *Current diabetes reviews*, *11*(4), 261-272.

Ekinci, E., & Omurca, S. İ. (2017). Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *9*(1), 51-58.

Farzindar, A., & Inkpen, D. (2015). Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, *8*(2), 1-166.

Fiksdal, A. S., Kumbamu, A., Jadhav, A. S., Cocos, C., Nelsen, L. A., Pathak, J., & McCormick, J. B. (2014). Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. Journal of medical Internet research, 16(10).

Finn, J. (1999). An exploration of helping processes in an online self-help group focusing on issues of disability. *Health & Social Work*, *24*(3), 220-231

Folias A, Brown AS, Carvalho J, Wu V, Close KL, Wood R. (2014). Investigation of the presence and impact on patients of diabetes social stigma in the USA. Diabetes; 15 (Suppl 1): 59-LB.

Gao, J., Liu, N., Lawley, M., & Hu, X. (2017). An interpretable classification framework for information extraction from online healthcare forums. *Journal of healthcare engineering*, *2017*

Guenther, N., & Schonlau, M. (2016). Support vector machines. *The Stata Journal*, *16*(4), 917-937.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483.

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?. *Information Processing & Management*, *54*(6), 1292-1307.

Hasan, A., Levene, M., & Weston, D. J. (2017, October). Natural language analysis of online health forums. In *International Symposium on Intelligent Data Analysis* (pp. 125-137). Springer, Cham

Hernandez M. (2014) Personal communication with Jeff Hitchcock, 9.

Ibrahim, N. F., Wang, X., & Bourne, H. (2017). Exploring the effect of user engagement in online brand communities: Evidence from Twitter. *Computers in Human Behavior*, *72*, 321-338.

International Hypoglycaemia Study Group. (2015). Minimizing hypoglycemia in diabetes. *Diabetes Care*, *38*(8), 1583-1591.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169-15211.

Jiang, Y., Song, X., Harrison, J., Quegan, S., & Maynard, D. (2017, September). Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism* (pp. 25-30).

Kang, H. J., Kim, C., & Kang, K. (2019). Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA). *Processes*, *7*(6), 379.

Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer science*, *2*(2), 194-200.

Kazmierska, J., & Malicki, J. (2008). Application of the Naïve Bayesian Classifier to optimize treatment decisions. Radiotherapy and Oncology, 86(2), 211-216.

Kennington, C., & Schlangen, D. (2015, July). Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 292-301).

Khunti, K., Davies, M., Majeed, A., Thorsted, B. L., Wolden, M. L., & Paul, S. K. (2015). Hypoglycemia and risk of cardiovascular disease and all-cause mortality in insulin-treated people with type 1 and type 2 diabetes: a cohort study. *Diabetes care*, *38*(2), 316-322.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, *19*(2), 65.

Kremic, E., & Subasi, A. (2016). Performance of random forest and SVM in face recognition. *Int. Arab J. Inf. Technol.*, *13*(2), 287-293.

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. (2016). Assessing the Big Five personality traits with latent semantic analysis. Personality and Individual Differences, 102, 229-233.

Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In *null* (pp. 32-36). IEEE.

Lederman, R., Fan, H., Smith, S., & Chang, S. (2014). Who can you trust? Credibility assessment in online health forums. Health Policy and Technology, 3(1), 13-25

Lee, S., Song, J., & Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, *51*(1), 1-10.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Levandoski, A., & Lobo, J. (2018). Document and Topic Models: pLSA and LDA.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016, July). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 165-174). ACM.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, *48*(2), 354-368.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. 2016. An Overview of Topic Modeling and Its Current Applications in Bioinformatics. SpringerPlus, 5(1): 1608. S

Lucini, F. R., Fogliatto, F. S., da Silveira, G. J., Neyeloff, J. L., Anzanello, M. J., Kuchenbecker, R. D. S., & Schaan, B. D. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International journal of medical informatics*, *100*, 1-8.

Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. Quantitative Finance, 15(6), 999-1012.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. Natural Language Engineering, 16(1), 100-103.

Manning, C., Socher, R., Fang, G. G., & Mundra, R. (2017). CS224n: Natural Language Processing with Deep Learning1.

Martin, J. H., & Jurafsky, D. (2018). Speech and language processing Chapter 5: Logistic Regression

McCarthy, R. V., McCarthy, M. M., Ceccucci, W., & Halawi, L. (2019). Predictive Models Using Regression. In *Applying Predictive Analytics* (pp. 89-121). Springer, Cham.

Mead, N., Varnam, R., Rogers, A., & Roland, M. (2003). What predicts patients' interest in the Internet as a health resource in primary care in England?. *Journal of health services research & policy*, *8*(1), 33-39.

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*, 28.

Mohammad, A. H., Alwada'n, T., & Al-Momani, O. (2016). Arabic text categorization using support vector machine, Naïve Bayes and neural network. GSTF Journal on Computing (JoC), 5(1), 108.

Momtazi, S. (2018). Unsupervised Latent Dirichlet Allocation for supervised question classification. *Information Processing & Management*, *54*(3), 380-393.

Mukku, S. S., Choudhary, N., & Mamidi, R. (2016). Enhanced Sentiment Classification of Telugu Text using ML Techniques. *SAAIP@ IJCAI*, *2016*, 29-34.

Nagpal, A., & Gabrani, G. (2019, February). Python for Data Analytics, Scientific and Technical Applications. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* (pp. 140-145). IEEE.

Nath, C., Huh, J., Adupa, A. K., & Jonnalagadda, S. R. (2016). Website sharing in online health communities: a descriptive analysis. *Journal of medical Internet research*, *18*(1).

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine* learning research, 12(Oct), 2825-2830.

Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web* (pp. 91-100). ACM.

Porter, M. F., "An Algorithm for Suffix Stripping",*Program*, 130-137, 1980.

Pranckevičius, T., & Marcinkevičius, V. (2016, November). Application of logistic regression with part-of-the-speech tagging for multi-class text classification. In 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE) (pp. 1-5). IEEE.

Purwarianti, A., Andhika, A., Wicaksono, A. F., Afif, I., & Ferdian, F. (2016, August). InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-5). IEEE.

Qi, X., Wang, X., & Lv, S. (2019, March). Heuristic Theme Clustering for Online Social Media. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence* (pp. 96-100). ACM.

Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. arXiv preprint arXiv:1410.5329

Ravert, R. D., Hancock, M. D., & Ingersoll, G. M. (2004). Online forum messages posted by adolescents with type 1 diabetes. *The Diabetes Educator*, *30*(5), 827-834.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016, April). Forecasting student achievement in MOOCs with natural language processing. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 383-387). ACM.

Rousseau, F., Kiagias, E., & Vazirgiannis, M. (2015, July). Text categorization as a graph classification problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1702-1712).

Sainani, K. L. (2014). Logistic regression. PM&R, 6(12), 1157-1162.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, *1*(3), 261-377.

Sedgwick, P. (2013). Logistic regression. Bmj, 347, f4488

Sharaf, A., Gupta, A., Ge, H., Naik, C., & Mathias, L. (2018). Cross-Lingual Approaches to Reference Resolution in Dialogue Systems. *arXiv preprint arXiv:1811.11161*.

Shi, H., Gerlach, M., Diersen, I., Downey, D., & Amaral, L. A. (2019). A new evaluation framework for topic modeling algorithms based on synthetic corpora. *arXiv preprint arXiv:1901.09848*.

Shree, S. B., & Sheshadri, H. S. (2018). Diagnosis of Alzheimer's disease using naive Bayesian classifier. Neural Computing and Applications, 29(1), 123-132

Sillence, E., Briggs, P., Harris, P. R., & Fishwick, L. (2007). How do patients evaluate and make use of online health information?. *Social science & medicine*, *64*(9), 1853-1862.

Süzek, T. Ö. (2017). Using latent semantic analysis for automated keyword extraction from large document corpora. Turkish Journal of Electrical Engineering & Computer Sciences, 25(3), 1784-1794.

Tang, P. C., Overhage, J. M., Chan, A. S., Brown, N. L., Aghighi, B., Entwistle, M. P., ... & Perkins, A. J. (2012). Online disease management of diabetes: engaging and motivating patients online with enhanced resources-diabetes (EMPOWER-D), a randomized controlled trial. *Journal of the American Medical Informatics Association*, *20*(3), 526-534.

Taylor, J., Pagliari, C., & Osborne, M. (2015). Understanding the social dynamics of Twitter, Facebook and Diabetes. co. uk and their value implications for patients and health researchers.nhanced resources-diabetes (EMPOWER-D), a randomized controlled trial. Journal of the American Medical Informatics Association, 20(3), 526-534.

Taylor, J., Pagliari, C., & Osborne, M. (2015). Understanding the social dynamics of Twitter, Facebook and Diabetes. co. uk and their value implications for patients and health researchers.nhanced resources-diabetes (EMPOWER-D), a randomized controlled trial. Journal of the American Medical Informatics Association, 20(3), 526-534.

Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science, 57, 821-829.

Üstüner, M. (2013). Destek vektör makineleri yöntemi ile arazi kullanımı sınıflandırılmasında kernel fonksiyonlarına ait karşılaştırmalı parametre duyarlık analizi: Rapideye ve spot örneği (Doctoral dissertation).

Vadivel, P. S., Krishnan, S. N., & Yuvaraj, D. (2018). AN EFFECTIVE DOCUMENT CATEGORY PREDICTION SYSTEM USING SUPPORT VECTOR MACHINES, MANN-WHITNEY TECHNIQUES. *International Journal of Pure and Applied Mathematics*, *118*(22), 895-900.

van der Eijk, M., Faber, M. J., Aarts, J. W., Kremer, J. A., Munneke, M., & Bloem, B. R. (2013). Using online health communities to deliver patient-centered care to people with chronic conditions. Journal of medical Internet research, 15(6).

van Rossum, G., & de Boer, J. (1991). Interactively testing remote servers using the Python programming language. *CWi Quarterly*, *4*(4), 283-303.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Vrana, S. R., Vrana, D. T., Penner, L. A., Eggly, S., Slatcher, R. B., & Hagiwara, N. (2018). Latent Semantic Analysis: A new measure of patient-physician communication. Social Science & Medicine, 198, 22-26

Wang, C. , & Blei, D. M. (2013). Variational inference in nonconjugate models. Journal of Machine Learning Research, 14 (1), 1005–1031.

Wang, P., Bai, G. R., & Stolee, K. T. (2019, February). Exploring Regular Expression Evolution. In 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER) (pp. 502-513). IEEE.

Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2 (pp. 90-94). Association for Computational Linguistics.

Wang, X., Zhao, K., & Street, N. (2017). Analyzing and predicting user participations in online health communities: a social support perspective. *Journal of medical Internet research*, *19*(4)

Wang, Y., & Liu, Z. (2007). Automatic detecting indicators for quality of health information on the Web. International Journal of Medical Informatics, 76(8), 575-582

Warrer, P., Hansen, E. H., Juhl-Jensen, L., & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *British journal of clinical pharmacology*, *73*(5), 674-684.

Williams, T., & Betak, J. (2018). A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. *Procedia computer science*, *130*, 98-102.

Williams, T., & Scheutz, M. (2015). A Domain-Independent Model of Open-World Reference Resolution. In *CogSci*.

Willis, E., & Royne, M. B. (2017). Online health communities and chronic disease self-management. *Health communication*, *32*(3), 269-278.

Wolf A, Liu N. (2014). The numbers of shame and blame: How stigma affects patients and diabetes management

Wood, J., Tan, P., Wang, W., & Arnold, C. (2017, April). Source-LDA: Enhancing probabilistic topic models using prior knowledge sources. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)* (pp. 411-422). IEEE.

Yi, X., & Allan, J. (2009, April). A comparative study of utilizing topic models for information retrieval. In *European conference on information retrieval* (pp. 29-41). Springer, Berlin, Heidelberg.

Yıldıztepe, E., & Uzun, V. (2018). Olasılıksal Yöntemler ile Türkçe Metinlerin Anlamsal Benzerliğinin Belirlenmesi. Sinop Üniversitesi Fen Bilimleri Dergisi, 3(2), 66-78.

Appendix A1- Confusion Matrix

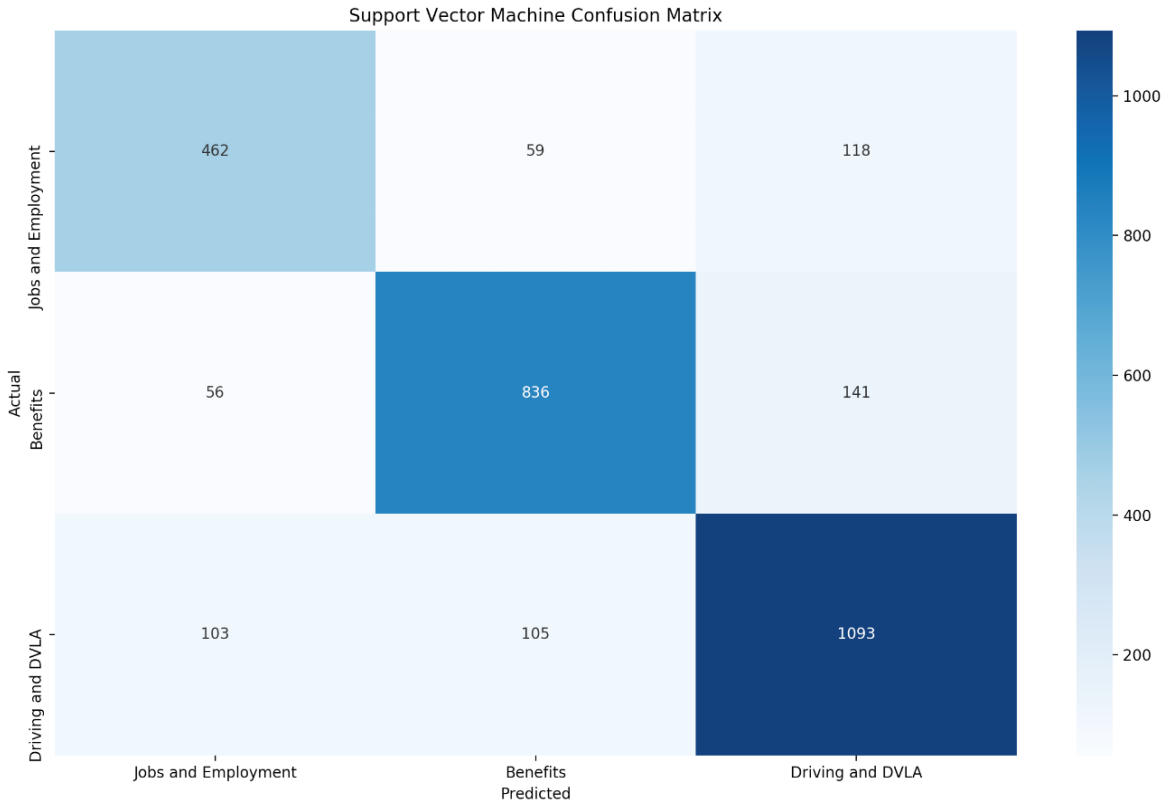

Figure 5
Confusion Matrix for SVM

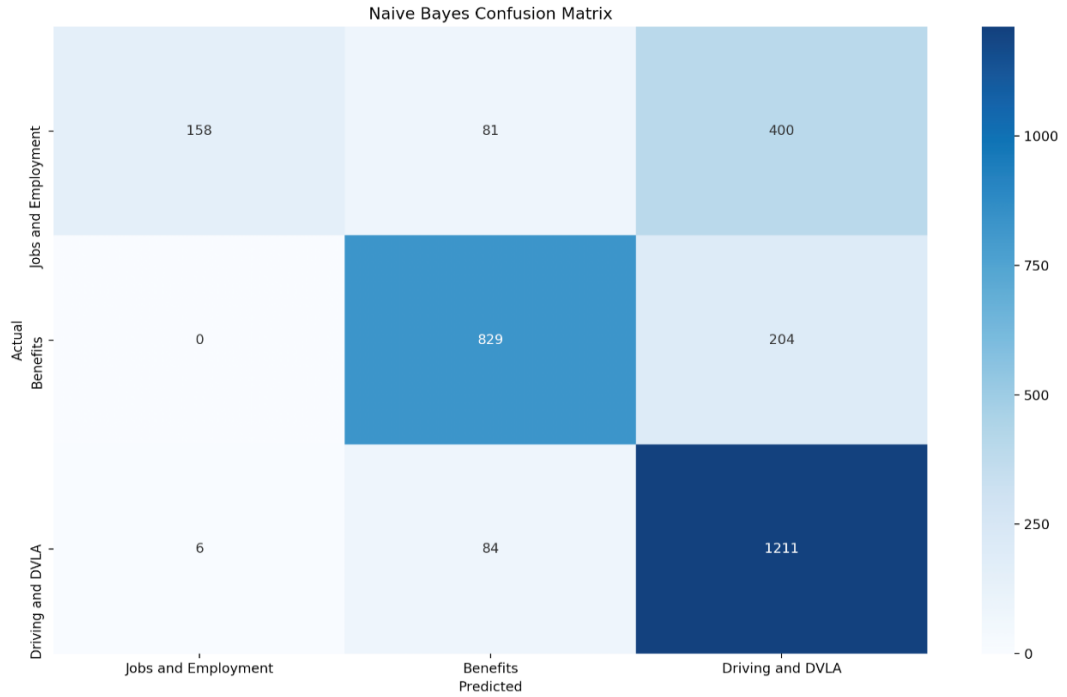

Figure 6
Confusion Matrix for Multinomial Naïve Bayes

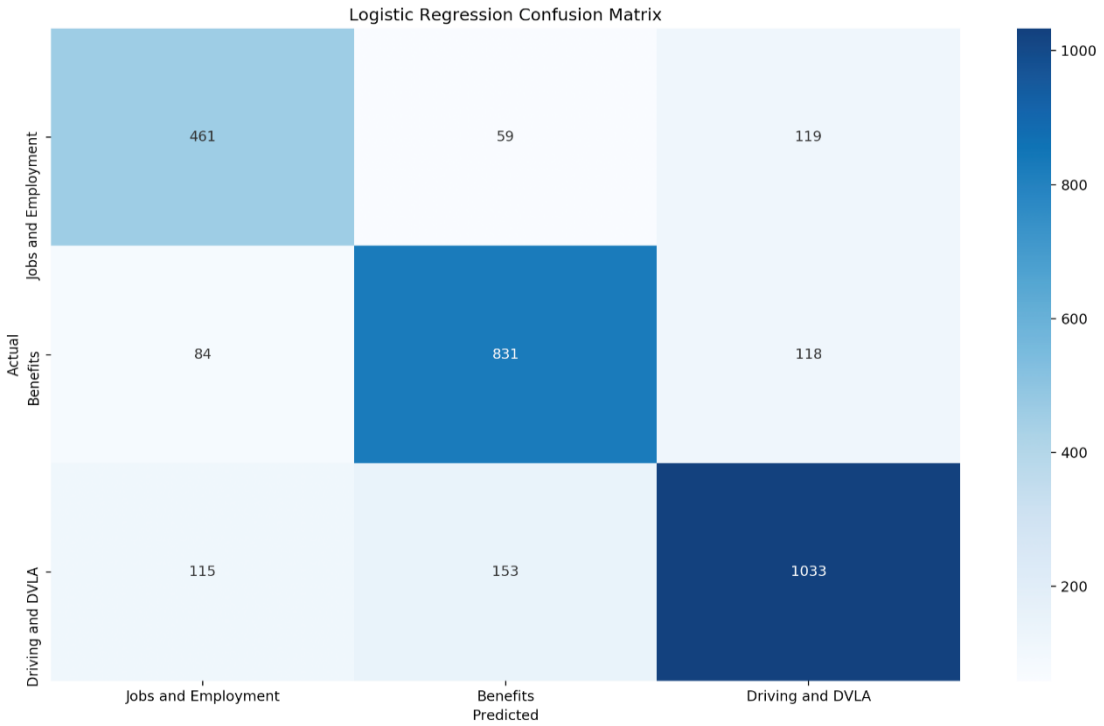

Figure 7
Confusion Matrix for Logistic Regression

Appendix A2 –Ethic Application Form

The University Of Sheffield.

Application 028336

Section A: Applicant details

Date application started:
Mon 10 June 2019 at 09:11

First name:
Nurbanu

Last name:
Aksoy

Email:
naksoy1@sheffield.ac.uk

Programme name:
Data Science

Module name:
Dissertation
Last updated:
10/06/2019

Department:
Information School

Applying as:
Undergraduate / Postgraduate taught

Research project title:
Examining people's discussion about living with diabetes by using natural language processing and text mining techniques

Has your research project undergone academic review, in accordance with the appropriate process?
Yes

Similar applications:
- not entered -

Section B: Basic information

Supervisor

Name	Email
Laura Sbaffi	l.sbaffi@sheffield.ac.uk

Proposed project duration

Start date (of data collection):
Mon 10 June 2019

Anticipated end date (of project)
Mon 10 June 2019

3: Project code (where applicable)

Project code
- not entered -

Suitability

Takes place outside UK?
No

Involves NHS?
No

Human-interventional study?
No

ESRC funded?
No

Likely to lead to publication in a peer-reviewed journal?
Yes

Led by another UK institution?
Yes

Involves human tissue?
No

Clinical trial?
No

Social care research?
No

Involves adults who lack the capacity to consent?
No

Involves research on groups that are on the Home Office list of 'Proscribed terrorist groups or organisations?
No

Indicators of risk

Involves potentially vulnerable participants?
No
Involves potentially highly sensitive topics?
Yes

Section C: Summary of research

1. Aims & Objectives

The overall purpose of this thesis is to have information about the lives of people with diabetes using posts from the website diabetes.ac.uk. This will be accomplished with the application of text mining techniques and NLP on the data. It also aims to contribute to in the health context such as diabetes, diabetes online forums and acquiring health information online.

There are 4 main objectives of this research which are:
I. To figure out problems relating to diabetes in work and issues about employment;
II. To investigate whether there exist any benefits of having diabetes in law and social life;
III. To explore diabetes patients' experiences of driving and Driver and Vehicle Licensing Agency;
IV. To extract the general concern of members in each sub-title.

2. Methodology

NLP and text mining methods will be used in this study to understand the essential meaning of all the posts belonged three decided titles. NLP will be used for especially to comprehend individuals' language and will also be used text pre-processing. At the end of the research, the general issues about these titles will be detected. In addition, Python programming language will be used for implementation.

3. Personal Safety

Have you completed your departmental risk assessment procedures, if appropriate?

No

Raises personal safety issues?

No

The personal information will not be used in the data, but also it will be anonymized. In addition, the study does not require group work or university labs. Therefore, there is no predictable risk for the researcher.

Section D: About the participants

1. Potential Participants

The data will be provided by the organization and all posts written under certain titles will be sent to me. Therefore, there are no determined characteristics for participants.

2. Recruiting Potential Participants

As stated in the above question, one-to-one communication will not be made with the participants and all the posts written by the users will be sent to me by the organization.

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by CiCS? No

- not entered -

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) No

Since the data is collected from an online website, it is not possible to reach all users for obtaining their consent. However, NDA has been signed between the website owner and me.

4. Payment

Will financial/in kind payments be offered to participants? No

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

As personal information of the participants will not be used, there is no (predictable) potential harm/distress.

How will this be managed to ensure appropriate protection and well-being of the participants?

The data will be anonymised by the organization before it is sent to me. No one, including the researcher, will be able to know about the owners of the data.

Section E: About the data

1. Data Processing

Will you be processing (i.e. collecting, recording, storing, or otherwise using) personal data as part of this project? (Personal data is any information relating to an identified or identifiable living person).
No

Please outline how your data will be managed and stored securely, in line with good practice and relevant funder requirements

The data in this study will not be shared with third parties and will only be used for this dissertation by me.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
No

Consent forms relevant to project?
No

Additional Documentation

Document 1063610 (Version 1) All versions
Non-Disclosure Agreement

External Documentation

- not entered -

Section G: Declaration

Signed by:
Nurbanu Aksoy
Date signed:
Mon 10 June 2019 at 11:31

Offical notes

- not entered -

Appendix A3- Ethical Approval

The University Of Sheffield.

Downloaded: 15/08/2019
Approved: 10/06/2019

Nurbanu Aksoy
Registration number: 180223132
Information School
Programme: Data Science

Dear Nurbanu

PROJECT TITLE: Examining peoples discussion about living with diabetes by using natural language processing and text mining techniques
APPLICATION: Reference Number 028336

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 10/06/2019 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 028336 (dated 10/06/2019).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely

Emma Cooper
Ethics Administrator
Information School