A Text-Mining Based Approach to Capturing the NHS Patient Experience

A thesis submitted for the degree of Doctor of Philosophy

By

Mohammed Bahja

Department of Computer Science,

Brunel University

February 2017

ABSTRACT

An important issue for healthcare service providers is to achieve high levels of patient satisfaction. Collecting patient feedback about their experience in hospital enables providers to analyse their performance in terms of the levels of satisfaction and to identify the strengths and limitations of their service delivery. A common method of collecting patient feedback is via online portals and the forums of the service provider, where the patients can rate and comment about the service received. A challenge in analysing patient experience collected via online portals is that the amount of data can be huge and hence, prohibitive to analyse manually.

In this thesis, an automated approach to patient experience analysis via Sentiment Analysis, Topic Modelling, and Dependency Parsing methods is presented. The patient experience data collected from the National Health Service (NHS) online portal in the United Kingdom is analysed in the study to understand this experience. The study was carried out in three iterations: (1) In the first, the Sentiment Analysis method was applied, which identified whether a given patient feedback item was positive or negative. (2) The second iteration involved applying Topic Modelling methods to identify automatically themes and topics from the patient feedback. Further, the outcomes of the Sentiment Analysis study from the first iteration were utilised to identify the patient sentiment regarding the topic being discussed in a given comment. (3) In the third iteration of the study, Dependency Parsing methods were employed for each patient feedback item and the topics identified. A method was devised to summarise the reason for a particular sentiment about each of the identified topics.

The outcomes of the study demonstrate that text-mining methods can be effectively utilised to identify patients' sentiment in their feedback as well as to identify the themes and topics discussed in it. The approach presented in the study was proven capable of effectively automatically analysing the NHS patient feedback database. Specifically, it can provide an overview of the positive and negative sentiment rate, identify the frequently discussed topics and summarise

individual patient feedback items. Moreover, an API visualisation tool is introduced to make the outcomes more accessible to the health care providers.

ACKNOWLEDGEMENTS

I would not have been able to complete this thesis without the aid and support of the kind people around me. First, I would like to express my earnest and heartfelt gratitude to my first supervisor, Professor Mark Lycett, for his advice, comments, and motivation throughout my Ph.D. journey. His guidance, broad knowledge and critical thinking have been of great value to me throughout this research. In addition, deep appreciation is given to my second supervisor Dr David Bell for his great feedback and support.

I would like to give thanks to my external advisor, Dr Mohammed Al Madhoun, for his valuable feedback. Words cannot express my gratitude to him.

I also acknowledge and thank *The NHS Choices* Service Desk Team for their magnificent help and kind support in providing for me the NHS Choices dataset from their back-end database engine.

A special thanks to my family, parents, brothers and sister, for providing me with unfailing support and continuous encouragement throughout my years of study. Words cannot express how grateful I am to my father and mother for all of the sacrifices that they have made on my behalf. Their praying for me was what sustained me thus far.

I would also like to thank all of my friends who supported me in my writing, and incentivised me to strive towards my goal.

Finally, I must express my very profound gratitude to my beloved wife Nour, who has just started her PhD journey, for her endless love and support at all times.

DECLARATION

The following papers have been published (or submitted for publication) as a direct result of the research discussed in this thesis:

- "Identifying Patient Experience from Online Resources via Sentiment Analysis and Topic Modelling", IEEE/ACM - BDCAT2016.
- "A Text-Mining Based Approach to Understanding the NHS Patient Experience", Information Systems Frontiers (Paper has been submitted).

ABBREVIATIONS

API: Application programming interface

ANN: Artificial Neural Networks

AUC: Area under the ROC Curve

DP: Dependency Parser

DSR: Design Science Research

K-NN: K-Nearest Neighbour Classifier

LDA: Latent Dirichlet Allocation

LSI: Latent Semantic Indexing

ML: Machine Learning

NB: Naıve Bayes Classifier (NB).

NHS: National Health Service

NLP: Natural Language Processing OAEI:

OM: Opinion Mining

POS: Parts of speech

R: A programming language and software environment for statistical computing and graphics

ROC: Receiver Operating Characteristic

SA: Sentiment Analysis (SA)

SoA: Strength of Association model

SVM: Support Vector Machines Classifier

TM: Topic Modelling

URL: Uniform Resource Locator

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1 Overview

User experience is a crucial aspect of any product or service. There is an immense focus by the product makers and service providers to provide the highest quality of user experience to the consumers. In the healthcare industry, the patients are the primary consumers of the provided service and their user experience is often referred to as patient experience.

The widespread use of intranet and internet enabled devices has enabled means of capturing user experience mainly through online portals, such as company websites, review sites, social media among others. In the healthcare industry, hospitals are increasingly collecting feedback from the patients about their experiences in hospital through their feedback sections on their websites. This has led to a large collection of patient experience data that have become increasingly difficult to analyse manually by hospital staff.

Advanced technologies, such as text mining and natural language processing, provide dependable methods of automated analysis of text data for identifying user sentiment and feedback topics. These technologies have been extensively explored in the research community for user experience analysis. However, they have not been explored enough for the analysis of patient experience. In this thesis, the study presented aims at addressing the research gap on the automated analysis of patient experience via Text Mining.

This chapter introduces the thesis as well as outlines the outcomes of the study on patient experience analysis via text mining approaches. The research motivation and the research problem being addressed in the thesis are first discussed. This is followed by presentation of the aims and objectives of the research, with the approach used in this project being also discussed. Finally, the structure of the thesis is described.

1.2 Research Motivations

1.2.1 Importance of Patient Experience

Providing health care to all is a major goal of countries across the globe especially within the developed countries where providing optimal health care is on top of the agenda of national priorities. In addition to maintaining health facilities, training a competent work force and creating informative policies, the success of a health care system is positively correlated with that of patient care and experience (Coulter, Fitzpatrick and Cornwell, 2009a). In recent years, the trend has been moving towards considering patient experience as an important indicator of performance and quality of health care organisations (Jason *et al.*, 2014). Much research has gone into the importance and active role that patient experience plays in the overall success of a health care system (Goodrich & Cornwell, 2008)

Health care within England, Scotland and Wales is mainly provided by the National Health Service (NHS), a public health service established shortly after the Second World War, which caters for all residents of the UK. The NHS puts a strong emphasis on being comprehensive, universal and free at the point of delivery, but recently, also on the patient experience, which the NHS describes as a core dimension of good quality care (Coulter, 2012). It is an imperative task for the NHS to measure patient experience in order to improve care, monitor health care performance, enhance strategic decision making and record progress for health care organisations (LaVela and Gallan, 2014).

With a static financial budget and the ever-growing demand on health care, the NHS faces many challenges whilst striving to achieve a great quality of healthcare. If patient experience is not of sufficient quality or if inappropriate care is provided, it leads to poor outcomes for patients and improper use of resources, which in turn leads to a higher burden on the NHS (Alderwick *et al.*, 2015). After recent economic and political upsets within the UK, further financial burdens upon the NHS is one that the UK could afford. As a result, understanding patient experience is therefore considered as being an investment rather than a cost, as it leads to an increase in positive health outcomes. This sentiment can also be described as follows:

> "*Robust patient experience initiatives provide critical information on the operation of organisations, their outcomes and opportunities for improvement. As such, there should not be a trade-off between efficiency and patient experience, as excellent experience can lead to improved efficiency"* (Coulter *et al.*, 2014).

Therefore, understanding patient experience enables hospitals to identify their weaknesses in providing healthcare. It provides opportunities for them to reflect on their functioning and thus, make efforts towards addressing the limitations in relation to the service provided. Further, understanding the patient experience can also contribute towards making the hospital processes more efficient, which in turn will lead towards better utilisation of resources and the addressing of patient concerns. For instance, patient feedback about aspects, such as longer waiting times or parking issues, can be used by the hospital to identify the underlying causes. This will allow for those responsible to pinpoint what the issues are and address them accordingly.

1.2.2 NHS Outcomes Framework

Providing high quality patient experience is a priority for the NHS organisation. In this context, the NHS developed a NHS Outcomes Framework that monitors its progress in England in providing positive patient experience. It publishes quality improvement and outcome measurements for monitoring the performance of the organisation, which is updated annually. Each year, the NHS Outcomes Framework provides appropriate measures for that year, which is then used to implement improvements to the existing indicators and develop new ones. The framework has five different domains for the indicators and a summary of these is listed below (Department of Health, 2016):

- *Domain 1*: Preventing people from dying prematurely and reducing the number of avoidable deaths e.g. premature deaths, especially within children and people with learning disabilities;
- *Domain 2*: Enhancing and improving the quality of life for patients with long-term conditions; ensuring patients feel supported when managing their health conditions;
- *Domain 3*: Helping patients recover from injury and illness and how to prevent them from occurring in the future; especially improving in the areas of injuries, traumas, strokes and dental health;
- *Domain 4*: Ensuring that patients, service users and carers have a positive experience of care; improving the experience of a patient in a number of key healthcare areas, such as outpatients and accidents and emergencies;
- *Domain 5*: Implementing patient safety by providing a safe environment in order to deliver better health outcomes and reduce the incidence of avoidable harm.

In the 2015-2016 NHS Outcomes Framework report, four indicators were included that were primarily focused on ensuring that the patients have a positive experience of care, these being: patient experience of primary care, patient experience of hospital care, friends and family test, and the patient experience characterised as poor or worse. The areas for improvement that the NHS needed were identified as: people's experience of outpatient care, hospitals' responsiveness to personal needs, people's experience of accident and emergency services, access to primary care services, women and their families' experience of maternity services, the experience of care for people at the end of their lives, experience of healthcare for people with mental illness, children and young people's experience of healthcare and people's experience of integrated care (Group, 2014). In sum, the purpose of NHS Outcomes Framework is to receive patient feedback and improve care as well as providing evidence that positive outcomes are being achieved (Coulter *et al.*, 2014).

Research conducted by the Picker Institute who conducted extensive work with patients have found the most important areas of patient experience for a patient include but not limited to (Picker Institute, 2009)

- Having fast access to trustworthy and reliable health related advice;
- Effective and efficient treatment administered and delivered by healthcare professional they can trust;
- Respect, empathy and emotional support;
- Support and more involvement from carers and family members;
- Continuity of care.

The Picker Institute has produced a detailed guide broken down into a number of sections where they elaborate upon their findings. The guide is targeted especially at health care professionals who care about patient experience and want to bring about improvements.

1.2.3 Web Resources for Collecting Patient Experience

Understanding patient experience is an important goal for the NHS, as this can be used to develop indicators and measures for improving the public health infrastructure and policy. The best method of understanding patient experience is by directly collecting patient feedback and understanding this information first hand from the point of view of the patient. There are many advantages of this approach and these include (Coulter, Fitzpatrick and Cornwell, 2009b):

- Enabling understanding of the current problems in healthcare delivery;
- Informing how to deliver continuous improvement, in particular, through the redesigning of services;
- Helping professionals to reflect on their own and their team's practice;
- Monitoring the impact of any changes;
- Facilitating benchmarking between services/organisations;
- Comparing organisations for performance assessment purposes;
- Informing clinicians involved in referrals about the quality of services;
- Informing commissioners and patients about the quality of services;
- Informing patients about care pathways;
- Helping patients to choose high quality providers;
- Enabling public accountability.

Thus, there has been an increased effort in recent years oriented towards collecting direct patient feedback and some of the widely-used methods of collecting patient feedback include interviews, questionnaires and personal surveys with the patients. These methods are considered to be effective as they provide the patients better privacy and anonymity while providing feedback. Another method used but not quite as effective as elaborated upon below is face-to-face conversations with patients.

Many factors contribute towards a patient's experience and as these involve dynamic processes, it is important to understand that not all aspects of patient experience can be completely captured by a set of questionnaires and/or interviews. Furthermore, in direct face to face interviews/conversations or when providing feedback forms for the patients, this can put patients in an "awkward" position, i.e. make them feel the need to be polite about their experience so as not to offend the healthcare providers.

With technology on the rise and people finding themselves more connected with computer devices, one recently more popular and effective method of collecting a patients' feedback is to provide an online forum, where they can provide their views on the health care they have received on a website feedback form at their own convenience. The online forum approach is being widely adopted as this allows the users to choose anonymity while providing feedback and thus, relieves them of the fear of being rude in face-to-face encounters. Further, it is also convenient for patients to provide feedback through websites, because they can do so at their own space, in comfort and it gives them a chance to edit their responses as well as providing multiple feedback. Moreover, collecting feedback via online resources facilitates better storage, access and management of the patient feedback database, which is much more challenging with paper, based patient feedback collection and processing.

In 2008, the NHS created a website called *NHS Choices* that invites patients to leave their feedback about their experience with the healthcare providers of the hospital, which they can do using two approaches. One method is to provide ratings to a given hospital on different metrics. Regarding which, on the feedback website there are several questions about different aspects of patient care, which the users can rate on a scale of 1 to 5 stars. Moreover, in addition to rating their visit or the span of their care for different metrics, they can also leave feedback in the form of comments in the comments section of the website. Thus, the NHS has a large database of patient feedback covering most hospitals across the United Kingdom (UK) and the number of comments and feedback across the database runs into the hundred thousand.

1.2.4 Analysing a Web Based Patient Experience Database

The process of collecting patient feedback via website source provides the healthcare professionals easy and convenient access to patient feedback data that can be used for analysis purposes. The hospital administration and researchers can use the feedback database to identify the strengths and weaknesses of their healthcare service and further employ them to address the identified issues.

For instance, a study by Greaves and colleagues examined hospital-level associations between web-based patient ratings on the NHS Choices website and objective measures of quality (Greaves *et al.*, 2012a). They performed a cross-sectional observational study of 166 NHS acute hospital trusts in England using 10,274 patient web-based ratings. The study reported that 68% of the patients agreed to recommend the hospital to a friend. Furthermore, they found a significant association between positive recommendations of the hospital and lower hospital standardised mortality ratios, lower mortality from high-risk conditions and lower readmission rates (Greaves *et al.*, 2012a). This research established a significant relationship between patients' online feedback and objective measures of clinical quality as well as showing that web-based ratings are a useful tool for patients and health care workers. Thus, online resources of patient feedback can be effectively used for identifying various measures related to patient care and clinical quality. However, at the same time, due to the enormity and the complexity of the patient feedback database, with time, it becomes increasingly challenging to analyse. The hospital staff will find it increasingly difficult to go through all the patient reviews as this require an enormous amount of manual work, which can be quite demanding and even ineffective.

Online reviews are generally written as free text and do not adhere to a structure or format. This makes analysing and understanding the patient experience more challenging than when dealing with closed questions, as the possibilities regarding the feedback content are endless. Techniques, such as keyword searching can help in searching for topics in patient feedback; however, this does not recognise positive and negative feedback. Further, the number of reviews could be so overwhelming that it would be challenging for the hospital to identify various features of patient feedback from the review analysis (Ganu, Marian and Elhadad, 2010). It is essential for a hospital to know how much feedback is positive from the patients and what are the aspects or features of the care that are affecting their experience. Identifying sentiment and extracting the topics from the patient feedback manually in the context of online free text, such as with the NHS Choices website, is nearly impossible given the complexity involved. However, this is important because this form of surveying is becoming increasingly popular and being seen as becoming indispensable by healthcare providers.

To address the challenge of analysing a large patient feedback database, automated methods would enable the analysis process to be more sustainable and time effective. The recent advancements in information analysis technologies, such as text mining and natural language processing, has resulted in them being widely applied to analyse the user experience of products by various companies. However, there have been very few studies that have explored how such language processing methods could be utilised to analyse patient experience. In the study presented in this thesis, text mining and natural language processing methods are explored to provide a framework for the automated analysis of a patient experience database, which is aimed at identifying patient sentiment about the care provided in a particular hospital. The outcomes of the study show that through the use of these processing methods, the themes and topics hidden in given patient feedback can be automatically identified and also, that the associated sentiment can be captured. The aims and objectives of the study are presented in the next section.

1.3 Research Aims and Objectives

Given the above, the aim of the project can be described as:

To provide a framework for an automated analysis of patient feedback database that investigates the effectiveness of using NLP methods to better capture and understand patient opinion about healthcare service from textual feedback.

The aim of this project was implemented using the patient feedback database obtained from the *NHS Choices* website and was underpinned by the following research objectives:

- Identify and evaluate the relevant scientific literature, methods, tools and technologies to develop a critical understanding of the literature;
- Investigate Sentiment Analysis models to achieve an automatic approach of accurately identifying patient sentiment from patient feedback;
- Explore and apply Topic Modelling approaches to the patient feedback database to identify topics and themes hidden in this feedback. Subsequently, to associate the predicted topics with the sentiment identified through the Sentiment Analysis methods applied earlier;
- Identify and evaluate the natural language processing (NLP) methods to identify keywords that can automatically summarise patient feedback, thereby enabling a quick and automated survey of the database
- Provide visualisation techniques where relevant to facilitate an interactive and user friendly interface for the healthcare professionals to understand and analyse the patient feedback

1.4 Research Approach

An appropriate research methodology is important to follow during the course of a scientific research project in order to achieve the desired outcomes. For this project, the Design Science Research (DSR) approach was followed. DSR has become a widely utilised approach for research within many fields, especially within Information Systems (Kuechler and Vaishnavi, 2008). Under this paradigm, a set of methods focusing on the design, implementation and testing of an artefact with the aim of improving its performance in relation to a given issue is deployed (Lukka, 2003). It is an iterative process, whereby the artefact development cycle is repeated such that its performance regarding the matter concerned is continuously being enhanced. An important characteristic of DSR is that it focuses mainly on problem solving (Orlikowski and Iacono, 2001).

The DSR method was deemed appropriate for the project, because the thesis research focuses on a real-world problem which required solving in practice (Hevner *et al.*, 2004). The project is aimed at building innovative solutions or artefacts that can automatically capture the sentiment of the patient from their feedback and also extract topics discussed within it. Further, the process of building the artefact is a multi-stage process and the learning outcomes of each stage are utilised in the next stage towards building the final artefact. Lastly, DSR is openly linked to prior theoretical knowledge (Lukka, 2003).

As mentioned above, the research is carried out in three main iterations. In the first, the patient sentiment in their feedback is automatically identified. In the second, topics from the patient feedback are extracted and the sentiment identified from the first iteration is then associated with the identified topic. In the third iteration of the research, the possible underlying reasons for the sentiment regarding each of the topics identified are automatically produced. In sum, in this research, using the DSR paradigm, the outcomes and knowledge acquired at each stage of the research iteration is transferred to the next stage for further development of the artefact.

The DSR method is carried out in five phases. Different chapters in this thesis pertain to the different stages of the DSR method. These being:

a) **Awareness of the Problem -** This stage involves identifying the problem and understanding its complexities and details, i.e. describing the problem on hand, elaborating upon it and also predicting what would happen if the problem is not addressed (Wieringa, 2009). The awareness of the problem stage is elaborated in Chapter 2 where a detailed literature review is carried out. This chapter reviews the available research articles and outlines the current shortcomings in patient experience analysis in order to obtain an awareness of the problem.

b) **Suggestion -** Based on the understanding of the problem, suggestions are made on how to carry out research to identify potential solutions and the solution, which would be deemed best fit. Suggestion can be seen as the creative step where functionality of the solution is envisioned. The suggestion phase is described in Chapter 2 where the NLP methods that can be utilized for patient experience analysis is identified and described.

c) **Development -** This is where the design is implemented and developed. The main tasks here are implementing the identified solution and developing an artefact by incorporating the solution into the context of the problem. However, the novelty remains in the design and not the actual development. The development phase is seen at each iteration of the research because the design and development of the artefact is carried out at each iteration of the research. These are outlined in Chapter 4, 5, and 6.

d) **Evaluation -** Evaluating the performance of the artefact with relevant performance measure metrics, such as accuracy, specificity, etc. Further, based on the performance measures obtained, if required, adapt the artefact to provide a higher performance measure or discard it and find another solution that achieves better performance. The evaluation of the artefact is done after the artefact development and hence is carried out at each iteration of the research. The evaluation process is elaborated for each iteration in Chapter 4, 5, and 6.

e) **Conclusion -** Finally, at the end of the research cycle there is the disseminating of the artefact or the outcomes of the research process via the deployment of the developed application and the production of reports and other relevant documentation. A detailed conclusion of the research is provided in the conclusion chapter, Chapter 7.

There are many applications of DSR as recorded in a number of literatures (Vaishnavi and Kuechler, 2007; Hevner *et al.*, 2004; Lukka, 2003; March and Smith, 1995; Kasanen, Lukka and Siitonen, 1993). From these works, it can be established that one of the core processes involved in DSR is establishing awareness of the problem and intended solution along with the research potential and theoretical contribution. Another process is that of the development of the artefacts/solutions that serve a purpose and to evaluate this when the solution is implemented and tested to see how well it works. A third process is theory building i.e. identifying and analysing its theoretical and research contribution. A more detailed comparison of these literatures has been performed by Rocha *et al.*(2012).

Compared to Natural Science Research, which focuses on being problem-orientated, DSR is more solution-orientated. This approach therefore ensures a practical and expert software prototype, which in return delivers rich data for theory testing and evaluating.

1.5 Thesis Structure

The rest of this thesis is structured as follows:

Chapter 2 provides a detailed literature review relevant to the thesis topic. The chapter first introduces patient experience via the description of user experience. Next, there is discussion on the science of natural language processing with a special emphasis on Sentiment Analysis, Topic Modelling and Dependency

Parsing methods. The Sentiment Analysis approach is discussed in detail and the relevant literature regarding its application for user experience as well as patient experience is covered. Similarly, the Topic Modelling and Dependency Parsing approaches in this context are also discussed. Finally, there is a consideration of the advantages, challenges and limitations of the available Text Mining approaches.

Chapter 3 provides a detailed discussion on the research methodology followed for the thesis. The Design Science Research (DSR) methodology is introduced and its adaptation for the research is discussed. A common framework approach is first introduced and then its adaptation to each iteration of the research is provided. Next, the five phases of the research process: awareness of the problem, suggestion, development, evaluation and conclusion are explained in detail.

Chapter 4 discusses the first iteration of this research, which pertains to the application of the Sentiment Analysis methods to the *NHS Choices* database. A detailed introduction to Sentiment Analysis approaches is provided and the most commonly used methods are discussed. The chapter also describes the patient feedback database that will be used throughout the research and the associated pre-processing provided in each chapter. Three Sentiment Analysis approaches that are shortlisted and implemented for the study are: the Support Vector Machine (SVM) along with the Naive Bayes (NB) and Strength of Association (SoA). Their implementation is then described in some detail. Finally, the outcomes of the research iterations are discussed in terms of the performance of each Sentiment Analysis model in relation to their level of accuracy in predicting patient sentiment.

Chapter 5 discusses the second iteration of the research. In this chapter, the Topic Modelling method is introduced with a special focus on the Latent Dirichlet Allocation (LDA) method being provided. Further, the unigram and bigram methods of Topic Modelling are discussed and the manner of its implementation regarding the NHS patient feedback dataset is provided. The chapter shows that using the Topic Modelling methods, the topics and themes present in the patient feedback can be automatically identified. Further, for each topic, its associated sentiment is identified by applying the sentiment scores obtained by the SVM method from iteration one of the research to identify the sentiment of each predicted topic drawn from the patient feedback.

Chapter 6 reports the third iteration of the research. In this chapter, the application of Dependency Parsing on the patient feedback database is provided. Two different approaches of Dependency Parsing, namely, the openNLP method and the coreNLP method, are applied to the NHS database to find "noun-adjective" pairs from the patient feedback. These pairs provide an automated summary of the feedback and indicate what sort of adjective is associated with a given noun in a patient's sentence. Finally, the performance validation of these approaches and the visualisation of the outcomes also discussed.

Chapter 7 concludes the thesis. It summarises the research findings and draws conclusions. It also discusses the research contributions of this research and evaluates whether the research objectives were met or not. The limitations of the research are also covered and potential future works to extend the present study are provided.

Chapter 2: Literature Review

2.1 Overview

This chapter provides a detailed literature review of the topics related to this research. A detailed discussion of text mining methods relevant to this research, including Sentiment Analysis, Topic Modelling and Dependency Parsing is provided. This chapter discusses the rationale behind selecting these methods of text mining for analysing patient experience.

The chapter is organised as follows: Section 2.2 introduces user experience and considers the patient experience topic that is central to this research. Section 2.3 introduces natural language processing methods and gives an overview of the various available methods. Section 2.4 introduces the Sentiment Analysis (SA) method, which is followed by detailed discussion of its approaches in section 2.5, its levels in section 2.6, its applications in section 2.7 and its challenges in section 2.8. A literature review of Sentiment Analysis for user experience is provided in section 2.9. Section 2.10 discusses Topic Modelling methods and finally, section 2.11 discusses Dependency Parsing methods.

2.2 User Experience

2.2.1 Definition and Significance

For a given product or service, the user experience is extremely important, as it is a measure of how satisfied the end user is with the functionality of the product or the service. The International Standards Organization (ISO) defines the user experience as *"…a person's perceptions and responses that result from the use or anticipated use of a product, system or service…"* (Steyvers and Griffiths, 2007).

It includes various parameters, such as the end user`s preferences, emotions, perceptions, beliefs, their behaviours as well as their physical and psychological responses before, during and after the use of a given product. The user experience is further dependent on other factors, such as the user, the product and the context of the use (Uys, Du Preez and Uys, 2008).

Understanding user experience is very important in the present consumer driven society where there is a huge consumption of products and services. It helps the service providers and the product designers to identify the strengths and weaknesses of their products (Uys, Du Preez and Uys, 2008). User experience helps in collecting a precise assessment of the quality of the product or service provided. It also facilitates the adaptation of the product functionalities to suit the contextual requirements of the users. In sum, to facilitate expansion of product reach and enhance the sustainability of a given product, it is essential that product designers collect and analyse user experience data to identify and overcome the limitations of their product (Blei, 2012).

2.2.2 User Experience Metrics

Understanding or quantifying user experience is challenging since it is subjective. People are diverse, with each person`s preferences, expectations, and experience being different (Tullis and Albert, 2008). Therefore, to collect reliable user experience information, it is necessary that such data is collected from a large pool of users and analysed. Further, it is also important that the right approach be used while collecting user information and one approach for it is to use metrics that are designed to collect precise user experience.

The *user experience metrics* topic is vast and there exists a wide range of metrics, which capture different aspects of user experience. For instance, task success is a metric, which captures the degree of success of the task execution, thereby determining whether there has been a positive user experience (Rodden, Hutchinson and Fu, 2010). Some of the most commonly used metrics are the time on task, success rate and user errors (Tullis and Albert, 2008). Given there are a variety of user experience metrics available, the choice is dependent on the application.

To capture the user experience for a service, such as customer service in a restaurant or a grocery store, collecting feedback directly from the customers via direct interaction with them is an effective approach to understanding their opinions about the service provided (Meyer and Schwager, 2007). However, some of the metrics mentioned above are more relevant in capturing the user experience for a given device, such as computers or web applications. There is an increasing online presence of various service providers collecting customer feedback about their service via online sources, such as on their websites, from review forums, through blogs or other similar portals (Meyer and Schwager, 2007). This approach provides an easy, efficient, and quick way of collecting customer feedback about a service.

Despite the advantage of easy and quick access to customer feedback provided by online sources, a disadvantage is that the volume of data needing to be read to understand customer opinion can be vast and distributed over multiple sources. Moreover, with the increasing usage of social media and the popularity of reviewing services, the volume of data to be analysed is complex, often unstructured and distributed widely. An effective approach to extracting customer opinion about a given product from a large volume of data is to use *Natural Language Processing* methods. Using these methods, it is possible to identify the opinion or the sentiment of the customers for a given product. In the next section, a brief introduction to *Natural Language Processing* is given followed by a detailed description of some of its methods.

2.2.3 Patient Experience

As mentioned above, user experience depends on the context of the service provided. For instance, in an online gaming application, it is dependent on factors such as the speed with which the game is streamed, the graphics of the game, the user interface and ease of communication between multiple players, amongst others. Thus, the user experience for a particular product or service is unlikely to be valid for another product or service.

Over the years, healthcare has evolved into a major service sector. Currently, healthcare is provided as a service to the patients and the patients have become its consumers. Similar to other services, the user experience of the patients in healthcare is important for enabling improvements to the service. In the healthcare domain, the user experience is often termed as *Patient Experience*. This is a broad concept, also covering terms such as patient perceptions, opinions and satisfaction. There is no standard definition of patient experience, however, a widely accepted definition has been provided by The Beryl Institute, USA:

"…The sum of all interactions, shaped by an organisation's culture that influence patient perceptions across the continuum of care…" (Ramage, Manning and Dumais, 2011).

Patient experience is more than patient satisfaction, being dependent on a patient's expectation as compared with the personalised care provided and whether this expectation was positively fulfilled (Hair, 2009). Patient experience is a conglomerate of all events that happen over the period of the care and is closely related to the patients, their families, their life style as well as their emotional and physical well-being when they access the health care system (Jason *et al.*, 2014).

Several factors influence patient experience and the key ones include:

Communication – The way information about the patient`s condition is disseminated to him or her and how well his or her issues are listened to are important factors influencing patient experience. When there is not sufficient communication from the hospital staff, the patient experience is known to be negative (Szyca, Rosiek, Nowakowska, & Leksowski, 2012).

Hospital Environment – This factor includes the hospital infrastructure, the patient management, friendliness of the staff along with the nature of the communication between the staff and the patient. These aspects influence the hospital environment and impact on the patient experience (Lane *et al.*, 2016).

Quality of care – The quality of care does not just refer to the diagnosis, for it also includes the attitudes of the staff while providing the care to the patient. Positive attitudes influence the quality of the care provided and the overall patient experience (Edmund *et al.*, 2014).

For the healthcare service providers, it is of utmost importance to achieve the highest levels of patient experience, because of the associated risks to the patient's health and well-being. Consequently, collecting patient experience on a regular basis is highly necessary so that healthcare service providers can identify the areas of strengths and weaknesses that are affecting the patient experience.

2.3 Natural Language Processing

Natural language processing (NLP) applies statistical predictive modelling techniques to understand language for various applications, such as speech recognition, text to speech synthesis, and text mining among others (Berger, Pietra and Pietra, 1996). NLP has been a field of immense research interest with significant works dating back to 1950s and 60s. With the recent advances in computing technologies and extensive research progress in machine learning techniques, NLP is making significant progress and is a widely researched topic for various applications (Manning and Schütze, 1999). NLP, in short, is considered as one of the best ways to understand the language used along with uncovering the sentiment behind it.

NLP is used in several tasks, but in the context of this thesis, the discussion is restricted to two main topics, Sentiment Analysis and Topic Modelling, because only these two approaches are explored throughout this research. However, Dependency Parsing will also get a brief introduction in this chapter in order to show the range of NLP techniques.

- Sentiment Analysis – This approach is targeted at extracting subjective information from a chunk of text data and as the name suggests, it focuses on identifying the sentiment of the given text.
- Topic Modelling – Under this approach, the given text is segmented and classified based on the identified topics embedded in the text (Collobert and Weston, 2008).
- Dependency Parsing – This approach analyses the grammatical structure of a sentence, creating a relationship between "head" words and words that change these heads.

In the following sections, detailed background and a literature review on the topics of Sentiment Analysis and Topic Modelling are provided.

2.4 Sentiment Analysis - Introduction

Sentiment Analysis is a natural language processing method, which aims to identify the sentiment or opinion contained in a given piece of data. As opinion is subjective, it can be said that Sentiment Analysis extracts the *subjectivity* in the given text. For instance, the text *"I feel great today"* refers to a sentiment of being positive due to the presence of the word *"great"*. Whilst Sentiment Analysis refers to identifying the emotion in the text, opinion mining pertains to detecting the *polarity* in the text. Polarity here refers to whether the text falls into a binary category, such as happy or sad, positive or negative. However, it is common to use Sentiment Analysis for polarity identification and thus, the two concepts fall under the same category of natural language processing and are used interchangeably (Cambria, 2013).

Sentiment Analysis, also referred to as opinion mining, refers to computational study of people`s opinion, sentiment, attitude, and emotion towards an entity. The entity can be about another individual or a public figure, a product, such as cinema or electronic device or service providers, such as restaurants and hospitals. The Sentiment Analysis process can be generally represented by the block diagram shown Figure 2-1.

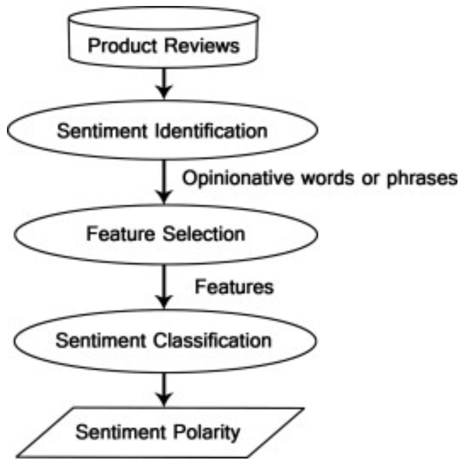


Figure 2-1: Illustration of Sentiment Analysis Process

The first step in the Sentiment Analysis process is to check the database at a lower level, such as sentence or document level, to identify the sentiment at that level. The text in the chosen level is then checked of the existence of an opinion or sentiment. The identification of sentiment is performed based on the presence of words or phrases that are likely to refer to stating an opinion, sentiment or an emotion. If the sentence is identified to have a sentiment or an opinion, then it will be subjected to the feature selection process, which is a more complicated approach. During this process, the sentiment identified is associated with the feature that is being discussed. Finally, the sentiment is then classified into a chosen classification type. For instance, the sentiment can be classified into a binary classification, such as positive or negative (Liu, 2015; Medhat, Hassan and Korashy, 2014b).

According to Liu (2007), a sentiment is dependent on five factors: entity, aspect, orientation of the opinion about the aspect, the opinion holder and the time at which the opinion was given. For example, in the statement, *"the speed of my broadband connection today is amazing"*, the entity is the broadband, the aspect is the speed, the opinion is positive, because of the word amazing, the opinion holder is I and the time when the opinion is stated is today. On an abstract level, Sentiment Analysis can be considered as classifying the opinion or subjectivity in the text into different classes, most frequently, into binary classifications, such as positive or negative. Whilst various methods or approaches for Sentiment Analysis exist, there is no formal classification of these (Vohra and Teraiya, 2013). In the following section, a classification is given based on the different types of approach they consider for extracting the text sentiment.

2.5 Approaches to Sentiment Analysis

The different approaches to Sentiment Analysis can be broadly classified into four categories, namely, keyword spotting, lexical affinity, statistical approaches and concept level Sentiment Analysis, based on their methods and target. Each category is discussed below.

2.5.1 Keyword Spotting

Keyword spotting is a commonly used approach in Sentiment Analysis. The functioning principle of this approach is to identify the sentiment in the text based on the presence of sentiment-conveying words, such as *"happy, sad, depressed, angry, and afraid",* etc., which are called as *"affect words"*. The approach is simple and easy to understand as well as to implement. However, there are two inherent weaknesses and the first is the inability to identify *negation*. In sentences, such as, *"the boss wasn`t happy with the results"*, it is likely that the sentence is classified into positive category due to the presence of the keyword *"happy"* and might well not consider the negation word, *"wasn't"*. Thus, this approach is not effective in such scenarios. The second weakness of this approach is the reliance on *"affect words"*, which restricts its ability to identify the emotions from texts that might not include the listed affect words and yet, convey emotions (Cambria *et al.*, 2013a).

2.5.2 Lexical Affinity

Lexical affinity is a slightly more sophisticated approach, whereby a probabilistic *"affinity"* is assigned to certain words. For instance, words such as *'divorce'* and *'accident'* are assigned a higher probability of negative polarity compared to words such as *"lucky"* and *"sunny"* which would be assigned a higher probability of positive polarity (Hoeksema, 2010). The accuracy performance of this approach is slightly better than keyword spotting; however, it also suffers from certain drawbacks. The negation affect is also a factor for this approach, as lexical affinity cannot effectively identify it. Further, the lexicon treatment performs weakly in relation to identifying the context or domain of the text message (Hu *et al.*, 2013).

2.5.3 Statistical Approaches

The statistical approaches use methods such as Bayesian statistics and machine learning methods, where predictive models are applied to identify the sentiment (Cambria *et al.*, 2013a). These can further be categorised into supervised learning and unsupervised learning approaches. Regarding the former, the predictive models are trained using a labelled training dataset and then applied on new data for Sentiment Analysis. The unsupervised approach is used when there is no dataset available for training and predictive models designed for a different training dataset are used. In addition, there are dictionary based and corpus based approaches for Sentiment Analysis using statistical models. The statistical approaches have relatively higher prediction accuracy and are widely researched methods for Sentiment Analysis. However, one of their drawbacks is the need for a large training dataset to develop the predictive model. Further, due to the reliance on semantic keywords, non-affect keywords are only used in a limited way for prediction (Medhat, Hassan and Korashy, 2014a).

2.5.4 Concept Level Sentiment Analysis

Concept level Sentiment Analysis focuses on the semantic aspects of text by relying on large semantic knowledge bases. This involves using semantic networks or web ontologies, which rely on implicit features of natural language concepts rather than depending on such things as affect keywords (Cambria *et al.*, 2013a). This approach uses feature detection and extraction methods to deconstruct the text into positive or negative emotion, as closely as possible. For instance, in the context of hotel reviews, the term *"small room"* will be considered for a negative polarity, whereas *"small queue"* for a post office is considered as being positive. Hence, concept level analysis considers the context in which the keywords are used and classifies the sentiment accordingly (Medhat, Hassan and Korashy, 2014b).

2.6 Levels of Sentiment Analysis

The sentiment or opinion in a text can be of two types. *Regular* opinion is that whereby the sentiment is about a single product and is relatively straightforward to classify the sentiment. For instance, *"Paris is beautiful"* is a straightforward positive sentiment especially because of the word *"beautiful"*. The other type of sentiment or opinion is *Comparative* opinion, where one product is compared with another or others. For instance, *"Paris is more beautiful than Venice"* is comparative and the sentiment again is a relatively positive one (Liu, 2012). Apart from the classification based on the approaches, Sentiment Analysis can also be classified based on the level at which it is implemented, i.e. sentence, document and entry and aspect levels which are briefly described in the following subsections.

2.6.1 Sentence Level

At this level, the analysis focuses on determining the sentiment at the sentence level. In other words, the analysis is performed for every sentence compared to analysis being performed on the whole document. The expression is identified as any of the following three outcomes: positive, negative or neutral (Poria *et al.*, 2014).

2.6.2 Document level

At this level, the task is to analyse the content of the entire document to determine whether its sentiment is positive or negative. That is, the overall sentiment is evaluated. For instance, in a given review, the document level analysis identifies whether it is positive or negative. The document level Sentiment Analysis is useful for analysis product reviews such as movies, electronic items, etc. However, this type of analysis performs weakly with documents where there are comparative reviews comparing multiple products (McDonald *et al.*, 2007).

2.6.3 Entity and Aspect level

The sentence or document level Sentiment Analysis simply classifies the sentiment into positive or negative and hence, is not suitable for identifying what exactly the review favours or dislikes. In other words, it is not possible to identify the target product that the review is focused on. At entity or aspect level Sentiment Analysis, along with identifying the sentiment, the focus is also on identifying the target. This level of analysis determines which product or entity the sentiment pertains to (Liu, 2012).

The aspect level Sentiment Analysis is also known as feature level analysis. This type of target determining for the sentiment helps in analysing it more precisely. For instance, the text *"the facilities in the hotel were great, but the staff service was not great"*, speaks about different aspects of the hotel. The sentiment is positive about the hotel's facilities; however, it is negative about its service. In this case, the sentiment can be classified into positive and negative for the two aspects separately. Thus, aspect level Sentiment Analysis is more granular than sentence or document level analysis (Agarwal *et al.*, 2011).

2.7 Machine Learning based Sentiment Analysis Techniques

The last decade has seen a tremendous increase in the application of machine learning methods in various fields. The advancements in computing devices' speed and complexity have enabled implementation of machine learning methods for many different applications and Sentiment Analysis methods have greatly benefitted from the adoption of these approaches. A classification of machine learning approach based Sentiment Analysis methods is illustrated in Figure 2-2.

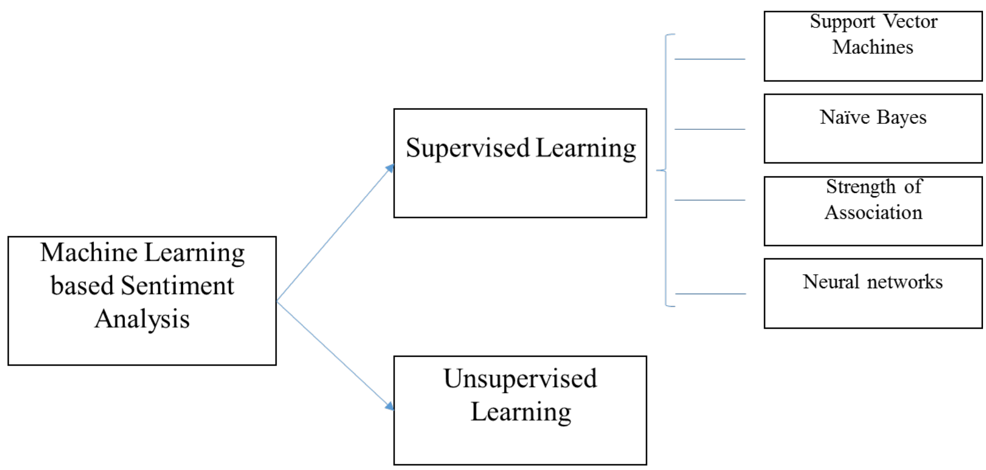


Figure 2-2: Classification of machine learning based Sentiment Analysis methods

As explained in Chapter 2, machine-learning approaches are broadly categorised into supervised and unsupervised learning. The supervised learning methods require a large database of labelled training documents based on which the machine-learning model learns to predict the sentiment. The unsupervised learning methods are used when such databases are not available. In the following, only supervised learning methods are discussed, which are widely popular in sentiment classification. Supervised machine learning techniques require a set of labelled documents, $D = \{X_1, X_2, X_3, \dots, X_n\}$. Based on the features contained within, each document is labelled to a class that is used to train the prediction model. This is then used on an unknown document to predict the class label to which the document may belong. There are many types of supervised learning methods that are used for sentiment classification (Cambria *et al.*, 2013b). Some of those most popular that are used in this research are briefly explained below.

2.7.1 Strength of Association (SoA)

The strength of association (SoA) approach is a popular text mining method that aims to determine how closely the words in a given chunk of text are associated with each other. This model uses a dictionary approach for data prediction. It is based on the information theory concept of "pointwise mutual information'", a technique that measures the similarity between two sets. The model scores individual words, according to the frequency they appear within each individual review type to classify the comment type, thus enabling the assessment of the contribution of each individual word's sentiment to the final score.

The SoA method allows for the determining of the relationships between words in a text based on their occurrence in documents. The strength of association between two words can be calculated using the probability principles, whereby the probability of the occurrence of one word, A, when another word, B, exists is computed. The SoA method uses the binomial distribution to compute the association between words (Agarwal *et al.*, 2015). The SoA between two words, A and B, in a corpus of documents, D, can be calculated using:

$$SoA = \log_e \left(\frac{1}{P(k)} \right);$$

The term $P(k)$ is the probability that the words A and B co-occur in the given corpus. The SoA will give a measure of how likely it is that A occurs in the document when there is B (Agarwal *et al.*, 2015).

In essence, the SoA method is used to identify the strength between words, whilst in Sentiment Analysis it is used to identify the strength of association of the words with a positive or negative sentiment. This enables the prediction model to identify the probability of a given document belonging to a positive or negative sentiment (Agarwal *et al.*, 2015).

2.7.2 Naïve Bayes Classifier

The Naïve Bayes classifier (NB) is a probabilistic classifier, which uses a mixture of models for classification and is widely popular for sentiment classification. Given a document and based on the distribution of words in the document, the NB approach computes the probability of a document belonging to a class. This model calculates boundaries according to the distribution of the words across the labels, whilst at the same time considering the joint probability of the words occurring independently together. Specifically, NB considers each word independently of one another and then tries to estimate the posterior distribution of a review being positive or negative, according to the joint distribution of the words in the review. The probability is computed using the Bayes theorem to predict that a given word belongs to a specific sentiment.

$$P(sentiment|words) = P(sentiment) * \frac{P(words|sentiment)}{P(words)}$$

$P(sentiment)$ is the prior probability of the occurrence of the sentiment, $P(words|sentiment)$ is the probability that a given set of words is classified as a particular sentiment and $P(words)$ is the probability of the occurrence of a given set of words. The NB model assumes independence between words. This assumption is a limitation, because the nature of the English language is dependent on phrases forming ideas and establishing context, nevertheless, this assumption is required for fitting NB models (Liu *et al.*, 2013).

2.7.3 Support Vector Machine model

The support vector machine (SVM) is the most popular machine learning approach that has found immense applications in various fields. It comprises a single layer neural network that performs well for datasets that have less than 100,000 cases. This model maps each observation onto an N-dimensional plane, where N is the number of variables (or in this instance, words). It then finds the two closest observations from each class and tries to find a linearly separable plane that maximizes the distance between them. These observations are called support vectors. With this boundary in place, new observations (or comments) are mapped into this N-dimensional space and labelled, according to the region they fall within.

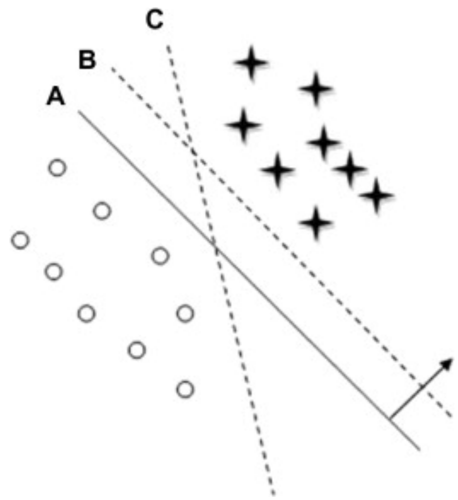


Figure 2-3: Illustration of SVM method based classification.

In Figure 2-3, an illustration of the SVM technique is given. It can be seen that there are two clusters of data and three hyperplanes, A, B and C. The hyperplane A provides the best separation between the two clusters, because the distance to the data points in this hyperplane is the largest. Thus, this distance enables to classify the data points into different clusters and then obtain a classification of the data. The SVM method is suitable for applications, such as sentiment classification and text mining, because of the sparse nature of the text, which allows for features that tend to correlate with each other to be put in linearly classifiable categories and thus, used effectively in decision-making (Vinodhini and Chandrasekaran, 2012).

2.8 Applications of Sentiment Analysis

The task of Sentiment Analysis approaches is to identify the sentiments or opinions embedded in a given text or document. This can be used in various applications, some of which are discussed below.

2.8.1 Review analysis applications

With the increasing popularity of social media, reviews of products are widespread and easily available. Product makers can collect the reviews and analyse them to identify the overall sentiment of the people about the product and subsequently, address the issues and thus, improve their public relations. For instance, movie reviews can be used by filmmakers to identify the general mood of the people and thus analyse their movie from a public perspective. They can then make movies according to people's interests. The presence of review aggregating websites for different types of products (e.g. rotten tomatoes website for movies) could use Sentiment Analysis to consolidate reviews and classify the sentiment towards the products accordingly (Chen and Tseng, 2011).

2.8.2 Sentiment Analysis as an enabling technology

Sentiment Analysis can be used as a subcomponent for other technologies. For instance, recommendations systems are able to use Sentiment Analysis approaches to identify the sentiment towards products and based on the outcomes favour a particular product. Sentiment Analysis can be used for applications, such as ad-placements, based on the sentiment found in the content of webpages and the user interaction with the web-systems. Further, Sentiment Analysis can be utilised in decision tree systems, where based on the sentiment identified, the decisions regarding a certain system can be made (Feldman, 2013).

2.8.3 Organisational applications

For business organisations, Sentiment Analysis is not just useful for identifying general public opinion about the product, for it can also be the avenue for getting more detailed feedback, such as the aspects of a product disliked by people. Additionally, business organisations can deploy it to identify the requirements of people in relation to products, to overcome certain issues with a particular product, to identify features that are desired by consumers and/or to develop new applications or products to cater for current public taste.

There a whole host of applications of Sentiment Analysis that can benefit governmental organisations. For instance, electoral candidates could assess the mood of the public and based on this devise campaign strategy to be in line with public demands. The opinion of the public about government performance could also be identified by these methods. Based on the sentiment identified, the policy making of the government can also be influenced in terms it needing to address widely held misgivings. Further, the intelligence agencies could identify and track negative communications, which might be detrimental to national security and transparency (Pang and Lee, 2008a).

2.9 Sentiment Analysis for User Experience

Possibly the most popular application of Sentiment Analysis is analysing user experience from sources, such as user reviews, feedback forums, social media, among others. Sentiment Analysis enables product manufacturers and service providers to understand the feedback about the user experience from a large pool of users from different locations in a quick, easy, and inexpensive way. Consequently, in the current consumer driven market, Sentiment Analysis is widely employed to understand user experience. In this section, a literature review of the research focusing on Sentiment Analysis regarding user experience is provided.

As mentioned in the previous sections, Sentiment Analysis has various approaches and methodologies. Most Sentiment Analysis approaches focus on classifying the sentiment into positive or negative sentiment and hence, a significant body of work on this can be found in the literature. In the polarity classification literature, the Sentiment Analysis approaches can be broadly categorised into knowledge based and learning based ones, which are discussed next.

2.9.1 Knowledge based approaches

The knowledge-based approaches, also referred to as lexical approaches, have been widely utilised for the polarity classifying of user reviews. These approaches mainly use linguistic models or other forms of knowledge database, such as a dataset of positive or negative words and phrases, for polarity classification. The principle characteristic of this approach is generating and using a dictionary to identify the sentiment of the word. Typically, the approaches here are manual or semi-automated (Melville, Gryc and Lawrence, 2009). Figure 2-4 illustrates the typical workflow of a knowledge-based approach for Sentiment Analysis.

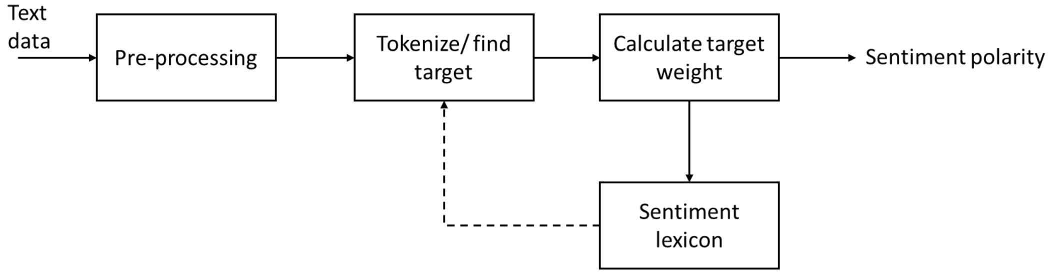


Figure 2-4: A knowledge based approach of Sentiment Analysis.

Numerous studies can be found in the literature in relation to this approach. For instance, Fang et al. (2015) proposed a method for sentiment classification of product reviews on Amazon.com (Fang and Zhan, 2015). This involved extracting sentiment sentences and tagging the extracted words into parts-of-speech followed by phrase identification including negation identification. Then, they computed a sentiment score for the sentiment *'tokens'* which consisted of the positive or negative word and its part-of-speech tag. In their method, they used over 11,000 word tokens and 3,000 phrase tokens. Based on the occurrence of the tokens and the sentiment score computed, they then classified the review into positive, negative or neutral.

Knowledge based approaches use a dictionary for polarity classification and such dictionaries are made available by researchers. For instance, Hu et al.(2004) put forward an approach for classifying user experience by using a list of 2,000 positive words and 4,000 negative words. Gann et al.(2014) analysed twitter data to list close to 7,000 token words and assigned a sentiment score to each token. The sentiment score classifies the token into positive or negative based on a score computation model, which considers the occurrence of the token in positive and negative tweets. Such dictionaries or datasets for polarity classification are widely available and well explored in Sentiment Analysis research (Liu, 2015).

The lexicon or knowledge based approaches have been tested for analysing various user reviews including movie reviews, product reviews on websites, such as Amazon.com and user reviews on social media, such as Twitter in various works, such as (Yu *et al.*, 2013; Sarvabhotla, Pingali and Varma, 2011; Wang *et al.*, 2011; Li, Zhang and Sindhwani, 2009). However, current research is rapidly moving away from knowledge or lexicon based approaches due to their limitations, such as the need to have domain-specific dictionaries for identifying positive and negative sentiments as well as their relatively lower accuracy in sentiment classification compared to the more advanced learning based approaches (Chen and Liu, 2016; Taboada *et al.*, 2011).

2.9.2 Learning based approaches

The advancements in learning based methods, such as machine learning and artificial neural networks, has triggered extensive research. In the recent years, Sentiment Analysis research has involved widely exploring learning based methods, which has been employed in user experience analysis. In Figure 2-5, the typical workflow of learning based approaches for Sentiment Analysis is illustrated.

Machine learning approaches, such as Support Vector Machines (SVM), Naive Bayes and the Artificial Neural Networks (ANN) approaches, have been explored for Sentiment Analysis and polarity classification. For instance, Kim *et al.*(2006) used SVM methods to identify the helpfulness of user reviews on products on Amazon.com. Their research indicated that review features, such as character length, unigrams and product rating are important for identifying the user experience.

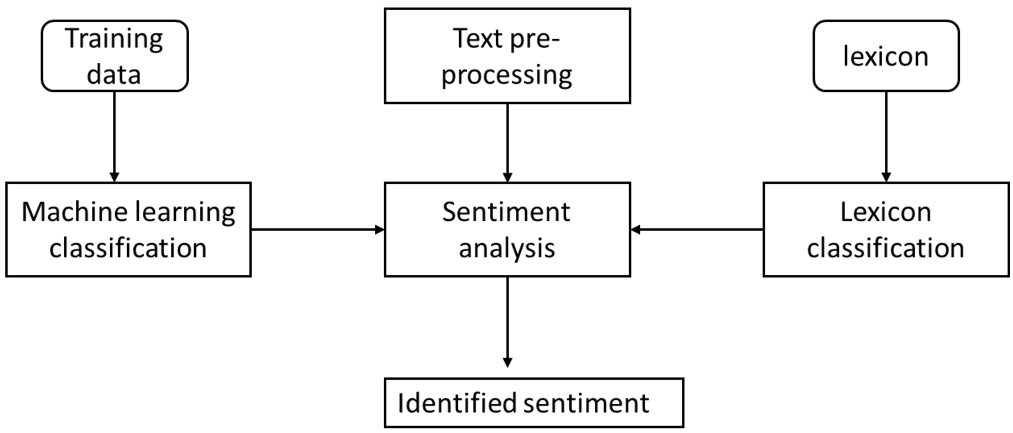


Figure 2-5: Learning based Sentiment Analysis approach

Kennedy and Inkpen (2006) presented Sentiment Analysis of movie reviews using SVM methods and valence shifters that considered words, such as negations, intensifiers, and diminishers. Their results indicated a high accuracy of classification of movie reviews. A similar approach for movie review analysis was proposed in Mullen and Collier (2004). Machine learning approaches have proven to be effective in sentiment classification of user experience and have been examined for various topics, including hotel reviews (Duan *et al.*, 2013), movie reviews (Singh *et al.*, 2013), social media (Ortigosa, Martín and Carro, 2014; Ghiassi, Skinner and Zimbra, 2013; Wöllmer *et al.*, 2013; Lin and Kolcz, 2012) and product reviews, such as games (Ryan *et al.*, 2015), e-cigarettes (Cole-Lewis *et al.*, 2015; Myslín *et al.*, 2013) and treatment drugs (Adrover *et al.*, 2015; Feldman *et al.*, 2015).

The machine learning methods, however, require manual intervention while building the model. At this stage, labels have to be identified for words and phrases to build the training dataset. The model is trained on the training dataset and verified on a test dataset, followed, which it is used on a regular dataset. The labels identified generally are keywords describing a sentiment. However, many reviews, especially in the social media, tend to be noisy with texts written in abbreviations, emoticons, symbols, etc. Recent works have also focused on training Sentiment Analysis models that consider noisy data, such as emoticons for sentiment classification. For instance, Liu *et al.*(2012), proposed a Sentiment Analysis model for Twitter data that considers not just the manually labelled data, but also the emoticons. Sentiment Analysis of other noisy data, such as hashtags, informal texts and figurative languages, have also been presented in the literature (Ghosh *et al.*, 2015; Kalamatianos *et al.*, 2015; Kiritchenko, Zhu and Mohammad, 2014).

2.9.3 Feature based Sentiment Analysis

Recent research on the Sentiment Analysis of user experience has moved away from just polarity classification (positive, negative or neutral) to feature being based, also referred to as fine-grained or multi-grained analysis. Consider the following user feedback for a laptop:

"My new laptop is great, it has an excellent battery life, very fast processor and a large memory, but the monitor size is small and is quite heavy."

In the above user review, the feedback is about multiple features of the laptop. A feature based Sentiment Analysis approach identifies the feature-specific user experience. This approach is useful for product manufacturers and service providers in identifying the best-selling and worst performing features of their product. Feature-based Sentiment Analysis enables other potential customers to choose a particular product based on their priorities. For instance, a customer looking for a laptop with a faster processor and who is not particularly bothered about its weight will be able to identify a suitable laptop based on the recommendations from this approach.

There is huge amount of research interest in feature based Sentiment Analysis. Guzman and Maalej (2014) presented a fine-grained analysis of app reviews from both Android and iOS app store. They were able to identify specific features liked by users for a given app. For instance, their Sentiment Analysis of Dropbox app reviews identified the most reviewed features as upload file, move file, file name, move photo, etc. Similarly, Ganu *et al.*(2009) presented an approach to analyse restaurant reviews and identify user experience for various features, such as food, price, ambience, among others. Similar approaches have been presented for hotel reviews (Duan *et al.*, 2013), usability and user experience features from game reviews (Hedegaard and Simonsen, 2013), audio-visual feature reviews from YouTube videos (Poria *et al.*, 2016), movie reviews (Zhuang, Jing and Zhu, 2006), and other phenomena (Greaves *et al.*, 2013a; Manary *et al.*, 2013; Greaves *et al.*, 2012b; Liu *et al.*, 2009; Collier *et al.*, 2008).

2.10 Sentiment Analysis for Patient Experience

Understanding patient experience is crucial for assessing the quality of healthcare provided. Traditional approaches of collecting patient feedback and understanding patient experience involve direct interactions in the form of interviews and surveys using both qualitative and quantitative research methods. Several works in the literature about patient experience have involved using traditional qualitative methods (Andrew, 2014; Greaves *et al.*, 2013b; Pang and Lee, 2008b). These methods have significant limitations in terms of the high cost, time and effort involved in collecting patient feedback. Further, analysing feedback via these methods is laborious and runs the risk of human error and fatigue in identifying the correct themes from the feedback.

In the present healthcare system, the patient is a consumer and healthcare has become a service. Digital technologies have been widely adopted for healthcare services and collecting patient experience information via digital media is a common occurrence. Hospitals are increasingly providing feedback portals on their websites for patients to provide ratings and reviews about the healthcare service provided. Thus, there is a huge amount of data about patient experience that needs to be explored not just for identifying patient satisfaction with the service, but also to identify features and themes in the patient experience data that can enable the healthcare service providers to adapt and improve the quality of care given.

Few works have explored Sentiment Analysis approaches for capturing patient experience. Regarding which, Alemi *et al.*(2012) and Greaves *et al.*(2012c) presented their work on Sentiment Analysis of patient experience in Maryland, USA. They applied SVM, decision trees, and Naive Bayes approaches of Sentiment Analysis and were able to classify patient comments into complaints/praise attributed to specific staff as well as feedback about other aspects, such as access, wait time, privacy, facilities, etc.

A similar study published in Lees (2011), analysed the patient experience data provided for US hospitals on Twitter. They assessed tweeted reviews for over 2,000 US hospitals and identified those related to patient experience. Their preliminary results showed that the patient experience present on the tweets did not match that regarding the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) ratings. Understanding patient experience is also necessary for pharmaceutical companies and they could employ Sentiment Analysis approaches for collecting patient feedback, as presented in Liu (2010).

Recently, the National Health Service (NHS) has built up a large database of unstructured, free text information about patient feedback and ratings for hospitals as well as their experience. This database provides a large volume of data that could be explored to understand patient experience. Greaves *et al.*(2012b) have presented their work in this direction by applying Sentiment Analysis approaches on patient feedback texts to identify their level of approval and ratings for given NHS hospitals.

However, as aforementioned, research on applying Sentiment Analysis approaches for identifying patient experience is very limited. This is particularly so in the UK context and to the best of this researcher's knowledge, the text mining approaches applied to the NHS patient feedback dataset proposed here have yet to be investigated. That is, there is a research gap in relation to this dataset's exploration and the use of proposed text mining approaches to understand patient experience. In sum, research and development of Sentiment Analysis approaches for patient experience could benefit both hospitals and patients and hence, more effort in this area is required (The Beryl Institute, 2014).

2.11 Topic Modelling

In the Sentiment Analysis approaches described above, the general focus is to identify the sentiment conveyed in the document about a particular product or aspects of it. In contrast, Topic Modelling is a natural language processing approach that focuses on identifying the recurring themes and topics in a given document or a set of documents.

2.11.1 Overview

Topic Modelling is a technique that identifies abstract *"topics"* from a document, which is useful for text categorisation and opinion mining (Rajagopal *et al.*, 2013). It is based on the idea that a document is a mixture of topics and each has a probability distribution over words. For example, a document comparing the performances of several phone devices discusses various features of the phone. Based on the distribution of words, such as resolution, clarity and pixels, in the document it can be ascertained that one of the topics in the document is about the *"display screen"* of the phone. Similarly, based on the occurrences of other words, relevant topics such as battery life, processor speed can be identified from the document.

A Topic Modelling approach calculates the probability estimate of a word for a given topic, $P(w|t)$ and the probability of a topic for a given document, $P(t|d)$, for all the topics and documents analysed (Uys, Du Preez and Uys, 2008). Figure 2-6 below further illustrates the Topic Modelling concept.

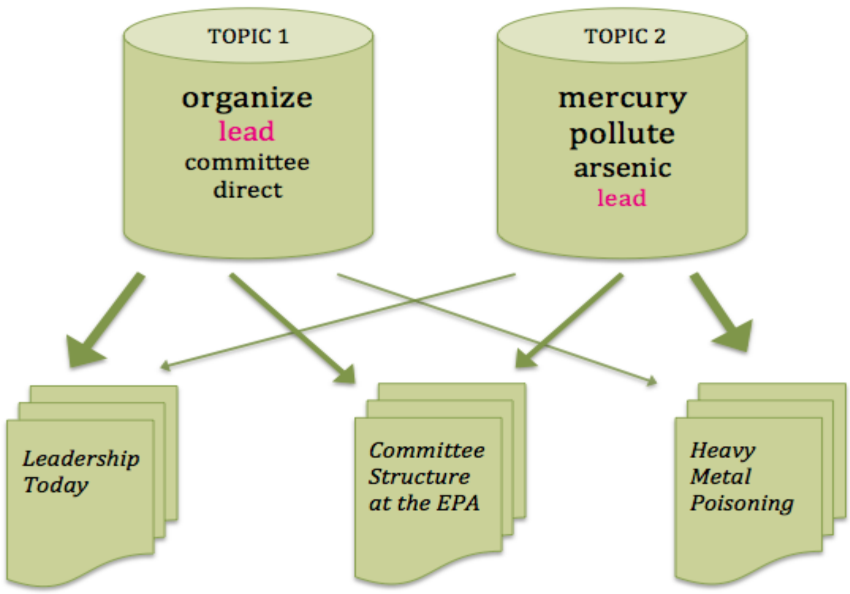


Figure 2-6: An illustration of the Topic Modelling method.

In the above figure, from a collection of documents two topics are identified. As can be seen, each topic is a collection of words that have different probabilities of occurrence in the document passages. The occurrences of the words in topic 1 can be used to identify the *"leadership"* topic discussed in the document, whereas the words from the topic 2 can be identified as a "mercury poisoning" topic. It is also interesting to observe that based on the occurrences of the word 'lead' in the two topics, a topic such as the *"committee structure of a metal manufacturer"*, can be identified. Thus, it should be noted that, Topic Modelling is a backward extrapolation approach that enables the identification of themes or topics from a collection of documents (Underwood, 2012).

As can be seen from the occurrence of the *'lead'* word in the above figure, the occurrences of a certain word or collection of words can be attributed to multiple themes in the document. The challenge is to identify the right topic from the collection of words and current research in Topic Modelling area has focused on improving the accuracy of identifying the right topic from a document using various probabilistic and statistical approaches.

The Topic Modelling approach is highly relevant in the current "Big Data" world, for it has the potential to change the way of doing online searches. It can enable exploring the documents based on various themes and also possibly seeing the associations between different themes in the documents as well as uncover how these have evolved over time. For instance, documents of political history of a given region can be used to identify the change in political discourse of that country over a given period. Moreover, Topic Modelling can be used to annotate, summarise and organise a large database of electronic documents automatically with minimal human intervention (Blei, 2012).

2.11.2 Topic Modelling Approaches

There are several Topic Modelling approaches and most use dimensionality reducing techniques, which aim to represent a document using fewer words. Some of the most popular ones are listed and discussed below and subsequently, other approaches are outlined.

- Probabilistic Latent Semantic Indexing
- Latent Dirichlet Allocation
- Correlated Topic Modelling

Probabilistic Latent Semantic Indexing (pLSI)

The Latent Sematic Indexing (LSI) approach uses linear algebraic approaches, such as singular value decomposition (SVD) and *"bags of words"*, to represent documents. It aims at extracting words that carry similar meanings, i.e. it uses synonyms and polysemy for topic identification. To represent a document in topics, the LSI approach involves using a rectangular matrix of words by coherent passages and each cell in the matrix consists of the number of times the word appears in the passage. Further, the matrix is represented as a vector using techniques such as SVD, such that every passage is a vector whose value is the sum of vectors standing for its component words. Next, similarities between words with other words and passages are computed using methods such as dot products, vector matrices and cosines (Su *et al.*, 2006).

The probabilistic variant of LSI is referred to as pLSI and uses probability and statistics to find generative data models. The pLSI approach models the words in a document as a sample of a mixture model, i.e. it is a sample of set of topics that is represented in the form of multinomial random variables. This approach was first proposed in Feinerer (2015). When given *T* topics, the pLSA approach finds the probability distribution of the words in a topic and further the probability distribution of these topics in a document. The generative process can be represented probabilistically as shown in the Figure 2-7 below (Aletras, 2014).

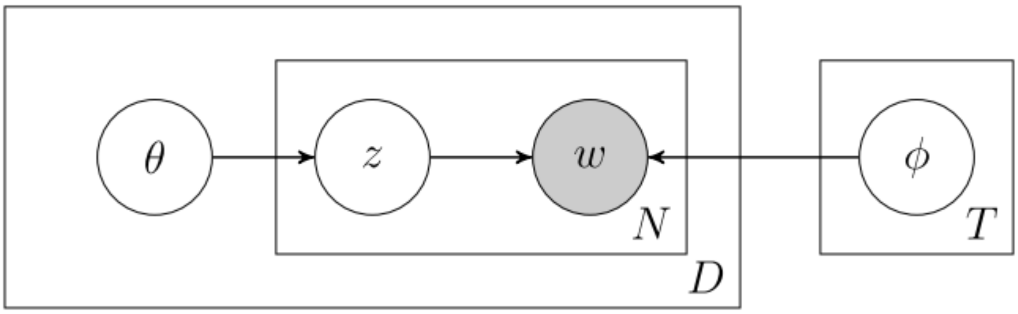


Figure 2-7: Graphical model representation of LSI method.

In the above figure, for a document $d \in D$ with probability$P(\theta_d)$, where θ is a vector of topic proportions and $\emptyset$ a scalar of topics:

Select a topic z with a probability $P(z|d)$

Generate a word w with probability$P(w|z)$.

The pLSA model assumes that each document has multiple topics and the probability of each consists of a weight for a given document. Based on this assumption, the topics in a document are identified.

A disadvantage of the pLSA approach is that the number of parameters of the model increases as the volume of data increases and this could lead to overfitting problems. Further, when the pLSA model is used on documents, which are not part of the training dataset, then the topic probabilities have to be assigned again. In other words, whilst for documents dealing with similar topics the topic probabilities might work effectively, for those relating to dissimilar topics the same topic probabilities might not give effective results.

Latent Dirichlet Allocation (LDA)

The Latent Dirchlet Allocation (LDA) approach is widely used in the literature, which is a generative model of Topic Modelling. It is considered beneficial with large, noisy datasets since this approach does not require the dataset to be annotated. The LDA method is also an unsupervised learning method, i.e. it is able to identify themes and topics from the dataset without supervision or topic selection by the users (Zhai *et al.*, 2012).

The LDA approach adopts the assumption that each topic is a distribution of words and that each document has a certain distribution of topics (Turney, 2001). The LDA approach can be represented at three levels as shown in Figure 2-8. A dataset is a collection of documents and each; the LDA uses Dirichlet distribution sampling to generate a 'topic' distribution. Once a distribution of the topics is generated, then these topics are repeatedly sampled to generate a probability distribution of words for them. Thus, a document can be described as a random mixture of topics, whilst these topics are a probabilistic distribution of words (Liu, 2013).

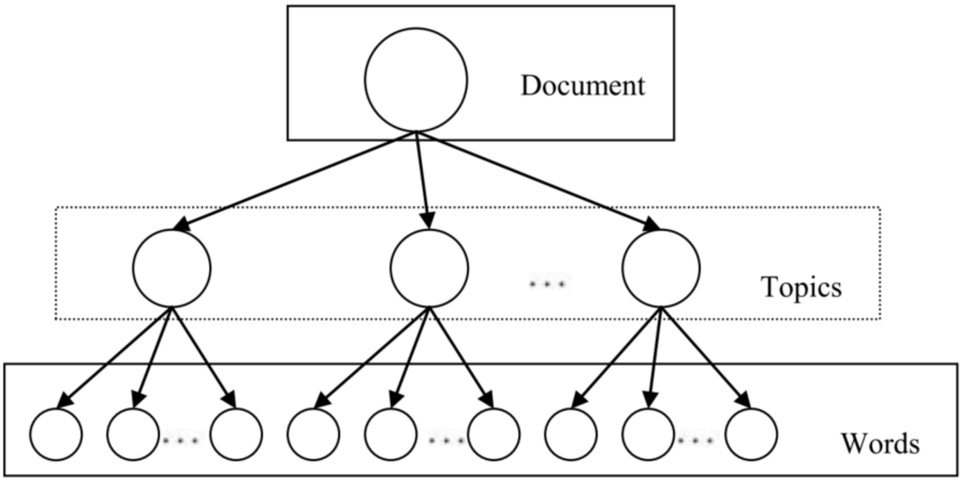


Figure 2-8: A network topology of the LDA approach (Liu, 2013).

In probability statistics terms, the LDA approach is a three-tier Bayesian model for developing a probability generative model to achieve Topic Modelling of the document. This approach is briefly described in mathematical terms and this is facilitated by Figure 2-9.

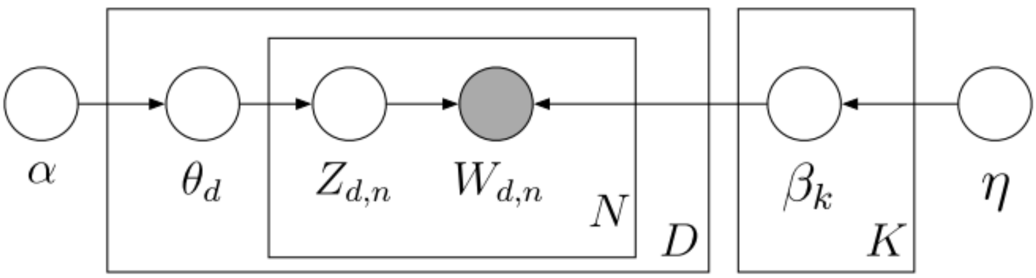


Figure 2-9: A diagrammatic representation of the LDA model

As can be seen in the above figure, there are three levels to the LDA model. The outer box represents the documents and the inner box represents the choice of topics and words in the document. The terms $\alpha\ and\ \beta$ are corpus level variables that are sampled once during the process of generating the corpus where θ is a vector of topic proportions and $\emptyset$ a scalar of topics. The parameter θ is sampled repeatedly from the Dirichlet distribution of each document in the database. The parameters α and β present the relationship between the topics in the document corpus and the probability distribution of the topic, respectively. These two terms are considered as being at the corpus or database level. At the document level, the parameter θ is sampled once per the document and it represents the proportion of the topics present in the document. The parameter z is a word level parameter and it refers to the proportion of the given document assigning the topics to each word in the document, whilst the parameter w represents the word vector of the document. To summarise, for each document, the Dirichlet topic distribution is obtained by the parameter θ based on the parameterα. A word w is generated by the parameter β and its topic z is generated using the topic distribution θ (Blei, 2012).

Further, the LDA approach can involve different variations with respect to the language modelling used. In other words, based on the number of words used for topic identification, the LDA approach can be varied. This is often referred to as n-gram modelling, which can be represented by the following equation:

$$p(w) = \sum_z p(z) \prod_{n=1}^{N} p(w_n|z).$$

One common approach is to use the unigram method, where single words are grouped in order to find a particular topic from the given document (Cheng *et al.*, 2014). Other approaches involve using multiple words. For instance, if$n = 2$, the LDA approach is said to be a bigram model that uses a group of “pairs of words” to define a topic. By extension, the n value can be increased to use n number of words for identifying a topic in a multi-gram arrangement.

One advantage of the LDA approach is that the statistical assumptions its makes for Topic Modelling enable it to uncover sophisticated structure in the texts. For instance, the “bag of words” assumption used in the LDA approach makes it invariant to the order of words in the document. Further, the order of documents in the corpus is also not a criterion for the LDA approach to extract topics from the document. This might not be suitable if the patient experience needs to be analysed longitudinally, i.e. over a period. However, in the current study, since the patient experience analysis does not consider the time factor, the LDA method suits its aims.

With LDA, the topics are distributions of words, which generate observations. The observations identified from the LDA model do not just identify a topic from the document, for each group of observations can be associated with multiple topics, thus providing analysts with multiple recurring topics in the document (Blei, 2012). In sum, the LDA method is suitable for analysing patient experience, because patients generally tend to discuss multiple topics in their feedback and this can potentially identify multiple topics from a document.

Correlated Topic Modelling (CTM)

One of the things LSA is unable to do is to model the correlation between the topics. The correlated Topic Modelling (CTM) approach helps in identifying the correlation between specific topics with others. The correlation information might help the users in identifying links or associations between specific topics from a database with other similar topics. For instance, a researcher searching for a particular topic could get more information about other topics associated with the researcher's topic, thereby equipping him or her with more information (Arora, Ge and Moitra, 2012).

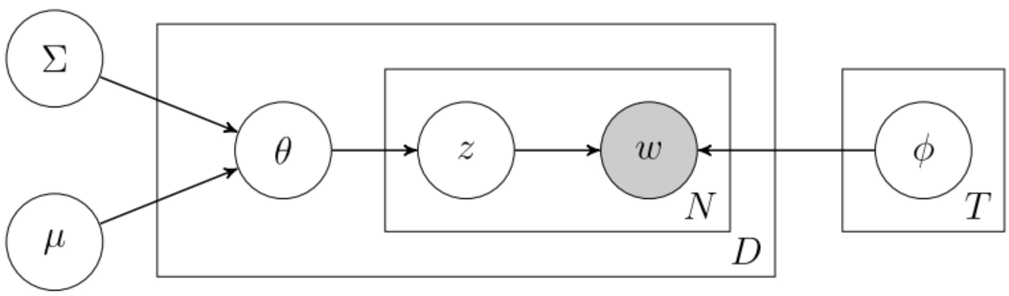


Figure 2-10: Graphical representation of CTM method.

The CTM approach uses logistic normal distribution to identify topics from documents. A covariance matrix used for parameterising the distribution is then used to identify the correlation between the topics. A topic graph is subsequently drawn, in which nodes represent the topics and their correlations with other topics depicted (Blei, 2012). The correlation approach provides more information to the user and thus enables better interpretation of the information. The CTM approach achieves a higher predictive likelihood than the LDA approach (Aletras, 2014).

Other approaches

Several topic-modelling techniques involve different approaches to those above. For instance, dynamic Topic Modelling (DTM) identifies how the topics have evolved over time. It involves dividing the corpus into segments, such as years or days and analysing the evolution of topic distributions to create a hierarchy of topic models (Wang, Blei and Heckerman, 2012). The Pachinko allocation model (PAM) is another type of Topic Modelling in which the various types of correlation between topics, including sparse, nested and arbitrary correlations, are captured. This type of Topic Modelling gives a deeper understanding of the dataset and there are variations to this type of modelling, such as hierarchical PAM (Li and McCallum, 2006). Other types of Topic Modelling are Topic Modelling using Bayesian interference (Purver *et al.*, 2006), and factor based Topic Modelling (Jiang and Saxena, 2013). In sum, with further advances in machine learning techniques and computing power, there is significant research progress being made regarding Topic Modelling approaches.

2.11.3 Topic Modelling for User Experience

In the literature, various studies applying different Topic Modelling approaches are available. These have extensively focused on analysing unstructured information to identify themes and topics from them (Uys, Du Preez and Uys, 2008). They include analysing documents, such as news sources (De Smet and Moens, 2009), natural language text (Lin *et al.*, 2012), data from social media sources (Schinas *et al.*, 2015), corpora documents (Rehurek and Sojka, 2010), government policy documents (Toikka, 2016), amongst others. However, in accordance with the thesis title, a brief literature review of studies applying Topic Modelling to capture user experience is provided below.

Several studies in the literature have applied Topic Modelling approaches to capture the user experience from a database of unstructured text or reviews. Most of their focus can be said to be similar to Sentiment Analysis approaches, i.e. they also aim to identify the user sentiment or user experience for a given product or service. However, Topic Modelling approaches are different because they do not just capture the overall sentiment of the users, for they also identify the user experience with respect to a certain aspect or topic of a given product or service.

For instance, Titov and McDonald (2008), presented a novel framework aimed at extracting rateable aspects of a given product from online user reviews. They presented a Topic Modelling approach that was an extension of the pLSA and LDA and termed it multi-grain LDA (MG-LDA). The proposed approach models two different types of topic, i.e. global and local topics. The global topic part of their model corresponds to identifying the global properties of the objects, such as the brand of a product type, whereas the local topic part of their model corresponds to uncovering user perception on various aspects of the product. The performance analysis results of their proposed model with other approaches show that they were able to capture more aspects of an object from the reviews and thus, were more representative of rateable aspects of an object.

Li et al. proposed a Topic Modelling approach called the *"Supervised User-Item Based Topic Model (SUIT)"* that incorporates user and product information into the Topic Modelling approach so as to identify the user sentiment (Li *et al.*, 2014). Their approach simultaneously utilised the textual topic and latent user-item factors to model the sentiment generalisation. The results of their studies showed that they were able to achieve better performance than other supervised Topic Modelling approaches.

Some commonly found studies on Topic Modelling focus on identifying topics and factors from movie reviews. The reason for the popularity of using movie reviews for NLP is that the review texts offer a wealth of information beyond ratings and thus, are good test datasets. Diao *et al.*(2014) proposed an approach that uses Topic Modelling and collaborative filtering to analyse movie reviews. This captured the interest distribution of the users and the content distribution for movies and thus, provided an aspect based Sentiment Analysis of movie reviews. Their approach was unsupervised and did not require aspect specific ratings or genres for analysis. Similar studies of applying Topic Modelling for movie reviews can be found in works such as that of Jakob *et al.* (2009).

There has been considerable amount of research interest in using Topic Modelling for identifying user experience. Topic Modelling studies have been shown to be useful in analysing user experience for products and services, such as hotel reviews (Rossetti, Stella and Zanker, 2016; Höpken *et al.*, 2013), electronic devices (Christidis and Mentzas, 2013), financial services (Laukkanen, 2006), pharmaceuticals (KUBRA, KHAN and FARHA, 2015) and product reviews on e-commerce websites (Kiran and Kumar, 2015; Xu *et al.*, 2011), banks (Kolyshkina, Levin and Goldsworthy, 2013), and other products and services. In general, it can be said that Topic Modelling for user experience focuses on identifying various aspects and factors of a given product or service review. The difference in the Topic Modelling approach from factor based Sentiment Analysis is that the former can identify the topics on the go from the user reviews without the requirement of predefined topics. In other words, Topic Modelling approaches can be unsupervised learning and can be more representative of various aspects of user experience from a given unstructured information text.

Considering the success of Topic Modelling in analysing user sentiments for products and services, it can also be applied to effectively capture the patient experience. Purver *et al*.(2006) used it with the LDA approach to capture aspects of doctor-patient communication in therapy including predicting symptoms and therapeutic relationships. The outcomes of their study showed that LDA was not effective in predicting the symptoms of schizophrenia therapy. However, it was successful in predicting the therapeutic relationship between the doctor and the patients.

Apart from the above stated work, there are no other works that have applied Topic Modelling to capture patient experience and hence, this is a significant research gap that needs to be addressed. Topic Modelling can be useful in performing effective aspect based analysis of patient experience so that the outcomes of such analysis can be used to improve patient care and hospital services. This can be particularly useful in public healthcare services, such as the NHS, for retaining the patients with local GPs by continually improving the service.

2.12 NLP Dependency Parsing

2.12.1 Overview

Natural languages are complex and difficult to analyse automatically as machines are not be able to identify the correct meaning of a given sentence from the numerous possibilities it can potentially have. The grammatical structure of languages is often complicated to understand as consideration of such matters as context, intentions, sarcasm needs to be given. In recent years, to overcome this challenge, Dependency Parsing techniques have been employed (Green, 2011).

Dependency Parsing enables the analysis of the grammatical structure of sentences and identifying the relationships between different words in them. The principle behind the *“dependency”* is that a sentence is such that the syntactic structure has binary asymmetrical relations between the words in the sentence. In NLP Dependency Parsing, a dependency relation is established in a sentence by defining entities named as the *head* and a *dependent* (Nivre and McDonald, 2008). Then, based on the relationship between the words in terms of grammatical structure, it identifies the subject, object, verb, nouns, and any other syntax in the sentence. Based on the identified grammatical structure parse trees, termed dependency trees, are created to analyse the sentence structure. These then act as decision trees for analysing and then, making decisions about the sentence structure (Koo, Carreras Pérez and Collins, 2008). Dependency parsers help in achieving a more accurate analysis of free word texts. There is a significant amount of research interest in NLP Dependency Parsing and the availability of datasets in various languages has helped fuel progress on this matter. The CoNLL group`s shared task body provides datasets that can be used to train Dependency Parsing algorithms (Buchholz and Marsi, 2006).

Dependency parser is different from another parsing method namely, constituency parser. This method breaks the text into sub-phrases. A given sentence typically gets separated into a verb and noun phrase. The dependency parser, on the other hand, connects the words in a given sentence according to relationships. The constituency parser is useful in analysing the sentences into subphrases. The dependency parser is useful to find relationships between words.

There are several techniques for Dependency Parsing, which are quite complex to explain and in any cases, are generally beyond the scope of this thesis. Thus, the two most commonly used Dependency Parsing techniques are only briefly discussed below.

Graph-based: This is a widely used technique and is based on graph theory. This technique uses a system of scoring the parsing decisions on a whole-tree basis. Typically, the graph-based approaches segregate the dependency tree into subgraphs with sizes varying from one node to multiple nodes and edges and each subgraph is scored. The scores are then summed up for the entire tree and that with highest score for a given sentence is made valid (Pei, Ge and Chang, 2015).

Transition-based: This technique anchors on each word in a sentence and applies a dependency link between the other words in the sentence by a series of right and left shift actions (Zhang and Clark, 2008). In each transition, it creates a stack and buffer for storing partially processed nodes and the remaining inputs. Scoring functions are then applied on the transitions to make decisions about the sentence structure. The scoring function applied on transitions are memory based learning or machine learning approaches, such as support vector machines (Nivre and McDonald, 2008).

2.12.2 Dependency Parsing for User Experience

Similar to other NLP techniques, Sentiment Analysis and Topic Modelling, discussed in the above subsections, Dependency Parsing has also been extensively studied for capturing user experience. For instance, Wu *et al.*(2009) introduced a phrase Dependency Parsing approach which involved parsing at a phrase level to identify relations between product features and customer opinions. Similarly, Somprasertsri and Lalitrojwong (2010) applied a combination of Dependency Parsing and ontological knowledge with a probabilistic based model, to execute feature based opinion mining from user product review effectively.

Mukherjee *et a l.*(2012) presented their study on applying Dependency Parsing to identify feature specific expressions of opinions in product reviews with different features and mixed emotions. Their designed system learns the set of significant relations to be used by Dependency Parsing and a threshold parameter that allows for the merging of closely associated opinion expressions. Their approach is a graph based method and was able to achieve a high level of accuracy for the product reviews they analysed in their study (Mukherjee and Bhattacharyya, 2012).

Various other studies have explored Dependency Parsing to capture user experience from product reviews (Agarwal *et al.*, 2015; Kumar and Raghuveer, 2013; Di Caro and Grella, 2013; Liu *et al.*, 2013; Joshi and Penstein-Rosé, 2009). Further, in the literature search it was also observed that there is a lack of research work studying the Dependency Parsing approach in capturing the patient experience. Thus, it is important to explore dependency parsing method`s effectiveness in analysing patient experience.

2.13 Challenges of Sentiment Analysis

Other text mining approaches in natural language processing identify specific *"themes"* in a given chunk of text and classify the information based on those identified. In Sentiment Analysis, whilst binary classification into positive or negative sentiment is the most popular application, achieving this simple classification is difficult due to the numerous challenges in identifying the true sentiment. This is because the construct of human emotions in text can be very complex and can have virtually infinitesimal permutations and combinations in relation to expressing sentiments. Some of the challenges for Sentiment Analysis are discussed in this section; however, this is not a comprehensive list but rather a selection of the most common challenges of Sentiment Analysis.

2.13.1 Target Entity Identification

One of the important issues in Sentiment Analysis or opinion mining is identifying the target entity. A piece of text can have multiple target entities, which is a common phenomenon in review analysis (Liu, 2015; Liu, 2010). For instance, a sentence such as:

"My new laptop is great; it has excellent battery life, very fast processor and a large memory"

In this single sentence, several features of the laptop are described and the target entities are multiple. The targets in the sentence include laptop, battery, memory and processor. The sentiment in the above sentence is clearly positive; however, the challenge is to identify to which component the specific positive sentiments belong. The following comment illustrates the issue of allocating sentiment to multiple parties in the same piece of text:

"…to describe the performances of the actor and actress awesome and awful, respectively…"

In the above sentence, it can be challenging to associate the appropriate opinion to the targets (actor and actress) as the analysis model needs to consider the word "respectively" and realise its significance in this context.

There have been efforts that have focused on addressing the target entity identification issue and the approaches are aimed at finding this are usually referred to as fine-grained sentiment analyses. Various studies can be found in the literature on the fins-grained analysis of text corpus. For instance, summarisation of twitter data using Sentiment Analysis methods that focused on identifying target entities was studied in Agarwal *et al*.(2011). Despite significant progress in this area, the challenge is achieving high accuracy of target entity identification.

2.13.2 Co-reference Resolution

The co-reference resolution is a well-known challenge for natural language processing. It refers to the problem of deciding which target entities in the text are co-referred. The usage of words such as *"it"* in a sentence can make it difficult to identify the noun referred to by *"it"*. For instance, in the above sentence example of the actor and actress, it is difficult to identify the target entities for each sentiment. This is an important challenge and is a topic of significant research interest (Varghese and Jayasree, 2013).

Approaches, such as supervised classification and clustering, are applied to train the analysis models to identify the co-reference. Generally, the supervised classification approaches use pairwise function to predict noun phrases that have co-reference. This is followed by applying clustering methods to produce the final cluster. The trained model is then used on the piece of text to identify the co-reference. Several studies have worked towards addressing these issues and made some progress. The major challenges pertaining to co-reference resolution can be summed up as achieving high accuracy, the availability of large datasets for training the models and the portability of the trained models from one context to another (Stoyanov and Cardie, 2006).

2.13.3 Identifying sarcasm, abbreviations in text

Another challenge in Sentiment Analysis is to identify sarcasm in the text. For instance, if the opinion holder is sarcastic in his opinion, then such subjectivity is very unlikely to be identified by the Sentiment Analysis model and may wrongly interpret the sentiment.

Further, some sentiments might appear as negative, but can on closer look be interpreted as being positive (Maynard and Greenwood, 2014). This is apparent in:

"... the wait for the next Star Wars movie is killing me..."

The above text may possibly be classified as negative by the classifier owing to the presence of the word 'killing'. However, it is clear that there is positive sentiment, whereby the opinion holder's use of the word *'killing'* means that he is very eager for the next Star Wars movie to arrive, given how it is used semantically.

Another challenge for sentiment analysis methods include the use of popular jargons, such as *'cool', 'what`s up', 'yo'* as these jargons do not usually have a specific context and would be difficult for the models to analyse into a sentiment. Further, usage of short words on social media such as *'ur', 'hru', 'cos'* and weak spelling make Sentiment Analysis challenging (Nakov *et al.*, 2016).

2.14 Summary

This literature review aimed to provide a detailed insight and critical analysis into relevant literature in the research area. A further aim was to give readers awareness of the problems and limitations of the various aspects and characteristics of Design Science Research (DSR). A point to note, when analysing "patient experience", this was equated to that of "user experience". The reasoning behind this was because patient experience is essentially users providing their feedback about the service they have received during their time as a patient. This statement can be considered as the definition of user experience where the user in this case is the patient.

This chapter has provided specific analysis into natural language processing methods for analysing patient experience data, namely, Sentiment Analysis and Topic Modelling, with a brief introduction to Dependency Parsing. The chapter also introduced the concept of patient experience beforehand in order to guide the reader into text mining and language processing methods. This therefore made conducting this literature review quite difficult in the sense that many areas had little literature to analyse and elaborate upon. However, this also made clear the fact that this was clearly an important research gap and one that this research will aim to address with new knowledge, further highlighting the novelty of this research.

As mentioned above, this research identified and shortlisted three potential natural language processing methods as solutions to address the research problem at hand. The first of these methods was Sentiment Analysis. Sentiment Analysis allowed for the identification of patient sentiment automatically from the data. The reason behind why this method was selected as a potential solution was because of its ability to give healthcare professionals a quick outline and overview of the general patient perception and feelings towards the level of service provided by the service provider i.e. hospital. Sentiment Analysis also had another benefit in that it allowed these healthcare professionals to comprehend the patient satisfaction rate they receive from patients and measure/compare this against their performance both as a provider and as an individual.

Topic Modelling approach was the next natural language process that was analysed. This was selected as an alternative to Sentiment Analysis, as Sentiment Analysis provides shallower understanding of patient satisfaction rates. Compared to this, Topic Modelling has the ability to identify automatically the topics that patients made the most comments on in their feedback. This is a great advantage to healthcare professionals, who could use this technique to understand areas that they are doing well in, but most importantly, identify those that require much improvement in order to enhance patient satisfaction rates in future feedback requests.

The third and final natural language processing method briefly analysed, as a potential solution was Dependency Parsing. This method was selected because of its ability to allow healthcare professionals to identify relations/associations between phrases within the feedbacks provided by patients which will further help to summarise patients' feedback and understand the reasons behind patients' sentiments.

In summary, this chapter identified the research gap in the area of patient experience analysis. In the literature research, it was identified that several recent NLP methods such as sentiment analysis, topic modelling, and dependency parsing have not been well explored in the context of patient experience text analysis. Healthcare is an important service sector. Therefore, it is necessary that research in NLP also explore the healthcare information systems to design and develop automated systems that allow analysis of text data such as patient feedback and reviews.

Chapter 3: Research Design and Approach

3.1 Overview

For successful research, a systematic approach is essential for obtaining reliable and reproducible outcomes. In this chapter, the research approach followed during the course of the production of this thesis is discussed. The Design Science Research (DSR) methodology is used as the general research methodology for this thesis. A detailed discussion of this method and justification for its adoption are presented in the following sections. Section 3.2 introduces the various types of research approaches in relation to information systems. Section 3.3 discusses, in detail, the DSR paradigm. In section 3.4, a DSR based framework for patient experience analysis is presented.

3.2 Research Approaches to Information Systems

Research in the area of Information Systems (IS) has gained significant attention and growth in the last couple of decades, due to the rapid adoption of such systems across various disciplines. Generally, a given research endeavour is underpinned by a research paradigm and an appropriate methodology. A research paradigm refers to the underlying philosophy behind a research activity that guides the research process. The research methodology is a set of processes or procedures that needs to be carried out to conduct research. It includes the various phases, methods, techniques and tools required to design, develop, implement the research as well as to analyse the outcomes (Mingers, Mutch and Willcocks, 2013).

Various types of research paradigms are extensively used in IS research. Two commonly used methods are the Behavioural Science Research and Design Science Research. The former includes approaches such as "Positivist" and "Interpretive". The latter includes several models proposed by different authors. A brief summary of the positivist, interpretive and design science research is presented in **Error! Not a valid bookmark self-reference.**and briefly explained below.

Table 3-1: Popular research paradigms in IS research.

Research Approach	Summary
Positivist	▪ Hypothesis created ▪ Relies on factual data ▪ Only objective measures accepted ▪ Outcomes must be reproducible
Interpretive	▪ No hypothesis created ▪ Collecting predominantly subjective data ▪ Subjective interpretations of data ▪ Multiple interpretations can be obtained
Design Science Research	▪ Problem solving paradigm ▪ Collects knowledge ▪ Develops artefacts to solve a problem ▪ Evaluates the artefact performance

3.2.1 Behavioural Science Research

A popular type of research approach is behavioural science research. It is most commonly used in natural sciences research often to describe or predict a phenomenon based on the inputs of human behaviour. The research outcomes obtained using these approaches are subsequently employed to develop or evaluate theories. The general approach used in behavioural sciences is to generate an initial set of research ideas based on personal experience of humans or from existing literature. The next steps include the procedure design, followed by observations of the collected data. The data obtained from the observations then undergo analysis and interpretation. The research outcomes obtained from this process are subsequently provided for other researchers in the field to stimulate further relevant inquiry (Bhattacherjee, 2012). Behavioural science research includes several methods. Two of the commonly used ones, Positivism and Interpretive, are briefly discussed below.

Positivism Research

In the positivist research approach, the underlying philosophy is that the research outcomes are supported by data collection to support hypotheses or assumptions made prior to the start of the research process. This approach adheres to the principle that only factual data or knowledge that is gained through observations and measurements alone is trustworthy. In other words, this perspective is completely dependent on the objective measurements only. Here, the role of the researcher is solely limited to data collection and interpretation via the objectives approach. The key aspect of positivist research is that it must be reproducible, i.e. the observations or objective measurements obtained must be repeatable. Moreover, this approach often involves modifying independent variables to observe its impact and then use it to identify a pattern, based on which predictions can be made. The positivist research approach is widely used in various fields other than IS. In particular, it is by far the most preferred approach in the physical and natural sciences areas of research (Cohen, Manion and Morrison, 2013).

Interpretive Research

As explained above, under positivism the collected data are used to support an assumption or hypothesis. In the interpretive research approach, there is generally no initial assumption or hypothesis and the collected data are used for building up the researcher's knowledge. In this sense, the underlying philosophy in interpretive research is subjective rather than objective. Under this approach, the phenomenon is studied in its natural environment and the researcher might have an impact on it while investigating it. The interpretive approach to research can generate multiple views of the phenomenon and all of these illuminate how people understand their world (Schwartz-Shea and Yanow, 2013).

3.2.2 Design Science Research

The Design Science Research (DSR) paradigm is a widely popular research approach in IS research. It is referred to as a problem-solving paradigm, which focuses on building "artefacts" that are aimed at addressing a problem. Moreover, the aim of DSR is to design and develop those *artefacts* that are aimed at solving a given problem and can be described as purposeful, innovative and novel. The artefacts address the problems or enhance existing solutions and thus, are important tools for arriving at research outcomes and reviewing these to decide how the artefact adopted can be further utilised (Peffers *et al.*, 2006). The DSR process follows a systematic procedure in which the artefacts are developed with systematic creation, capturing, and communication of knowledge from the design process. Significantly, DSR also uses an iterative process, whereby the artefacts are reconstructed during each iteration and thus, can be described as a continuous learning process that enhances the artefacts quality incrementally (Gregor and Hevner, 2013).

The aim of the research presented in this thesis can be described as developing an artefact as a framework for automated analysis of patient feedback about healthcare services provided by hospitals. The DSR approach enables the development of an artefact that can be used as a tool to solve an existing problem. In the current case, the problem is to achieve an automated analysis of patient feedback to identify their sentiment and opinions about the healthcare service. This framework not only enables addressing the problem, for it can also be further used as a tool to enhance the healthcare service provided based on the analysis of the patient feedback.

As pointed out above, the DSR is an iterative approach that makes the artefact a continuous learning model. The patient feedback analysis carried out in the present study is also an iterative process in that for each successive iteration the patient feedback is more deeply analysed based on the outcomes obtained from the preceding ones. This iterative approach facilitates the development of a strong, effective automated patient feedback analysis system. In the following sections, the DSR paradigm is discussed in detail, followed by a description of the implementation of DSR technique deployed for developing the framework for the study.

3.3 The Design Science Research (DSR) Paradigm

3.3.1 Overview

In Information Systems (IS) research, DSR methods are being widely adopted. The DSR paradigm is a set of methods that focuses on the development of an artefact that aims to improve the performance of the given artefact. In DSR researchs, the artefacts are designed and developed to perform in the context of a given application and are often influenced by resource availability and constraints of the given application domain (Vaishnavi and Kuechler, 2015). Additional research, in terms of iterative improvements, enables generalising of the research outcomes to broader domains. Thus, the DSR approach seeks to achieve research outcomes that improve the artefact performance in a specific domain as well as intends to obtain a broader theoretical understanding, thereby making it applicable to different domains.

3.3.2 DSR Cycles

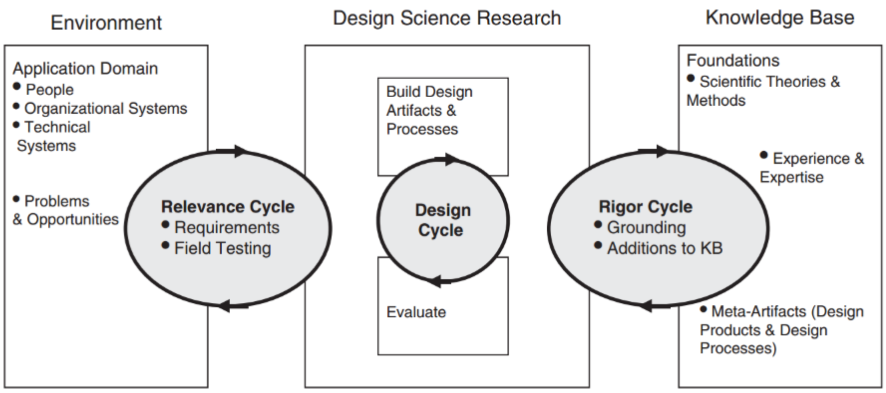


Figure 3-1: Design Science Research Cycles (Hevner, 2007)

According to Hevner (2007), DSR has three inherent research cycles. The relevance cycle bridges the contextual environment of the application with the design science activities. The rigor cycle connects the knowledge base that includes foundations of theories and methods relevant to the application to the design science. The design cycle refers to the iterative process of building and evaluating the design artefacts and processes of the research (Hevner, 2007).

3.3.3 DSR Steps

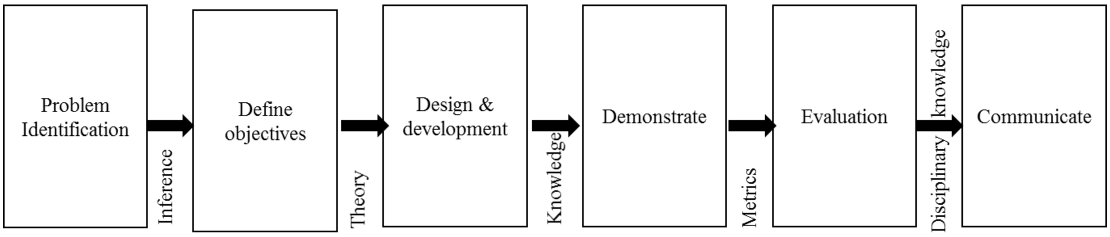


Figure 3-2: Design Science Research Steps (Peffers et al. 2008).

Work by Peffers *et al.(2008)* and Peffers *et al* (2006) chooses the more traditional path of steps in comparison to Hevner's cycles as shown in Figure 3-2 above. Their model consists of six steps: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. All steps are consistent with prior literature in the field, provide a process model for doing DSR, and provide a mental model for presenting and evaluating DSR.

3.3.4 Knowledge Base

In Figure 3-1 above, it can be seen that building a knowledge base is an important part of DSR, for it provides the materials from which the artefacts are constructed and evaluated. It comprises two important parts: *foundations* and research *methodologies*. The *foundations* determine the type of knowledge outputs that are produced in the given application domain for the design science. Vaishnavi *et al*.(2015) have proposed four outputs that determine the type of knowledge that can be derived from DSR, these types are listed in

Table 3-2.

Table 3-2: Types of knowledge outputs in DSR

Knowledge Output Type
Constructs
Models
Methods
Instantiations

Construct: It is a set of concepts that form the vocabulary of a given application that shapes the knowledge to the given problem and to providing solutions.

Models: The models use the constructs to define a set of statements or propositions that describes the relationship between the constructs.

Method: This type provides a set of guidelines that is used to perform tasks using the constructs and the models. They describe the steps that need to be followed to obtain solutions using the constructs and models.

Instantiation: Based on the constructs, models and methods, instantiation is the realisation of an artefact in its environment. It enables the operationalisation of constructs, models, and methods to obtain the artefact.

3.3.5 DSR Methodology

Vaishnavi & Kuechler developed a methodology for DSR, which is different from the methodology developed by Hevner (2007) and Peffers *et al.*(2008). The difference in this model is that Vaishnavi and Kuechler put more emphasis on the process of contributing to knowledge. The DSR process here is based on the knowledge built and comprises five main stages, as discussed below as shown in Figure 3-3.

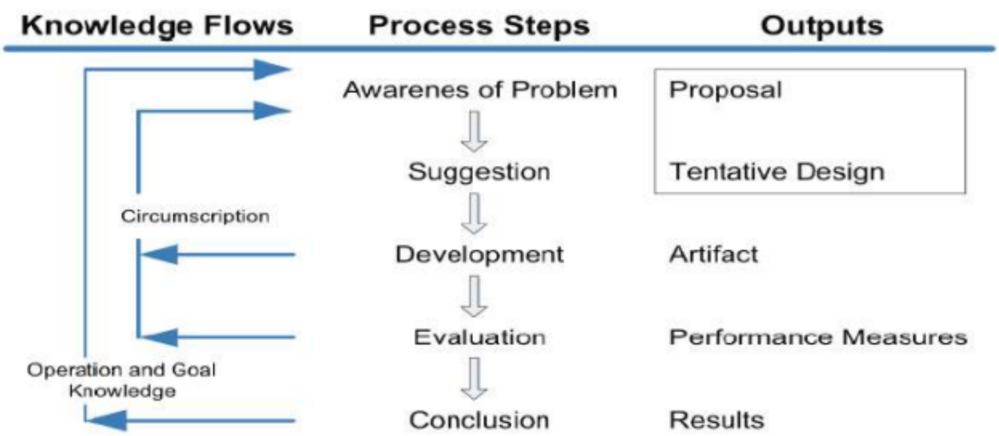


Figure 3-3: Design Science Research Phases (Kuechler, Park and Vaishnavi, 2009)

- **Awareness of the Problem -** This is the first phase in DSR. The awareness of problem aims at understanding it in the context of the application. This can be achieved from multiple sources, such as industry knowledge or from literature research, which enables identifying the research problem and articulating it. The outcome of this process leads to the development of the *proposal* of the research.
- **Suggestion -** Once the proposal is made from the previous phase, the potential solutions that can help in addressing the given research problem is explored. Various insights into the application domain are obtained during this phase and the specifications for the solutions are acquired. This leads to the development of a *tentative design* that represents suggested solutions for the research problem from the proposal.
- **Development -** This very important stage leads to the development of the first *artefact.* In fact, this stage forms the core of the DSR process. The artefact type can be any of the four types of outputs (constructs, models, methods, and instantiations).
- **Evaluation -** Once the artefact is developed, it is important that it be evaluated. These artefacts will be analysed according to a set of specifications that are identified during the suggestions phase. If the evaluation outcomes of the artefact are not satisfactory, then the process cycle is repeated such that the awareness of the problem and suggestions phase is revisited to enable a better development of the artefact. Consequently, DSR is an iterative process that leads to the development of an artefact that satisfies the requirements of the application.
- **Conclusion -** This is the final stage of the DSR process where the outcomes of the previous stages are disseminated to a wider audience. The conclusion involves summarising the knowledge gained in the DSR process and the lessons learnt that could be used for further research in the area. These can also be extended to other application domains to achieve

similar solutions for different research problems. The knowledge gained is often documented and presented to an appropriate audience.

3.3.6 Comparison of DSR Frameworks

At present, a number of DSR frameworks exist in literature that divide the design process into several phases and or cycles as discussed above, by defining a set of milestones to achieve within the design process (Peffers *et al.*, 2008). Table 3-3 compares the works of Vaishnavi and Kuechler, Hevner and Peffers et al. It is important to note that in performing this comparison and wider research into this field, it is quite evident that most research in this area of interest seem to focus greatly on which phases are essential for a DSR process but lack more deeper analysis into what activities are required to be performed at these different stages of the process.

Table 3-3: Comparison of DSR frameworks

Model	Problem identification and motivation	Objectives of a solution	Design and development	Demonstration
Hevner (2007)	Relevant/ Important Problems	Constructs and models.	Build/Develop	Evaluate/Justify
Peffers et al., (2008)	Problem Identification and Motivation	Suggestion	Demonstration	Evaluation
Vaishnavi and	Awareness of the	Suggestion	Development	Evaluation

Kuechler (2007)	Problem			

This table highlights key process elements/activities and compares the most common elements/activities found in each model. From this table, it can be seen that all three models accept defining the research model at the beginning of the DSR process as well as evaluating the design artefact at the end of the process. There are also some similarities between the models in the design and development aspect; however, a point to note here is that each model has further subdivided activities within this stage. The differences arise during the ‘Objectives of a solution’ stage. Hevner`s model uses selecting methods, building models and constructing to reach the objectives of the solution, whereas Vaishnavi and Kuechler`s model focuses on providing a suggestion as a solution that could be further used to development and evaluation.

3.4 DSR based Framework for Patient Experience Analysis

Based on the DSR paradigm, the aim is to address the research question, i.e. to provide an automated analysis of patient experience data via text mining methods. To achieve this, the five-phase design process steps mentioned in Figure 3-3 is adapted and used as the DSR approach for the study. This methodology is chosen because it enables a systematic approach to develop the framework. Moreover, the research to be conducted is carried out in three iterations, namely: sentiment analysis, topic modelling, and dependency parsing. The three iterations schema is presented in Figure 3-4. This DSR approach enables the identification of the problem for each iteration, finding a solution, developing and evaluating the performance of the solution methodically and hence, is suitable for the research goals.

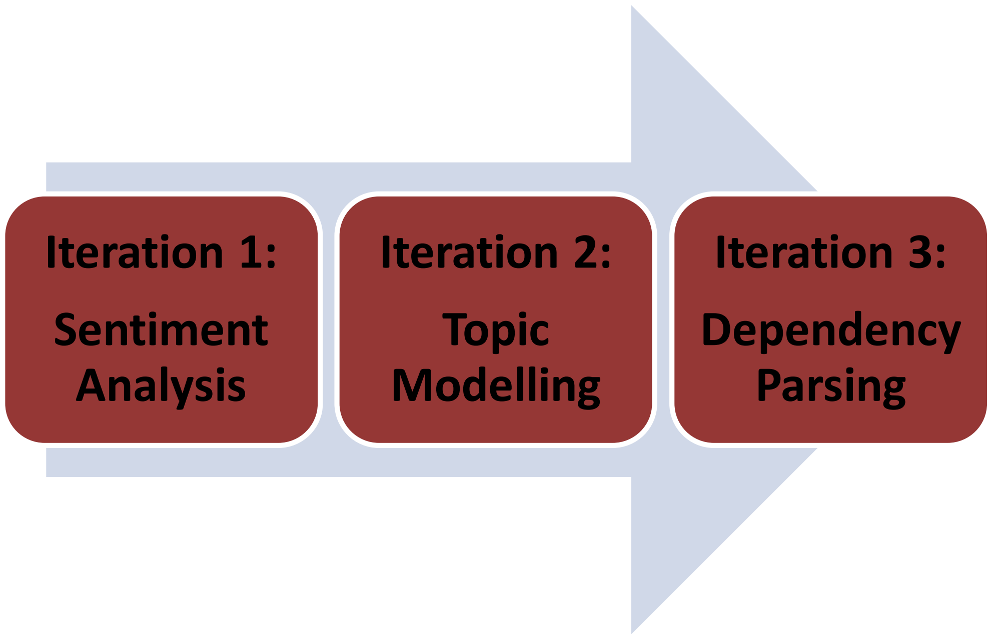


Figure 3-4: Schema of three iterations of the research presented in the thesis

The study presented in this thesis is carried out over three iterations. The first involves applying Sentiment Analysis models to identify the patient sentiment in the feedback data. For the second iteration, a Topic Modelling approach is applied to the patient feedback to identify themes and its associated sentiment from the sentiment identified from the first iteration. The third iteration pertains to applying natural language processing on each topic identified for every patient's feedback and finding the noun-adjective pair that summarises the sentiment for that topic in the given comment.

For each iteration in the study, a common DSR framework is applied and every successive one is dependent on the outcome of that preceding it. Each of the five steps will be followed for each iteration of the study. It should be noted that after the evaluation stage, there is a provision to restart the process depending on the outcomes of the evaluation stage. This process will be applied for each iteration of the study. In the following subsections, the implementation of this framework for each iteration is explained.

3.4.1 Iteration One: Sentiment Analysis of Patient Feedback Data

The first iteration of the study presented in this thesis is to apply Sentiment Analysis (SA) approaches to the *NHS Choices* patient feedback database. The aim of this iteration is to find the sentiment of the patient feedback in their comments provided on NHS hospitals on the NHS website. The sentiment identification is performed by the application of Sentiment Analysis approaches that analyse and classify a comment into positive or negative sentiment based on its contents. In Figure 3-5, the research approach adopted for this iteration is presented.

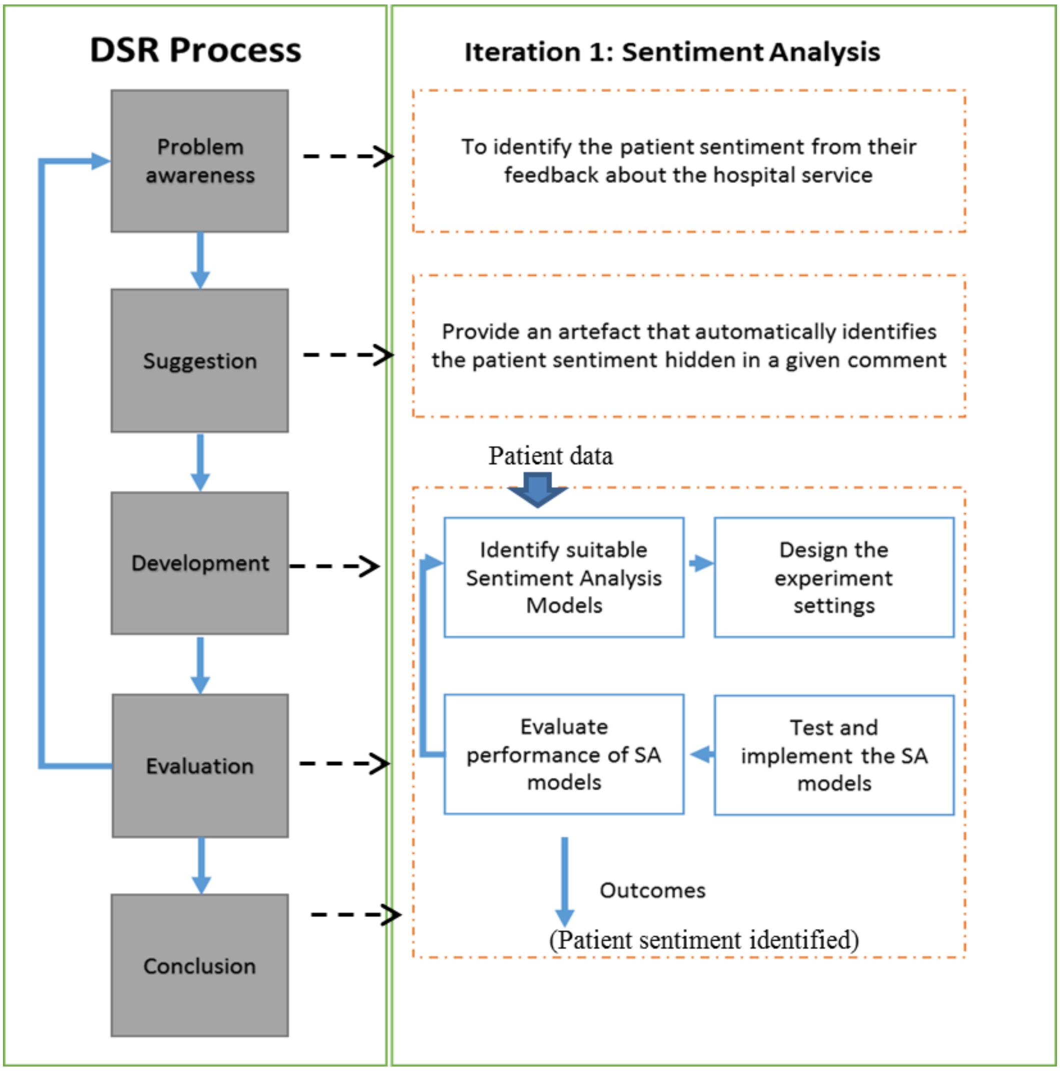


Figure 3-5: Design Science Research approach for the first iteration

The above figure shows the framework methodology for the first iteration of the study. The left part of the figure provides a schema of the DSR approach used and the right part of the figure maps the DSR stages to the stages of our research. The black dotted lines indicate the mapping of the DSR stages to the research steps. Here it can be seen that the problem is the need to identify the patient sentiment from the comments obtained for each NHS hospital from the NHS Choices website. The suggestion is to develop an artefact that can automatically identify the patient feedback or sentiment in the comment as positive or negative. To develop this artefact, there are several sub-steps. The first is to identify suitable Sentiment Analysis (SA) models from the literature research that can identify the sentiment in the comment. The next sub-step is to design the experiment settings where the SA models are implemented on the feedback database. This involves identifying the processes for cleaning and preparing the database for the analysis. This stage also involves identifying the tools and technology required to implement the SA models. In this case, to achieve implementation of the SA models and other computation, the *RStudio* environment is used where the implementation is mainly carried out in the R language environment.

The next step is to evaluate the outcomes of the SA model implementation. The evaluation is an analytical approach in this iteration, whereby it is focused on the accuracy of the SA model in identifying the correct sentiment in the patient feedback. If the sentiment identified by the SA model matches the ground truth sentiment, then the SA model is said to be accurate. The ground truth would be the associated ratings provided by the patient while providing the feedback. The ratings are classified into positive and negative rating and this classification is used to verify whether the sentiment identified by the model from the text matches the rating provided by the patient (this is explained in detail in Chapter 4:). The accuracy of the SA model is evaluated based on that which it achieves for the entire *NHS Choices* patient feedback database chosen for the study. The outcomes will be disseminated in the form of reports and publications. Further, they will also be used for the second iteration of the study presented in the thesis.

3.4.2 Iteration Two: Topic Identification from Patient Feedback

The next iteration of the study is to identify topics in the given patient feedback. The aim is to understand what specific area of the healthcare service the patient has discussed in their feedback, such as the maternity department, parking facilities or the waiting period in the hospital.

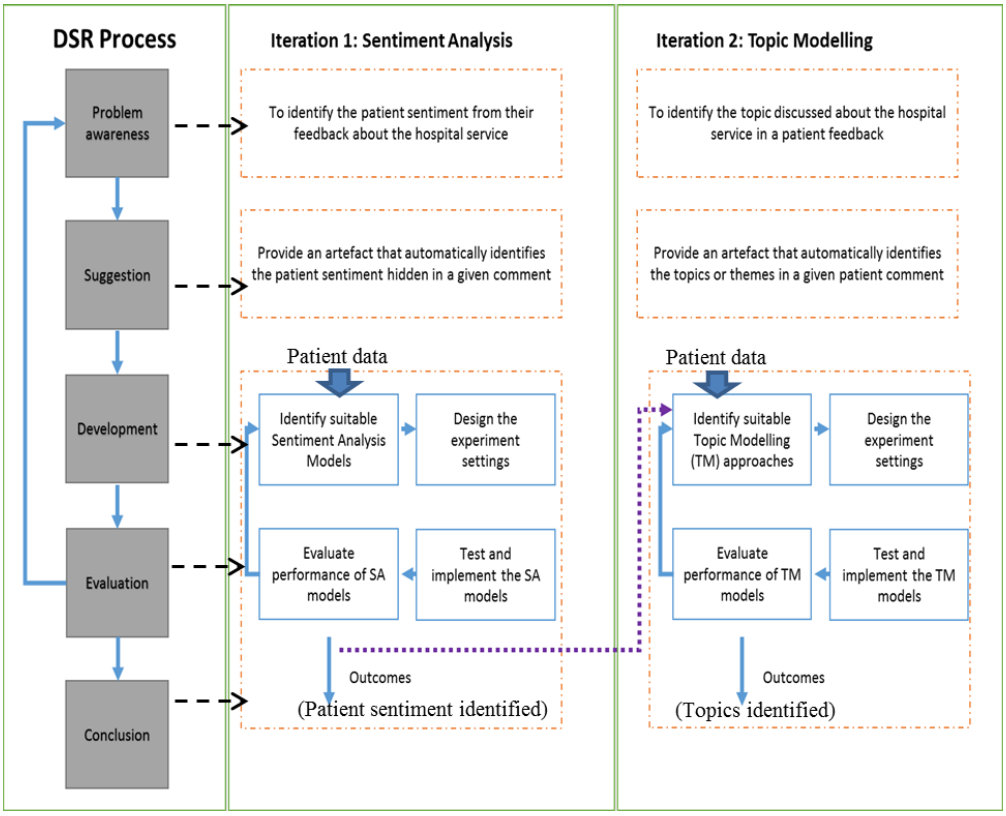


Figure 3-6: Design Science Research approach for the second iteration

In Figure 3-6, it can be observed that the DSR framework for iteration two is more or less similar to the iteration one approach. The problem in this iteration is to identify the topic being discussed by the patient in their feedback about the hospital on the NHS website. Therefore, in line with the problem, the artefact aimed at providing a model that automatically identifies the topics or themes being discussed by the patient in a given comment.

To develop the artefact, the first step is to identify a Topic Modelling approach that is efficient and suitable for the study. In this iteration, the Latent Dirichlet Allocation (LDA) Topic Modelling approach is used. Once the TM approach is identified, the next step is to implement the TM model on the database after designing the experiment settings. After the TM approach is implemented, and then the topics and themes for each patient comment are automatically identified. This reveals which topics are being discussed in the patient feedback. The next step is to identify the sentiment for each topic identified in the patient comment. To achieve this, in this iteration the sentiment identified for each comment from the SA models in iteration one is used. This is indicated by the purple dotted line. That is, for each comment, the identified sentiment is mapped to the relevant topics to obtain a sentiment score for that topic across the patient feedback database. A difference in iteration two when compared to iteration one, during the development stage, is that the Topic Modelling iteration study requires the outcomes of the SA model obtained from the first iteration. That is, this iteration of the study is dependent on the previous iteration.

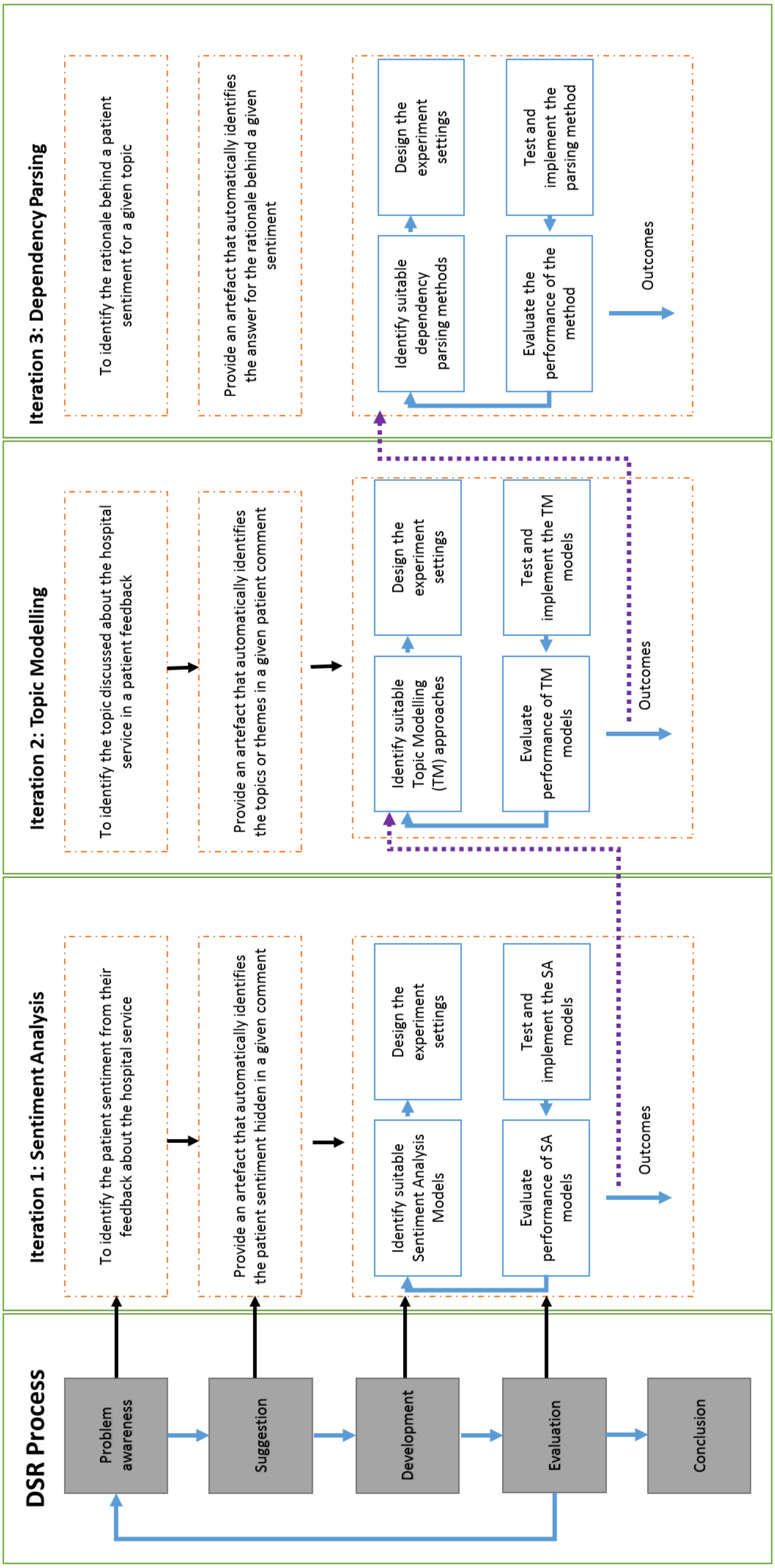
The evaluation of the topics identified in iteration two involves both analytical and observational evaluation. The former refers to when the sentiment score for each topic is computed based on the sentiment scores found from the iteration one. The observational part of the evaluation pertains to when the topics are identified from the TM method. The topics are bags of words that are likely to belong to a particular theme and they need to be manually analysed in order to assign a label to each of them. This is achieved by involving NHS experts to observe each topic’s elements and provide a label for each category.

Similar to iteration one, the outcomes of iteration two will be disseminated in the form of reports and publications. Further, the topics identified will also be visually represented using data visualisation approaches in the R programming environment to make them accessible to hospital staff. Finally, the topics identified by the TM approach in iteration two will be used for further study and analysis in iteration three for the Dependency Parsing investigation aimed at identifying the rationale behind the patient sentiment.

3.4.3 Iteration Three: Rationale Identification of Patient Sentiment

In the third iteration, the aim is to identify the rationale behind a particular sentiment of the patient in a given comment or review.

Figure 3-7: Design Science Research approach for the third iteration.



In other words, the purpose is to find out why the patient is happy or unhappy about a particular topic in a given comment.

In the above figure, it should be noted that the same DSR approach used in the first two iterations of the study is followed. The problem being addressed in this iteration of the study is to find the possible reason behind a patient`s sentiment for a particular topic in a given comment. To achieve this, the artefact developed will implement a Natural Language Processing method called Dependency Parsing. This identifies the keywords in a given comment that could potentially summarise the patient feedback for a given topic in the comment. Specifically, this is achieved by extracting a “noun-adjective” pair for each topic in a given comment and this pair is expected to provide the keywords in the comment that indicate the reason behind the patient`s happiness or unhappiness about the healthcare service provided.

Similar to the previous two iterations, once the suitable Dependency Parsing methods are identified and the implementation environments are finalised, the experiment settings are created and implemented. Further, it needs to be noted that the outcomes of the previous two iterations will be utilized in the development and evaluation stage of the third iteration. This is indicated by the purple dotted line. For this iteration, once again, the R programming environment is utilised and the openNLP and coreNLP methods available in the literature are adopted. Further, similar to how iteration two was dependent on the outcomes of the iteration one, iteration three is also dependent on those from the two preceding iterations. This is because, the “noun-adjective” pair that is extracted to summarise the patient comment is performed on a per topic basis. In other words, for each topic, the reason behind the patient sentiment is identified. In particular, the unigram topics identified from the TM method are used for Dependency Parsing in this study.

The evaluation of the topics identified in iteration two is both analytical and descriptive. The former refers to when the Dependency Parsing methods implemented automatically analyse the feedback data and identify a noun-adjective pair from the comments. The descriptive part of the evaluation pertains to the noun-adjective pair identified providing a summary of the patient comment for a given topic. Thus, it potentially describes the patient comment by a pair of words.

Similar to the previous two iterations, the dissemination of the outcomes of this iteration study will also be via reports and publications. Further, a visualisation of the outcomes of this iteration will be provided. In this visualisation, the users will be able to list all the comments for a chosen topic and then for each comment they will be able to visualise the most salient noun-adjective pair that potentially summarises the comment for the given topic, either negatively or positively.

Thus, the DSR approach is deemed suitable for all the three iterations presented in this thesis. It provides a systematic approach to formulating the problem, identifying potential solutions, implementing and testing the solutions and finally, evaluating and disseminating the outcomes. The iterative approach is a strong aspect of the DSR method and is particularly suitable for the current study each part of the study presented in this thesis is dependent on the outcomes of its predecessor study.

3.5 DSR Methods and Guidelines

Our research follows the DSR guidelines. These are presented and detailed below in Table 3-4.

Table 3-4: DSR Methods and Guidelines adapted from Hevner (2004)

Guideline	Description
Design as an Artefact	DSR must produce a plausible artefact at the end of the process whether it is a model, a method, a contrast or an instantiation.
Relevance of Problem	The main aim of DSR is to produce a technology centred solution for business problems.
Evaluation of Design	The quality of a DSR artefact must be

	demonstrated through a range of evaluation methods.
Contributions of Research	Effective DSR must provide notable and novel contributions in the areas of the design artefact in question.
Rigor of Research	DSR relies on the application of rigorous methods in both the development and evaluation of an artefact.
Design as a Search Process	Search of an effective design artefact requires use of available means to reach desired results.
Communication of Research	DSR must be presented effectively to both technology and business orientated personnel.

All guidelines are crucial for the success of this research. It is hard to agree that one guideline is more significant than the other is as all are equally important and need to be addressed in any DSR artefact. However, it can also be argued that this depends on the context of the research in question. This thesis will put emphasis on the last guideline i.e. communication of research as it is essential to understand that not all readers of thesis will have a technological background.

3.6 Summary

The DSR method is an iterative process that enables the development an artefact for solving a problem. It is suitable for this research, as the main goal is developing an artefact for analysing patient experience automatically. Further, the DSR methodology involves an iterative process and is suitable for this study, because the artefact will be developed in three iterations and each successive iteration study is dependent on the outcomes of the previous iteration. In the following chapters, this three-iteration study is presented.

Chapter 4: Identifying Patient Experience via Sentiment Analysis Approaches

4.1 Overview

In this chapter, the first study in relation to analysing patient experience via Sentiment Analysis approaches is presented. A Sentiment Analysis approach, in its simplest application, classifies given patient feedback into a positive or negative sentiment. Such analysis is useful when there is a large database of patient feedback because it enables to understand the patient feedback as either positive or negative in an automated way. Thus, in this part of our research, we test the performance of the most popular sentiment analysis models in identifying the patient sentiment. The three models: Support Vector Machine (SVM), Naiive Bayes (NB), and Strength of Association (SoA) are explored on *NHS choices* patient feedback database and their accuracy in identifying the overall patient sentiment in the database is evaluated. The outcomes of the study showed the SVM method showed the highest accuracy in identifying the patient sentiment from the database.

Section 4.2 discusses the experiment settings providing detailed information on the dataset used and the modelling approaches followed. Section 4.3 presents the performance evaluation of the three sentiment analysis models on the patient feedback database. The experiment results are presented in Section 4.4. The conclusion to the chapter is provided in Section 4.5.

4.2 Experiment Settings

The machine learning approaches described in section 2.7 have been widely successful in identifying patterns in data and classify them accordingly. The aim of the research presented in this chapter is to apply and evaluate the use of Sentiment Analysis approaches to identify patient experience from a large dataset. The application of Sentiment Analysis approaches is to free patient experience from a priori categorisation and start the process of automating what might be considered as a deeper (and more qualitative) understanding.

4.2.1 NHS Dataset Description

The NHS Outcomes Framework (developed in December 2010) monitors quality improvement and outcome measurements, with the intention of providing up-to-date information about healthcare quality and the patient experience. In the 2015-2016 Framework report, four indicators were included to ensure that people have a positive experience of care – primary care, patient experience of hospital care, friends and family test, and patient experience characterised as poor or worse. The areas for improvement include: people's experience of outpatient care, hospital`s responsiveness to personal needs, people's experience of accident and emergency services, access to primary care services, women and their family experience of maternity services, the experience of care for people at the end of their lives, experience of healthcare for people with mental illness, children and young people's experience of healthcare, and people's experience of integrated care (Group, 2014).

To summarise, the main purpose of NHS Outcome Framework is to receive patient feedback, improve care, and provide evidence that positive outcomes are being achieved. As aforementioned, the National Health Service (NHS) in the United Kingdom (UK) has an online portal called *NHS Choices*, which includes a platform where patients provide both ratings and reviews for a particular NHS hospital. The *NHS Choices* ratings system provides an outline of patient experience, rating is an optional feature that is collected for a specific set of parameters, such as 'cleanliness', 'dealt with dignity', among others (NHS Choices, 2016). Due to this process of collecting feedback for an a priori set of parameters, patient views are thus partially constrained and this could conceal other aspects of patient experience that are of value.

However, patient reviews provide opportunities for detailed patient feedback where opinions about various aspects of the patient experience can be described. Subsequently, by applying Sentiment Analysis approaches to individual patient reviews various hidden aspects of this feedback can be identified. Further, it is essential to analyse and understand patient feedback about hospital services because the quality provided has a significant impact on the patient experience. Studies in the literature have identified that along with the right diagnosis and treatment, other factors such as communication and hospital related services are important factors in patient satisfaction, which thus play a significant role in the retaining of patient loyalty.

On the *NHS Choices* website, patients can provide feedback about NHS hospitals under three main headings. Firstly, they are asked to rate, on a scale of 1 to 5, "how likely they are to recommend the particular hospital to family and friends?" This is the main question on the *NHS Choices* website in that the ratings provided are used to calculate the overall rating for a given hospital. Next, they are asked to provide ratings on five parameters: cleanliness, staff co-operation, dignity and respect, involvement in decisions and same-sex accommodation. The choice allocated ratings for these five parameters are optional to the users. Finally, the participants are given the option to provide a review on the hospital in their own words up to a maximum of 3,000 characters. Figure 4-1 illustrates the review section from the *NHS Choices* website.

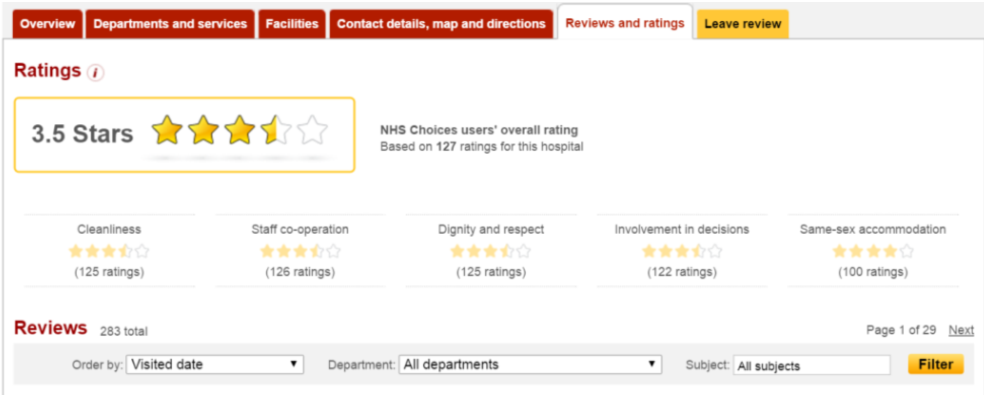


Figure 4-1: A screenshot from the NHS choices website illustrating the various parameters for obtaining patient experience feedback (NHS Choices, 2016)

The dataset collected covers the ratings and feedback comments of patients who received healthcare from hospitals across the United Kingdom. The data collected covers the period January 2010 to July 2015, comprising 76,151 comments with 56,818 labelled observations. The distribution of patient feedback on these ratings and a summary statistic are provided in the figures below. In Figure 4-2, the dataset distribution is pictorially presented. The distributions for both the mandatory ratings optional ratings described previously are provided. In addition, various statistics including the mean, standard deviation and the confidence bounds of the ratings are illustrated.

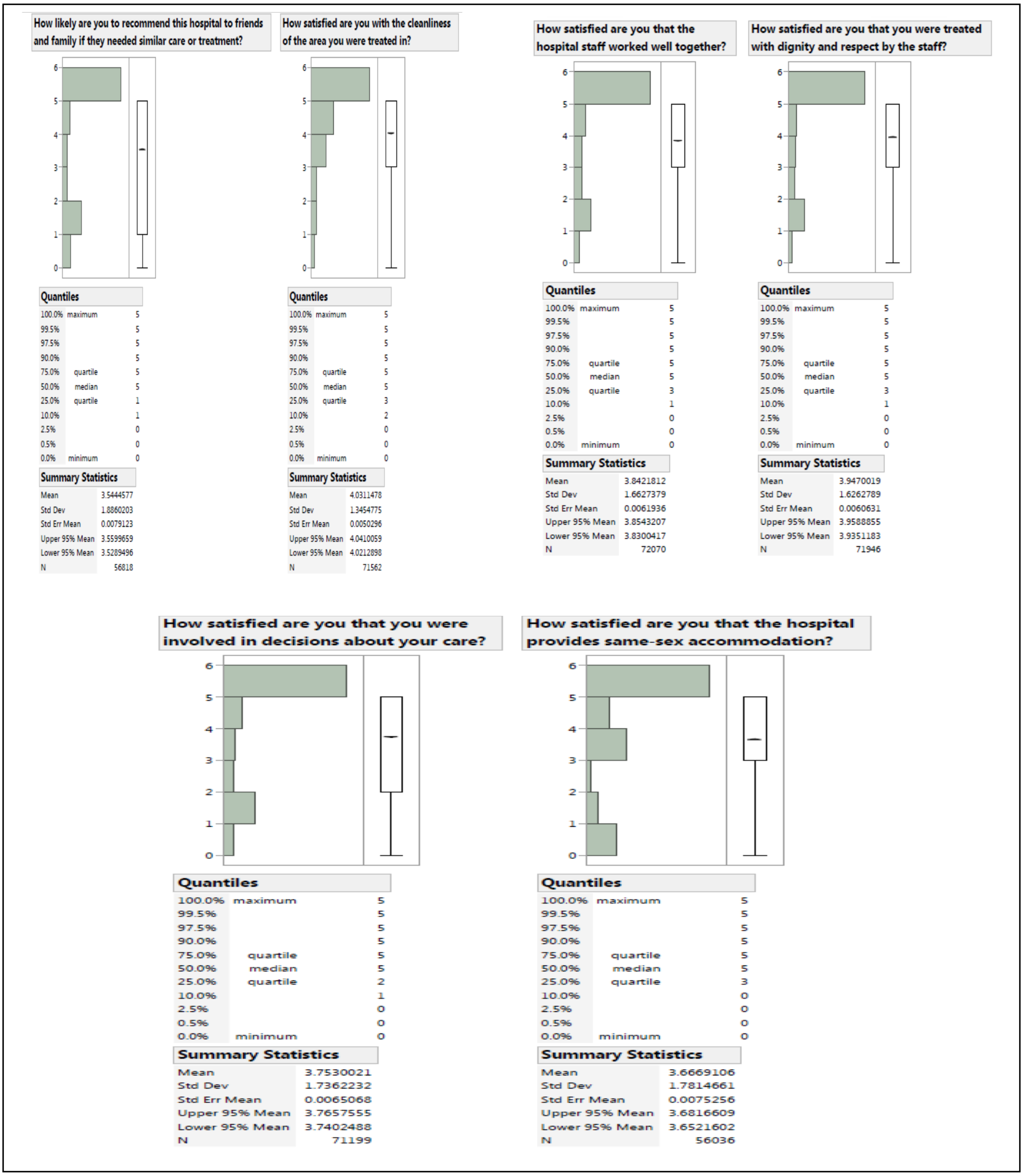


Figure 4-2: A graphical representation of the NHS patient experience dataset distribution

4.2.2 Data Pre-processing

Prior to analysis, the dataset was pre-processed using the R programming language as this language is widely used in data analysis and comes with standard statistical packages for implementation. Various packages were used, particularly, "tm" and "openNLP", because they have feature relevant text mining functions, whilst "Matrix" and "e1071" include several machine learning functions. A list of the packages used is shown in Figure 4-3.

```
## Loading required package: NLP
```

```
library(openNLP)
library(Matrix)
library(e1071)
library(plyr)
library(reshape2)
library(ggplot2)
```

Figure 4-3: Packages used for Sentiment Analysis using R

As a first step, the dataset undergoes a few processing steps, where only columns from the dataset that are relevant for the studies are selected. Specifically, for parsimony purposes, only relevant data fields were extracted and relabelled into Date, Comment, and Label columns in the database. The Date and Comment columns refer to the posted date and the content of the participant's comment: The Label column is used to hold the sentiment of the comment, classified as either *positive* or *negative*.

Having sorted the data by date, in line with standard procedures (Liu, 2015; Feldman, 2013; Feinerer, 2008). The next step was to partition the dataset into a training dataset, test dataset and a validation dataset. This helps to avoid assessing the model quality only on observations seen by the algorithm. In the training dataset, observations were labelled as either 'positive' or 'negative' according to the sentiment inferred from the comments and the associated ratings for each comment (explained below). In the test dataset, observations later than 1st January

2014 were used as an input to derive patterns from the training dataset using the text mining models. In the validation dataset, the values in the Label column were not defined.

Within the dataset, the ratings given by the participants were used as the actual data against which the performance of the text-mining model (i.e. in predicting the patient feedback) was tested. Those data were obtained from the ratings provided by the participants for the question, "How likely are you to recommend this hospital to friends and family if they needed similar care or treatment?”. Due to the skewed distribution of the numerical responses and limitations of the machine learning methods, in order to reduce complexity in the modelling procedure, the continuous scale patient feedback ratings were discretised. Following discretisation, the ratings scores of 1 and 2 were categorised as *negative* and those of 4 and 5 as *positive*. The rating score of 3 was discarded since it is a neutral rating and hence, did not portray any sense of polarisation (Liu, 2015). This categorisation serves as a binary sentiment label for which each text-mining model is trained and assessed.

After discretisation, a clean corpus was extracted from the reviews by applying functions in R to perform the following “cleaning” operations:

- Remove punctuation;
- Reduce all letters to lower case;
- Remove numbers;
- Reduce “*stopwords*” occurrence (e.g. the, who, take that, is, etc.);
- Eliminate whitespaces;
- Convert corpus to plain text format;
- Other minor formatting procedures.

The above operations help to create uniformity across the reviews. Consistency is crucial for model derivation, because small differences in words such as "Love" and "love" can reduce the predictive power of the model. Once the corpus was created, the text was reformatted into a term document matrix, which is a simple, specialised structure that assigns a row to each document (review) and a column for each word. In the cells corresponding to a document, either a 0 or 1 is assigned, thus indicating the presence or absence of the word. For each occurrence of a given word, the count is incremented to capture its frequency of occurrence. The term document matrix is nothing but a description of the frequency of the terms that appear in the document (review). An example of a document matrix is shown in Table 4-1.

Table 4-1: A table illustrating a document term matrix

Document number	Word 1	Word 2	Word 3
Doc 1	1	0	0
Doc 2	1	1	0
Doc 3	1	0	1

Such a matrix can be useful in identifying the frequency of occurrence of words and can be especially so in identifying occurrence of words with similar meaning or synonyms. This knowledge can be further used for identifying topics and themes during document or data clustering.

Sparsity is a natural by-product of this specialised structure, which can negatively affect both computational and model performance. For example, rare words, which occur in only one type of document, could be heavily weighted and thus, lead to biased results. Under certain approaches it is best practice to remove sparse words before performing analysis (Feinerer, 2013), so, a word needed to appear in at least five reviews to be considered for feedback classification in the current work. The five review requirement was chosen because with lesser number of reviews, infrequent terms were not removed as much. Thus, five review requirement gave the optimized result were not too many terms were removed and at the same the time the terms had a good occurrence frequency. Further, it also eliminates typos of the same word and avoids the same words being listed more than once. For instance, words such as ‘colour’, ‘color’, ‘colors’ may refer to the same word but may get listed multiple times thus creating duplicates in the matrix.

With these applied pre-processing steps, the dataset is now ready for application of the machine learning methods for Sentiment Analysis and this is described in the next subsection.

4.2.3 Modelling Approaches

For sentiment classification of the pre-processed data, a dictionary based linguistic approach was applied that uses a “bag of words” to identify sentiment bearing words followed by application of machine learning methods to identify the positive or negative sentiment portrayed by the patient feedback. Three machine learning approaches are evaluated in this study that were described in Section 2.7, namely, Strength of Association (SoA), Naïve Bayes (NB), and the Support Vector Machine (SVM) methods.

The first approach applied is the SoA approach, which scores individual words according to the frequency they appear within each individual review type. A pseudocode of this method is shown below.

```
1. Define function to count word frequency
2. For all words in the document
        Count the number of words and get sumAll
        Count the number of times a word ‘x’ appears and get sumWord
        Calculate the difference between sumAll and sumWord
        Divide the difference by sumAll
        Get the sumWord frequency
3. Store the word frequency information
```

The next approach to be applied is the SVM approach. As mentioned earlier, the SVM approach finds the two closest observations from each class and tries to find a linearly separable plane that maximises the distance between them to form a vector of observations called support vectors. With this boundary in place, new observations (or comments) are mapped into this N-dimensional space and are labelled according to the region they fall within.

Finally, the Naive Bayes (NB) approach is applied to the dataset. This approach provides a middle ground between the purely aggregated scoring approach and the optimised boundary approach. Naive Bayes calculates boundaries, according to the distribution of the words across the labels as well, but also tries to account for the joint probability of the words occurring independently together. It considers each word independently of one another and then tries to estimate the posterior distribution of a review being positive given the joint distribution of the words in the review, assuming independence.

Regarding the above-described methods, all three machine-learning ones are fitted to the training dataset. The next step is for the learning approaches to make predictions about the patient sentiment in the database.

4.3 Performance Evaluation of Sentiment Analysis (SA) Models

The performance evaluation of the Sentiment Analysis models is an important aspect of validating their accuracy and hence, choosing the right evaluation methods is important. Analysing the performance of each model in terms of its prediction accuracy is necessary to identify the best possible model suitable for the context. The performance evaluation metrics chosen for this study are discussed below.

4.3.1 Binary Classification Accuracy

As mentioned earlier, the main aim of the study is the sentiment classification of patient experience. In this stage of the study, the binary classification accuracy of the SA models is analysed and the outcomes will be used for further studies presented in the subsequent chapters.

As the name suggests, in binary classification, the patient experience is classified into two classes, i.e. either a "Positive" or a "Negative" sentiment. The classification performed by the SA models needs to be validated by the reference or the ground truth data. For the tests aimed at validating the performance of each model with the actual sentiment score obtained from the patients, the following question from the *NHS Choices* website was chosen:

"How likely are you to recommend this hospital to friends and family if they needed similar care or treatment?"

The above question is a mandatory question for the patients while providing feedback on *NHS Choices* website. The patient opinion is collected in "stars", where 5 stars represent excellent patient experience and 1 star represents poor or bad patient experience. The ratings obtained for this question were discretised as shown in Table 4-2 below. The ratings were discretised to a scale of 1 to 5 and classified into either a *positive* or *negative* category, as described earlier.

Table 4-2: The classifcation of patient feedback ratings obtained from the NHS Choices website

Sentiment	Score
Negative	1
	2

Neutral	3
Positive	4
	5

The sentiment classification of the SA models obtained from the above table is used as the "ground truth" sentiment score to verify the score obtained by the SA models. The model is considered to have predicted the right sentiment when it bins the patient experience into the same sentiment category as that obtained from the above table. Further, the classification accuracy of the SA model is the percentage of the correct predictions made by the model where the correctness is validated by the ground truth sentiment score.

4.3.2 Sensitivity, Specificity, and ROC curves

The binary classification accuracy can have certain limitations in analysing the performance of the SA models. As pointed out previously, the predictions might occasionally include false positives, i.e. a negative sentiment wrongly identified as a positive one; or false negatives, i.e. a positive sentiment wrongly identified as a negative one. Thus, to test the accuracy of the results obtained further, sensitivity and specificity measures are calculated. In the context of this study, the sensitivity was considered as being the percentage of *positive* sentiments identified correctly as *positive* sentiments. Whilst specificity is the percentage of *negative* sentiments identified correctly as so. A well performing SA model is expected to have a high sensitivity and specificity score.

Further, it is also necessary to visualise the SA model performance in terms of sensitivity and specificity and a popular way of doing it using Receiver Operating Curves, also widely referred to as ROC curves. These can be effective in illustrating the classification performance in that they facilitate choosing the right threshold. A typical ROC curve is shown in Figure 4-4.

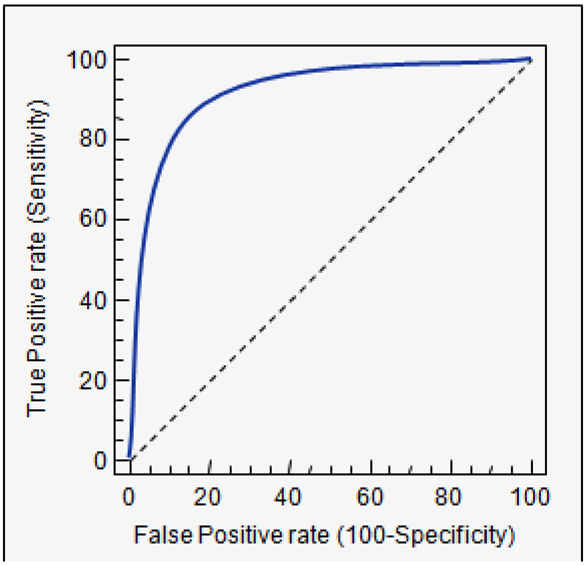


Figure 4-4: A ROC analysis curve. The AUC is the probability of making correct detection

The ROC curve can be described as an operating curve of true positive and false positive rates, i.e. sensitivity vs specificity. The false positive rate is represented as (100 – specificity) when used in a scale range of 0 to 100. The general inference from a ROC curve is the Area under the Curve (AUC) denoted by the area below the curve and the line joining the two ends of the curve (Prakash and Gupta, 2015). The AUC is a representation of the detection accuracy of the SA model: he higher the AUC, the higher the accuracy of the model. In the current study, the classification accuracy, sensitivity, specificity, and ROC curves were used to analyse the SA models. In the next sections, the performance evaluation of the models is presented.

4.4 Experiment Results & Discussion

The three machine learning based SA models described in Section 2.7: SoA, SVM and NB approaches, were tested in the studies. The model was fitted according to the steps described in Subsection 4.2.3 and the SA models were applied in the *RStudio* IDE that is based on the R language framework. An illustration of the commands that build the prediction model is shown below:

```
predict_elem <- predict_label_soa(data_test$comment[1],soa_matrix)
```

```
predict_elem_svm <-predict_label_svm(data_test$comment[1],svm_fit,svm_in_sample)
predict_elem_nb <-predict_label_nb(data_test$comment[1],nb_fit,nb_in_sample)
predict_elem
```

The performance validation of the SA models was performed in two main stages. In the first stage, a single fold test dataset was used for validation, whilst for the second stage; the dataset was divided over multiple folds as a test dataset and then used for cross validation of model performance. Details of these two stages are provided below.

4.4.1 Single fold Cross-Validation

In this stage, the standard practice of performing Sentiment Analysis and prediction was applied. The dataset here is divided into *training* and *test* datasets. The *training* dataset is used to train the SA models on identifying the sentiments from the patient experience. The *test* dataset is used to test the performance of the SA model in predicting the sentiment on a new dataset of patient experience on which they have not been trained.

Once the prediction commands are applied in the *RStudio* on the cleaned dataset, the classification results of each SA model are obtained. The models read the dataset and learn to identify as well as classifying the sentiment of the review, which is generally referred to as training the dataset. Further, the SA models are now trained for sentiment identification and classification. This implies that when a new review is presented to these trained SA models, they will be able to classify the sentiment. For this research, individual predictions on a brand-new review were carried out for all three SA models. New reviews, both positive and negative, were given to the SA model to test and the SA models were able to classify them into appropriate sentiment successfully.

In the above table, it can be seen that all the three SA models classified the sentiment as positive. This was possible because from the training dataset they were able to learn and identify words that are associated and are more likely to belong to a particular sentiment. In this example, the words like love, wonderful, thank you were picked up by the SA models and they associated them with a *positive* sentiment.

Despite all the three SA models being able to identify the sentiment of the new review accurately, it is necessary that extensive testing of the model performance be carried out. For this, the *test* dataset was used. The sentiment predictions of all three models were carried out on the *test* dataset as shown below.

```
predict_vect <- predict_label_soa(data_test$comment,soa_matrix)
predict_vect_svm <- predict_label_svm(data_test$comment,svm_fit,svm_in_sample)
predict_vect_nb <- predict_label_nb(data_test$comment,nb_fit,nb_in_sample)
```

With a large dataset, along with the accuracy it is also essential to check the miscalculation made by the SA models. Thus, the prediction output of each model is collected in terms of true positives (TP), false negative (FN), true negatives (TN), and false positives (FP). The accuracy and miscalculation estimates of the SA models on the test dataset are shown in the table below, which is termed a confusion matrix.

Table 4-3: The sentiment prediction and miscalculation of the SA models on the test dataset. The overall prediction accuracy is also provided in the last column

SA model		Positive	Negative	Prediction accuracy (f-score)
SoA	**Positive**	5,967 **(TP)**	222 **(FN)**	0.67 (67%)
	Negative	4,346 **(FP)**	3,354 **(TN)**	
SVM	**Positive**	10,115 **(TP)**	1,945 **(FN)**	0.84 (84%)
	Negative	198	1,631	

		(FP)	**(TN)**	
NB	**Positive**	10,036 **(TP)**	2,715 **(FN)**	0.78 (78%)
	Negative	277 **(FP)**	861 **(TN)**	

In Table 4-3, the true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP) predicted by each SA model are provided. These results are obtained after verifying each sentiment prediction undertaken by the SA model with the ground truth sentiment that was explained in Subsection 4.3.1 For every instance where the sentiment predicted by the SA model for an individual review matches the ground truth sentiment of that review then it is either a TP or TN depending on whether the sentiment is *positive* or *negative.* Through this process, the TP, TN, FP, and FN for each SA model is obtained. The prediction and the miscalculation values obtained for each model are then used to calculate the prediction accuracy of the SA model further. In the last column of Table 4-3, the prediction accuracy of each SA model is presented. It can be seen that the SVM model performs the best amongst the considered models, with a prediction accuracy of 84%, whilst the NB approach has a prediction accuracy of 78%. The worst performance is shown by the SoA approach that shows an accuracy of only 67%.

Further, the sensitivity and specificity of the models, as discussed in Subsection 4.3.2 , are also calculated using the equations below.

$$Sensitivity = \frac{number\ of\ TPs}{number\ of\ TPs + number\ of\ FNs}$$

$$Specificity = \frac{number\ of\ TNs}{number\ of\ TNs + number\ of\ FPs}$$

Using the above two equations, the *sensitivity* and *specificity* for all the three models calculated and the outcomes are shown in the table below.

Table 4-4: The sensitivity and specificity performance of the SA models

SA model	Sensitivity	Specificity
SoA	0.964	0.435
SVM	0.838	0.891
NB	0.78	0.756

In the above table, it can be seen that the sensitivity of the SoA model is the highest. In other words, despite the low prediction accuracy, it can be said that even at lower instances of identifying the positive sentiments, this model was able to identify and classify positive sentiments more precisely than the other two. The NB model has the lowest sensitivity of 0.78 and the SVM model has a sensitivity score of 0.838. Further, the SVM model has the highest specificity, i.e. the number of occasions of correct negative sentiments identified by this model is higher than the other two SA models.

A summary of the ground truth number of positive and negative reviews along with the predicted values is shown in Table 4-5.

Table 4-5: The total number of positive and negative sentiment reviews identified by the three SA models along with the actual ground truth

SA Model	Positive sentiment	Negative sentiment
Ground truth	10,313	3,576
SoA	6,189	7,700
SVM	12,060	1,829

NB	12,751	1,138

To visualise the performance of the SA models in the study, a segmented bar chart of the actual and predicted sentiment instances is shown in Figure 4-5.The figure below shows a bar graph for all three SA models along with the ground truth-values. It can be observed that the positive reviews identified by the SVM model is closer to the ground truth positive reviews, however, it identifies almost 40% fewer negative reviews when compared to the actual ground truth. The NB approach overestimates the number of positive reviews and grossly underestimates the number of negative reviews present in the dataset. The SoA approach approximately identifies almost 50% fewer positive reviews and estimates 50% more negative reviews than the ground truth.

Figure 4-5: Plot illustrating performance of each SA model considered in the study against the "actual ground truth" data

In this section, the study performed was a single-fold study, i.e. the dataset was divided into a training and test section and then used for validation of the performance of the model. It is essential that further cross-validation studies be carried out to test the robustness of the SA model's performance. The next subsection discusses such a study.

4.4.2 Multi-fold Validation Study

As mentioned earlier, the performance of the model is further assessed by using a multi-fold cross validation approach. Multi-fold validation pertains to the process of making multiple folds of the dataset as test and training datasets. For instance, a ten-fold cross validation study would involve dividing the dataset into ten different variations of training and test dataset, where the partition is randomly chosen, as illustrated in Figure 4-6.

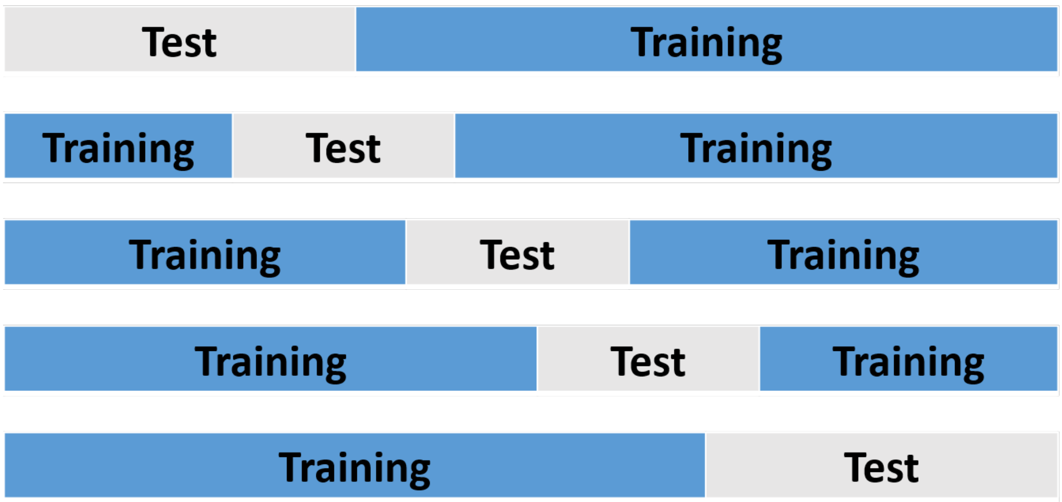


Figure 4-6: A multi-fold break-up of the dataset for cross-validation studies

In the current study, the cross-validation was performed using four folds, i.e. the dataset was partitioned into training and test datasets four times and at each instance, the partition was carried out randomly in *RStudio*. The results for each SA model considered are summarised in the following table.

Table 4-6: Performance results for the NB model in the four-fold cross validation study

1) **SoA model performance**

SoA model		Positive	Negative	f-score	AUC

folds					
Fold 1	**Positive**	3,215 **(TP)**	3,895 **(FN)**	0.607	0.897
	Negative	24 **(FP)**	2,874 **(TN)**		
Fold 2	**Positive**	3,266 **(TP)**	3,891 **(FN)**	0.608	0.902
	Negative	24 **(FP)**	2,826 **(TN)**		
Fold 3	**Positive**	3,214 **(TP)**	3,988 **(FN)**	0.598	0.895
	Negative	25 **(FP)**	2,776 **(TN)**		
Fold 4	**Positive**	3,148 **(TP)**	3,969 **(FN)**	0.599	0.897
	Negative	31 **(FP)**	2,859 **(TN)**		

It can be seen in Table 4-6Table 4-8 that the f-score of the SoA model is low when compared to the other models. The Area under Curve (AUC) scores is relatively consistent and higher in all the four-fold cross validations. The ROC curves of the SoA model in this multi-fold study are shown below.

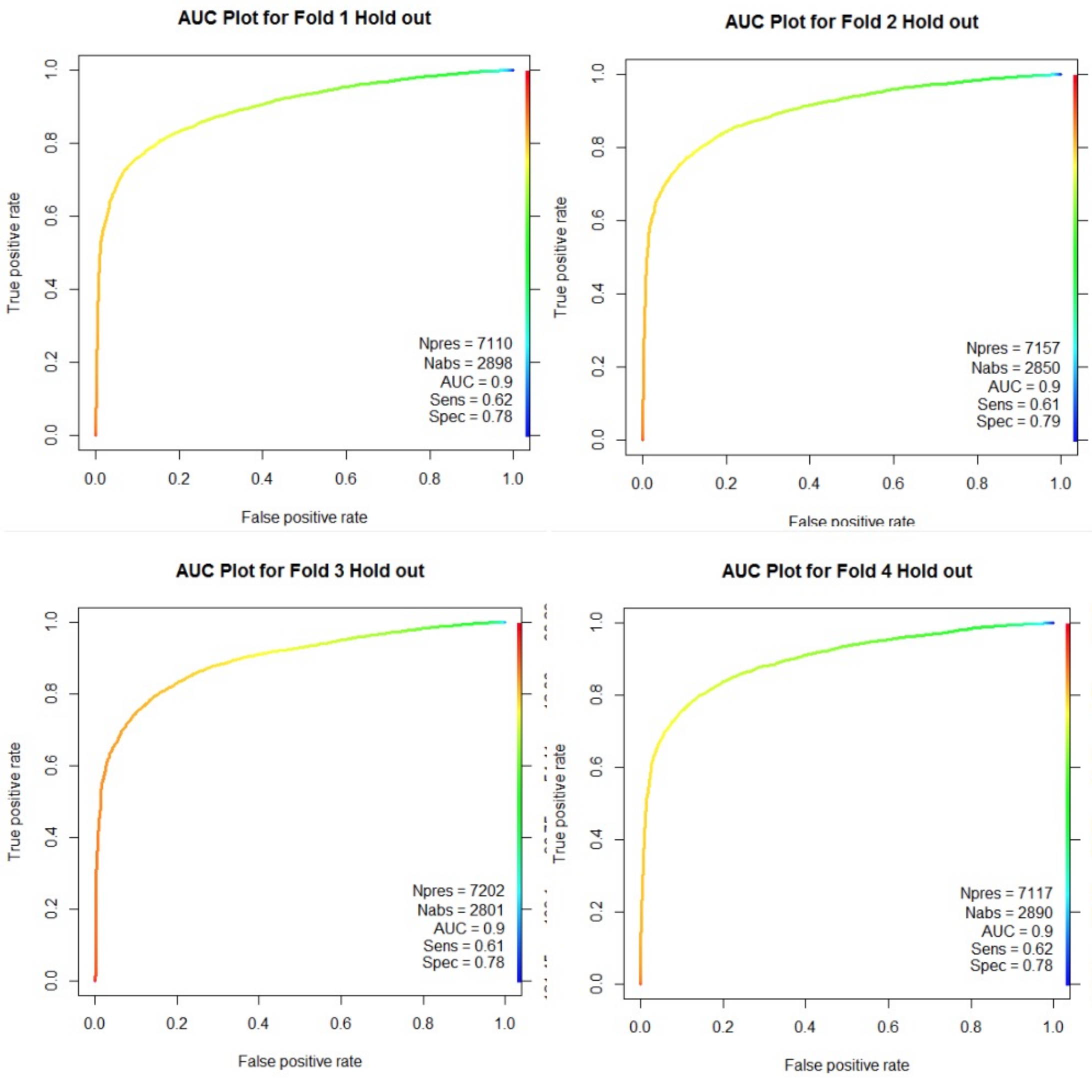


Figure 4-7: ROC curves for the four-fold cross validation study of the SoA model, where AUC = Area under curve, Sens = sensitivity, Spec = Specificity, Npres = number of present observations (i.e. number of positives), Nabs = Number of absent observations (i.e. number of negative reviews)

The above figure shows the ROC curves for the SoA model. It can be observed that in terms of the AUC, the SoA model shows good performance with a value of around 0.9, which is close to the AUC value of the SVM model.

2) SVM model performance

Table 4-7: Performance results the SVM model in the four-fold cross validation study

SVM model folds		Positive	Negative	f-score	AUC
Fold 1	**Positive**	6,999 **(TP)**	377 **(FN)**	0.782	0.921
	Negative	1,271 **(FP)**	1,679 **(TN)**		
Fold 2	**Positive**	7,079 **(TP)**	374 **(FN)**	0.79	0.926
	Negative	1,197 **(FP)**	1,676 **(TN)**		
Fold 3	**Positive**	6,983 **(TP)**	366 **(FN)**	0.780	0.924
	Negative	1,299 **(FP)**	1,678 **(TN)**		
Fold 4	**Positive**	7,010 **(TP)**	372 **(FN)**	0.782	0.924
	Negative	1,275 **(FP)**	1,669 **(TN)**		

It can be seen in Table 4-7 that the f-score and AUC performance of the SVM model are quite consistent in all of the four-fold cross validation study. The ROC curves of the SVM model in this multi-fold study are shown below.

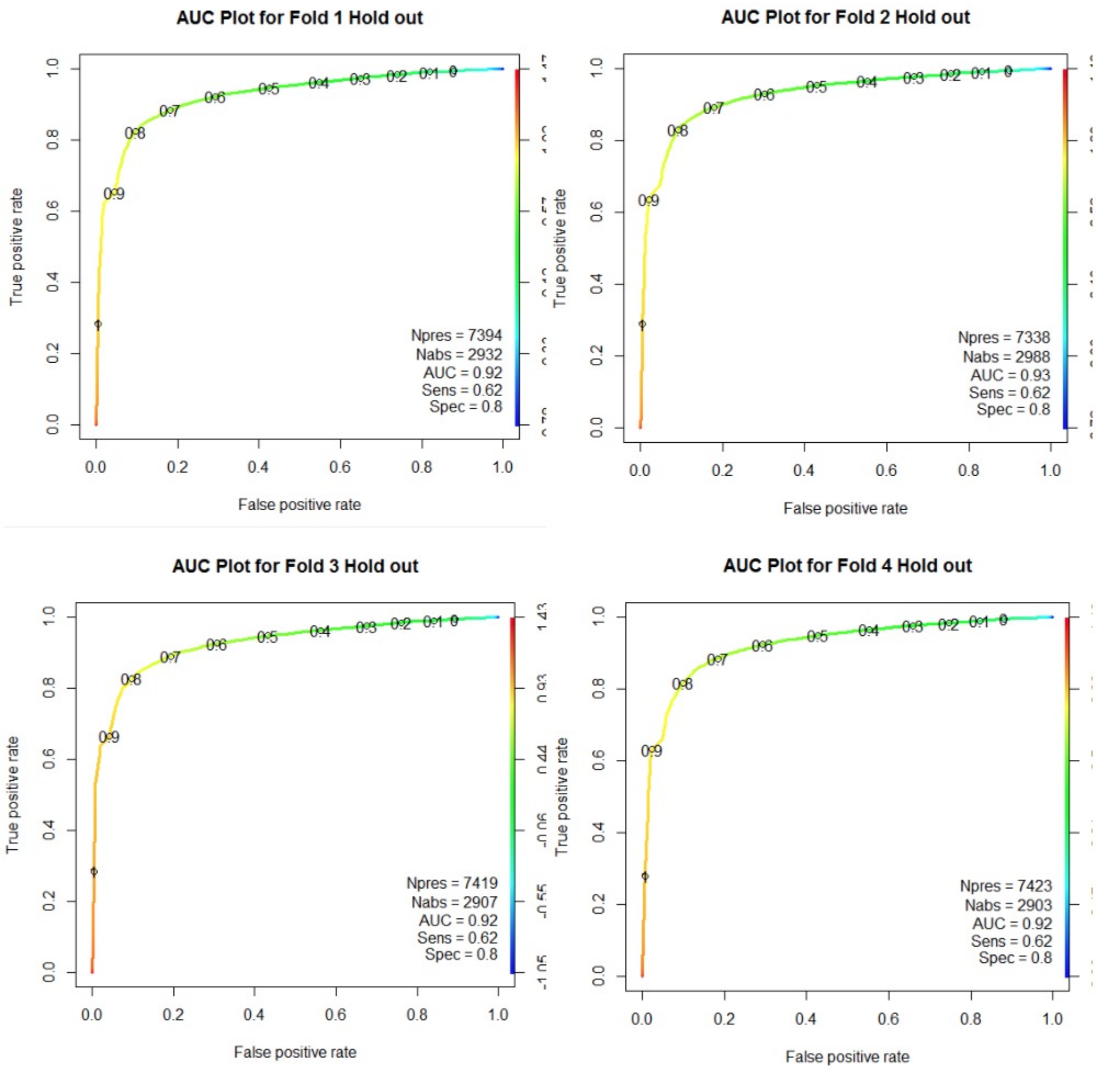


Figure 4-8: ROC curves for the four-fold cross validation study of the SVM model.

It can be seen that the SVM model shows a good performance in terms of the AUC curve, which is around 0.92 and since the curve is closer to the top left corner of the graph, this is an indication that the sentiment detection performance of this model is good.

3) NB model performance

Table 4-8: Performance results for the NB model in the four-fold cross validation study

NB model folds		Positive	Negative	f-score	AUC
Fold 1	**Positive**	6,418 **(TP)**	947 **(FN)**	0.787	0.812
	Negative	866 **(FP)**	2,095 **(TN)**		
Fold 2	**Positive**	6,384 **(TP)**	975 **(FN)**	0.787	0.815
	Negative	849 **(FP)**	2,118 **(TN)**		
Fold 3	**Positive**	6,507 **(TP)**	924 **(FN)**	0.789	0.815
	Negative	842 **(FP)**	2,053 **(TN)**		
Fold 4	**Positive**	6,515 **(TP)**	889 **(FN)**	0.790	0.813
	Negative	867 **(FP)**	2,055 **(TN)**		

It can be seen in Table 4-8 that the f-score and AUC performance of the NB model are also quite consistent in all the four-fold cross validation study, with the performance closely matching that of the SVM model in terms of the f-score. The ROC curves of the NB model in this multi-fold study are shown below.

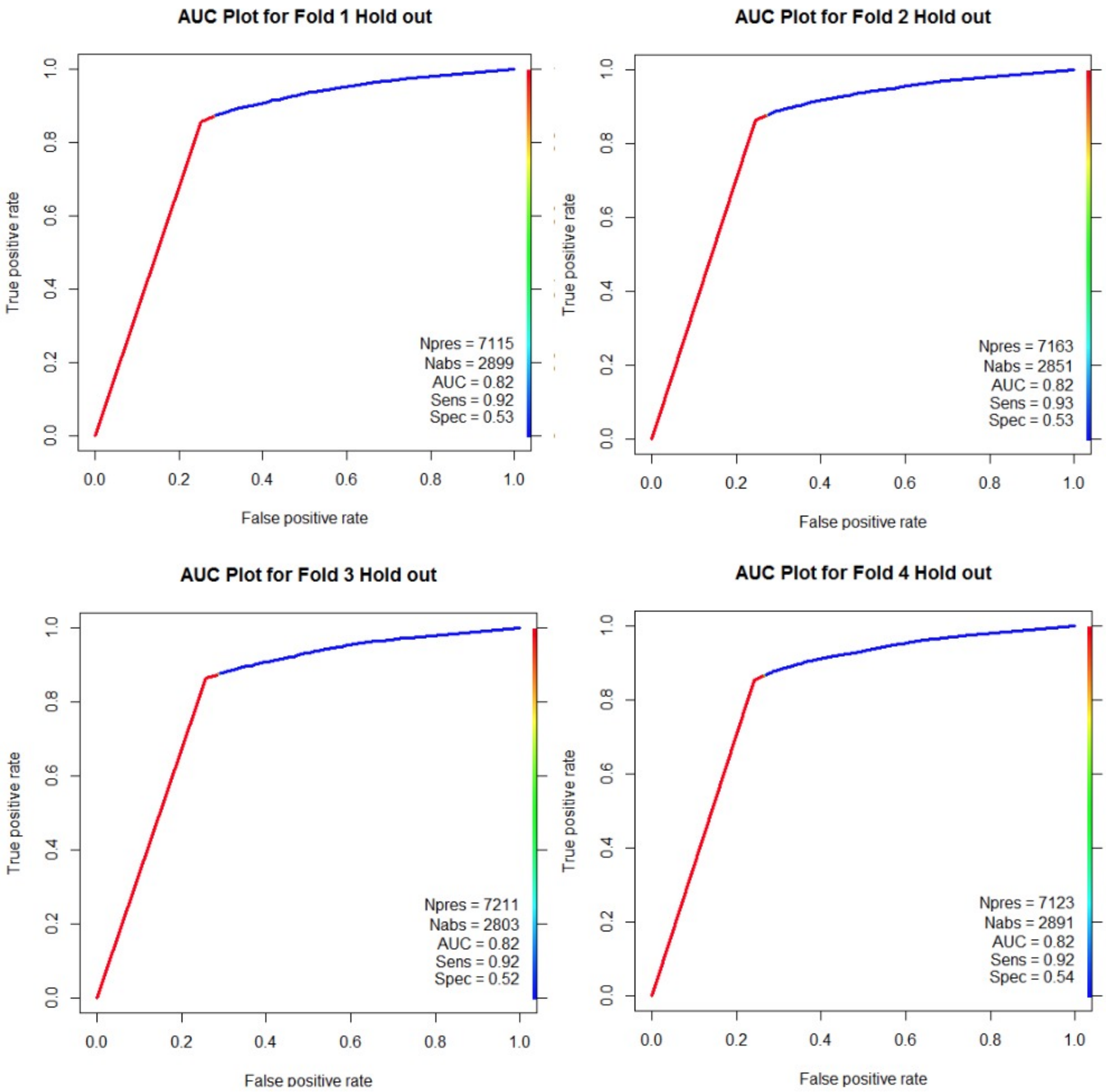


Figure 4-9: ROC curves for the four-fold cross validation study of the NB model.

The figures show that the AUC of the NB model is lower when compared with the SVM model. It should also be noted that the ROC curves are relatively farther from the top left corner of the graph, further indicating a poorer performance than for the SVM model.

4.4.3 Discussion

Subsections 4.4.1 and 4.4.2 have provided the results obtained from the studies for the first iteration. The prediction accuracy (f-score), sensitivity, and specificity evaluation metrics were used to analyse the performance of the SA models considered in the tests. In both the single fold validation and the four-fold validation studies, the SVM model showed better performance than other two SA models considered. The SVM models showed high performance in terms of the prediction accuracy, i.e. the f-scores. A summary of the f-scores, sensitivity, and specificity for all the three SA models in both the single and four-fold validation studies is shown in Table 4-9 below.

Table 4-9: Summary of the f-scores, sensitivity, and specificity scores for all the three models for both cross validation studies

SA Model		F-scores	Sensitivity	Specificity
SVM	**Single-fold**	0.84	0.83	0.89
	Four-fold	0.783	0.62	0.80
NB	**Single-fold**	0.78	0.78	0.756
	Four-fold	0.788	0.92	0.53
SoA	**Single-fold**	0.67	0.96	0.435
	Four-fold	0.603	0.62	0.78

The NB approach provided similar prediction accuracies in both the single and four-fold validation studies. Further, it should be noted that the SVM model performance in the four-fold validation study was similar to the NB approach. The SoA model had the lowest f-score in both instances of the validation study.

In terms of the sensitivity and specificity metrics, the performances were quite varied in both instances of the validation studies and it is difficult to point out a single SA model as being better than the others are. In the single-fold study, the SOA metric showed the highest score of 0.96, followed by the SVM and NB models, respectively. However, in the four-fold validation study, the NB model had a high sensitivity score and both the SVM and SoA delivered scores of 0.62. Similar vary performance in terms of specificity was also seen for all the three models in both instances of the validation study.

Whilst it is difficult to single out a particular model as the most suitable SA model for patient experience sentiment classification from this study, the performance results indicate that the SVM provides the best performance when compared to the other two such models in this study. These results would need further testing and validation on different types of patient experience datasets and be extensively studied for a stronger conclusion in relation to the best patient sentiment classification model.

This study has shown that the current SA models can be used to get an overview of patient sentiments from a given dataset. However, a higher prediction accuracy than that found in this study is desirable and achievable. Regarding which, using a larger and more diverse dataset for training, would improve the prediction accuracy of these models. Nevertheless, a prediction accuracy of almost 85% obtained by the SVM model can be beneficial for the hospitals and clinics in that it represents a 0.85 probability of being right.

The binary classification of the sentiment is the first stage or what could be termed the surface level Sentiment Analysis of the patient experience. The findings have shown that it is possible to predict reliably the sentiment of a comment with a fair amount of accuracy. The identification of the sentiment can give an overview of the general sentiment of the patients about their experience and enable the hospitals to realise their performance in satisfying the patients. This information, however, is limited as it would be beneficial to have a more detailed understanding of different aspects of patient feedback to recognize the areas of patient care that may need improvements. The eventual Sentiment Analysis should be more fine-grained analysis that would identify different aspects or features from the patient feedback database. The findings of this chapter can enable in achieving this goal.

4.5 Chapter Summary

This study has provided evidence that machine learning based SA approaches can be used to identify the overall patient sentiment from a patient feedback dataset. However, a binary classification of patient experience is of limited use to healthcare providers. A more fine-grained analysis of patient feedback is required to get a deeper understanding of the performance of hospitals in different areas of patient care. In the upcoming chapters, the patient feedback analysis is taken to deeper levels in that there is automatic identification and prediction regarding the different topics from the dataset.

Chapter 5: Topic Modelling Analysis of Patent Experience

5.1 Overview

In this chapter, the second iteration of this research is presented. This iteration of the study is aimed at providing an approach to perform a more fine grained analysis of patient experience by identifying themes and topics discussed by patients in their feedback using topic modelling methods, thereby identifying the strengths and limitations of the healthcare provided to the patients that can be used to improve the care provided.

In this chapter, the research work presented focuses on providing a fine-grained analysis of the patient feedback data to identify and understand various aspects frequently discussed by the patients in their feedback. The chapter is structured as follows. Section 5.2 presents a brief overview of the Topic Modelling concept, whilst Section 5.3 discusses the Latent Dirichlet Allocation method of Topic Modelling. The materials and methods used for this iteration are presented in Section 5.4, followed by analysis of the results in Section 5.5. The visualisation of the results obtained is discussed in section 5.6, followed by discussion of the findings of the study in Section 5.7.

5.2 Topic Modelling - Background

Sentiment identification is sufficient for obtaining a broad overview of the general sentiment of the patients for a given hospital. It informs the administration about their success ratio in providing satisfactory healthcare to the patients. However, often this information is not enough as it provides a limited insight into patient feedback. Further, the *NHS* Choices website collects patient ratings cover only five aspects of the service provided, i.e. cleanliness, staff co-operation, dignity and respect, involvement in decisions, and same sex accommodation. In the textual patient feedback, it is inevitable that the patients discuss various aspects of the service if they are not captured by these five aspects or provide important nuances regarding an existing aspect. In sum, it is essential that the hospital is also aware of the various aspects of the healthcare service that are appreciated by the patients and the service aspects in which it is underperforming. Such "fine-grained" analysis of patient reviews can be achieved by adopting approaches such as *Topic Modelling* methods.

A detailed discussion of Topic Modelling approaches and an associated literature review was provided in Section 2.9. It emerged from the latter that there has been a very limited number of studies exploring Topic Modelling approaches for analysing patient experience data. This research gap is being addressed in this work by applying the Topic Modelling method to the NHS patient feedback database.

5.3 Materials and Methods

The LDA model for the Topic Modelling approach was applied to the dataset. It was implemented using the R programming language using packages, such as tm, lda, matrix, plyr, and other relevant packages. The list of packages is shown in Figure 5-1.

```r
library(Matrix)
library(e1071)
library(reshape2)
library(ggplot2)
library(LDAvis)
library(RTextTools)
library(topicmodels)
library(tm)
library(lda)
library(Rmpfr)
library(networkD3)
library(plyr)
library(stringi)
library(openNLP)
```

Figure 5-1: List of packages used for LDA Topic Modelling implementation in R

5.3.1 Dataset

To implement the Topic Modelling approach, the same patient feedback database from the *NHS choices* website was used, which was described in detail in Subsection 4.2.1 it can be recalled that the dataset included in this study had 76,151 comments with 56,818 labelled observations. The cleaning and pre-processing of the dataset was carried out as explained in the Subsection 4.2.2 further, in the dataset, some of the comments that were not useful for the Topic Modelling analysis were removed. The comments considered in the study were accompanied with ratings provided to the mandatory question on the *NHS Choices* website - "How likely are you to recommend this hospital to friends and family if they needed similar care or treatment?”. The corresponding comments that did not have at least a single character were removed from the dataset. In addition, comments that had non-zero or missing ratings for that question were also removed. The cleaning process also included removal of stop words, punctuation, control characters, spaces and digits from the comments.

The cleaning process resulted in a reduced dataset of 59,843 comments. In the next step, the unigram and bigram modelling was applied prior to applying the LDA approach of Topic Modelling.

5.3.2 Unigram and Bigram Modelling

The next step in the study was to implement the unigram and bigram modelling on the clean corpus. First, unigram modelling was applied. In this step, the model grouped single words that probabilistically might belong to a certain topic. The unigram modelling identified single words that were most likely to belong to a particular theme or topic in the database. For instance, words such as ‘birth’, ‘baby’, ‘midwives’, ‘pregnancy’ and similar ones were identified and grouped as belonging to a particular theme from the database. For this, the single words from the database needed to be identified and stored as vectors. That is, the database needed to be processed before applying the LDA model.

The “vocab” is a vector that holds all the single words from the database that will be further used for LDA Topic Modelling.

Unigram modelling is widely used in the literature; however, single words may not be expressive enough on all occasions. Thus, bigram modelling was also tested in this study. For the bigram modelling, the dataset was sampled for “word pairs” that might belong to a particular topic. For example, words pairs, such as car park’, ‘parking space’, ‘blue badge’, ‘disabled-parking’ and other related terms could be grouped into a particular theme, such as parking infrastructure. In order to apply the bigram approach for the LDA model, word pairs needed to be created from the database and stored in the vectors for LDA implementation. The above pairs of words that are created are then stored in the “vocab” vector. The pre-processing process of both unigram and bigram modelling described just now prepared the dataset for implementation of the LDA approach that is described below.

5.3.3 LDA implementation for Aspect Identification

The next step was to apply the LDA model on the dataset and fit the LDA model on the dataset. This enabled listing various aspects or themes of the patient experience that could further be used for predicting the patient sentiment about those aspects. As a first step, the dataset was organised into a format required for the LDA model.

Next, the LDA model was applied to the prepared ‘documents’ using the package available from the R programming community. The initial α value was set at 0.02 and Gibbs sampling for the LDA model was used for sampling the observations from the dataset. This enabled obtaining a sequence of observations from a given multivariate distribution. In other words, from the Dirichlet distribution of the database, a sequence of words is obtained that are likely to belong to a certain theme or joint distribution.

The above step provided a matrix of several topics from the database, with each having words that belonged to that topic and ranked according to how close this association was. Along with the most likely words for a given topic, the implementation of the LDA model also provided the distribution of topics across the corpus. This could be further used to identify the most likely reviews for each topic.

At this stage, various themes that occurred in the patient feedback database that could be considered as the different aspects on which patients provide feedback about the healthcare service had been elicited.

5.3.4 Sentiment Classification of Topics

The next important step was to identify the sentiment score for each discovered topic. In other words, the various aspects of patient experience occurring in the given database had been identified, but how well a given NHS hospital had performed in providing quality patient experience for each aspect remained unknown. Consequently, the sentiment score for each aspect needed to be computed.

To achieve this, the sentiment scores from the analysis results of the SVM model obtained in the study described in the previous chapter were used. The methodology to calculate the sentiment score for each topic is shown below.

For a given topic, T_i, identifies the distribution of the topics in the patient feedback database to get the comments that discuss it, that is:

$$C = (1,2,3, \ldots n)$$

Where, n is the number of comments that belong to topic T_i, i.e. $n \in T_i$.

Next, for all the comments listed, get the sentiment score S corresponding to each comment that is obtained from the SVM method and then compute the mean of all the sentiment scores as given below.

$S_{T_i} = Mean(S_1, S_2, S_3, \dots S_c)$; c Is the number of comments belonging to topic T_i

Thus, the term S_{T_i} is the mean of the sentiment score for a given topic. Therefore, the *mean sentiment score* of each topic has been obtained. The mean of the sentiment score is calculated because the mean value gives an overview how the hospital is performing for that given aspect. For instance, for a topic or theme such as *waiting time*, if the sentiment score is high, it gives an indication that the patients do not have to wait for longer time in order to be served. Further, the sentiment scores of 1 and 2 are considered for the negative mean score and the scores of 4 and 5 are considered for the positive mean score. The neutral score of 3 is not considered for the computation of the *mean sentiment score.*

The *mean sentiment score* of each topic gives an indication of patient opinion about the performance of the hospitals in providing patient care for a given aspect. This gives an automatic fine-grained analysis of hospital performance in relation to providing patient care. In the next section, the results obtained from applying the above methods are discussed.

5.4 Results Analysis

The unigram and bigram LDA modelling has resulted in identification of several topics from the patient feedback database. In the study, the number of topics identified for each method for the first iteration of the study was kept at 30 topics, i.e. from the database 30 different patient experience topics were identified for both the unigram and bigram models. When the number of topics was increased to higher number such as 40, 45, and 50 topics, there was significant overlap between topics and a more cluttered representation of the dataset, both of which made the topics less manageable. Therefore, 30 numbers of topics was an optimized number of topics to fetch from the database.

The process of labelling the topic identified was carried out with the help of medical experts. The procedure involved selecting the top 25 words for each topic identified by the LDA model in *RStudio.* The medical experts reviewed the 25 words for each topic and then provided each with a label. The topics identified and their analysis is provided below.

5.4.1 Unigram Model

Topic Identification

In the unigram model, each topic had 25 ‘single’ words, which were the words that were more likely to belong to a topic. From the unigram model, 30 topics were identified. The Table 5-1 shows the topics identified by the LDA model and that were labelled by the medical experts. The process of labelling an identified topic is illustrated in Table 5-2.

Maternity Department	Emergency Room Service	Dental Check-up	Knee Surgery	Gynaecology visits
Hospital/Discharge Lounge	Waiting Room	Ophthalmology	GI procedure	Paediatric Visits
Emergency Call/Delivery	Patient Reviews	Ear Doctor Visits	Family Doctor	General surgeon
Cancer Treatment	Cardiovascular Treatment	Inpatients Bathroom Complaints	Operation Room	Elderly service department
Patient Appointments and Follow-ups	Non-intensive Surgical Procedure	Admission Ward	Hospital for Royal Families	Clinic Service Experience
Delivery Room	Patient's Good Experience	Parking Space Availability	Telephone Service Department	Reviews on Hospital Service

Table 5-1: The 30 topics identified using the unigram LDA Topic Modelling

In the above table, 30 different topics obtained using the unimodal LDA model are listed. It can be observed that the aspects are vastly different from each other and speak about different features of patient experience. For instance, it can be seen that a common recurring topic in the database is about specific departments, such as maternity department, ophthalmology, cancer treatment and cardiovascular treatment. It can be observed that aspects related to the infrastructure, such as parking space availability, bathroom facilities, are also occurring topics in the patient feedback database.

To illustrate further the topic identification, the list of the 25 words for three topics is shown in Table 5-2 below (Other topics are not shown here due to space limitations). A look into the 25 words for each topic shows that many words are related to a given topic and thus are grouped into a single topic by the LDA model.

Table 5-2: An illustration of the top 25 words for three topics identified by the unigram LDA model

Topic 1 - Maternity Department	Topic 2 - Emergency Room Service	Topic 3 - Dental Check-up
Birth	pain	oral
Baby	call	dentist
midwives	doctor	filling
midwife	antibiotics	nurse
labour	nurse	dental
maternity	hours	tooth
care	relief	didnt
delivery	home	told
ward	infection	pain
staff	gp	feel
pregnancy	severe	teeth
amazing	medication	life
section	back	mental
experience	blood	people
daughter	asked	back
support	night	health
fantastic	dr	wisdom
son	ambulance	mouth

child	painkillers	crown
born	urgent	worse
natal	morphine	wrong
felt	chest	molars
antenatal	agony	bad
unit	worse	years
recommend	days	hospital

In the above table, it should be noted that words such as 'baby', 'midwives', 'pregnancy', etc. were grouped into a single topic using the unigram LDA model. The label "maternity department" was provided by the medical experts as most words for this topic corresponded to pregnancy and related aspects. Similarly, for the "dental check-up" topic, it can be observed that the occurrences of words, such as dental, tooth, pain, wisdom, are grouped into a single group. Thus, this illustrates how relevant words are identified and grouped by the LDA model, which can then be used to label a topic with an appropriate theme.

Mean Sentiment Score of Topics

The next logical step after identifying the topics is to calculate the mean sentiment for each topic`s distribution in the database, which is calculated using the procedure explained in Subsection 5.3.4 The *mean sentiment score* gives an overall indication of the patient experience for the given topic across the dataset and Table 5-3 lists the 30 topics along with this score.

Table 5-3: Mean sentiment scores for the 30 topics identified by the unigram LDA model

Topic Name	*Mean Sentiment*	*Topic Name*	*Mean Sentiment*	*Topic Name*	*Mean Sentiment*
Maternity Department	4.715829	**Emergency Call/Delivery**	4.667739	**Patient Appointments and Follow-ups**	1.978535

Emergency Room Service	2.129049	**Patient Reviews**	4.864157	**Non-intensive Surgical Procedure**	3.088429
Dental Check-up	2.721543	**Ear Doctor Visits**	3.58316	**Admission Ward**	4.759339
Knee Surgery	4.897586	**Family Doctor**	4.767187	**Hospital for Royal Families**	4.590572
Gynaecology visits	2.83755	**General surgeon**	3.27205	**Clinic Service Experience**	4.821004
Hospital/Discharge Lounge	4.216019	**Cancer Treatment**	4.399123	**Delivery Room**	2.18617
Waiting Room	2.021315	**Cardiovascular Treatment**	4.824486	**Patient's Good Experience**	4.706137
Ophthalmology	4.118987	**Inpatients Bathroom Complaints**	2.036872	**Parking Space Availability**	3.143373
GI procedure	4.81401	**Operation Room**	4.877235	**Telephone Service Department**	2.061702
Paediatric Visits	4.394958	**Elderly service department**	2.205752	**Reviews on Hospital Service**	2.124943

In the above table, it can be observed that certain services, such as maternity department, ophthalmology, etc. have good scores, whereas the dental departments, gynaecology visits, delivery have lower sentiment scores.

Performance Validation with the Unigram Model

To analyse the performance of the unigram model further, the Sentiment Analysis scores obtained using the SVM model discussed in the previous chapter are used for comparison. The sentiment scores obtained from the SVM model are employed, because it provided the highest sentiment classification accuracy. Figure 5-2 illustrates the approach used to calculate the predicted sentiment score of each topic identified from the LDA model.

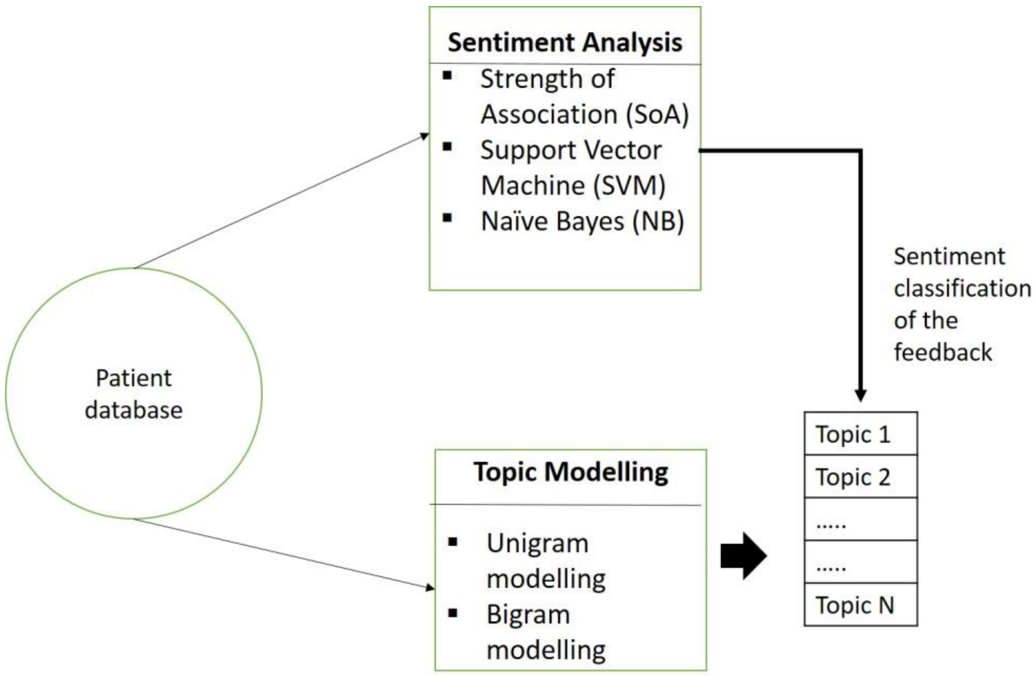


Figure 5-2: The topics identified from the Topic Modelling approach associated with the corresponding sentiment score identified from the SVM method

Since the SVM model made a binary classification of the sentiment into 0 (positive) and 1 (negative), the scores for each comment are in 0s and 1s. The neutral score was not considered. For each topic`s distribution over the patient feedback database, the sentiment scores of only those comments in which a given topic is present are identified. From those comments, the proportion of positive comments was computed based on the ratio of 1s to 0s. For example, of all the comments in which the topic “emergency room service” occurs, it was observed that only 35.5% were classified as positive (1s) from the SVM model. Thus, a *mean predicted sentiment score* for this topic was set to 0.355. Similarly, the *mean predicted sentiment score* for all the 30 topics was computed, which was then compared with the *mean sentiment score* obtained from the topic modelling and listed in Table 5-3. In order to compare the scores on the same scale, the *mean sentiment scores* were normalised to a range of 0 to 1. Thus, both the scores were in the same range.

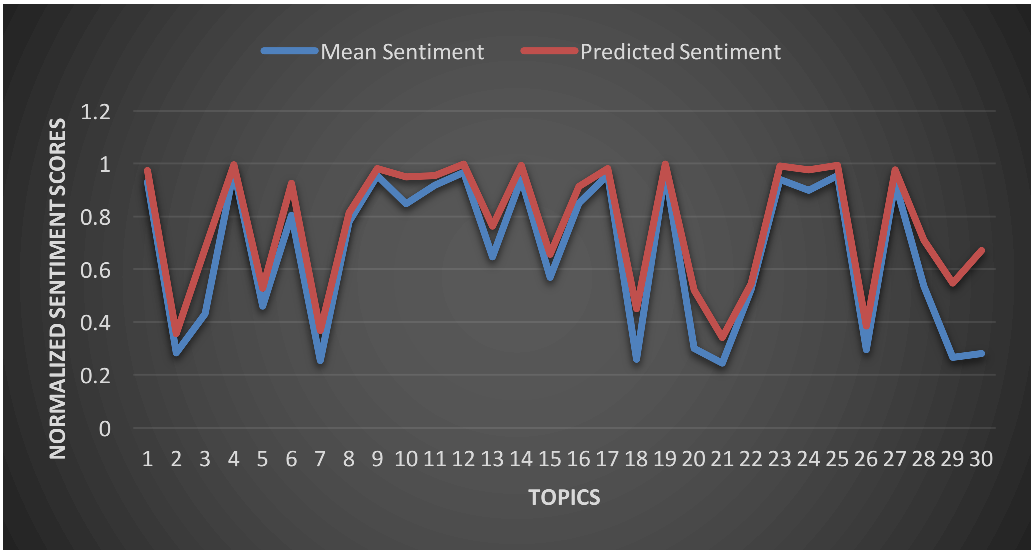


Figure 5-3: Mean sentiment scores vs the predicted sentiment scores for the 30 topics identified from the unigram model

In Figure 5-3, a visualisation of the scores of the *mean sentiment scores* compared with the *mean predicted sentiment score.* It can be observed that the two lines are very close to each other and overlap for a majority of the topics. Since the *mean sentiment scores* obtained for the topics is close to the *mean predicted sentiment score* obtained by the SVM model, it indicates that the our method of associating the topics obtained in second iteration with the sentiment scores obtained from the first iteration would give an insight into the patient feedback for each topic. Thus, this approach provides the user with the ability to identify topics in patient feedback and also understand the overall sentiment behind each topic in an easy and automated way.

Hospital Specific Heatmap Distribution of Topics

The information of frequently occurring topics in the patient feedback for a given hospital can help the hospital administration to find the areas of patient care they need to focus on. The NHS patient feedback database included comments from over 1,000 hospitals. The 30 topics identified from the unigram model have a distribution for the reviews across the 1000 hospitals. It can be beneficial to have knowledge of the most commonly occurring topics in patient reviews for each hospital and the frequency of a topic being reviewed in the patient comments for a particular one. This information can further help the hospital to identify negative feedback that is more frequently given about a particular topic and thus, set priorities on improving the service in that particular topic area. A screenshot of the visualisation of the topics occurring in the patient feedback for each hospital is given in the Heatmap chart in Figure 5-4 below and due to space limitations, only a screenshot is provided here.

Hospital	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17
Abingdon Community Hospital	0	0	0	0	0	0	0.166667	0	0	0	0	0	0	0.125	0.083333	0	0
Accrington Victoria Hospital	0	0.0625	0	0	0	0	0.03125	0	0	0.0625	0	0.03125	0.03125	0	0	0	0
Acomb Gables	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Addenbrooke's	0	0.027027	0.006757	0.081081	0.013514	0.006757	0.074324	0	0.033784	0.013514	0.067568	0.074324	0.040541	0.054054	0.02027	0.033784	0
Aintree University Hospital	0	0.045918	0.035714	0.020408	0.010204	0.02551	0.061224	0.015306	0.061224	0	0.020408	0.02551	0.015306	0.091837	0.020408	0	0.005102
Aire Court	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aire Court Community Unit	0	0	0.333333	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Airedale General Hospital	0.025362	0.032609	0.01087	0.036232	0.003623	0.007246	0.036232	0.003623	0.036232	0.014493	0.072464	0.07971	0.01087	0.057971	0.01087	0.014493	0.043478
Alder Hey Children's NHS Foundat	0	0.095238	0	0.047619	0	0	0.095238	0.047619	0.047619	0.047619	0.095238	0.095238	0.047619	0	0	0	0
Alderney Hospital	0	0	0.333333	0	0	0	0	0	0	0	0	0.111111	0	0.222222	0	0	0
Alexandra Hospital	0.04878	0.073171	0.036585	0.04878	0.036585	0.012195	0.060976	0	0.012195	0.012195	0.04878	0.036585	0	0.073171	0.02439	0.02439	0.012195
Alfred Bean Community Hospital	0	0	0	0	0	0	0.142857	0	0	0	0	0	0.142857	0	0.142857	0	0
Alfred Bean Hospital	0.166667	0	0	0	0	0	0	0	0	0	0	0	0	0	0.166667	0	0

Figure 5-4: A screenshot of the Heatmap visualisation of topic occurrence for each hospital covered in the NHS patient feedback database

In the above figure, it can be observed that for Abingdon Community Hospital the topic 7 has a value of 0.166, which indicates that the topic 7 is discussed in 16% of the reviews provided for this hospital. Similarly, it can be seen that 12.5% of the comments for this hospital pertained to topic 14. In sum, using such Heatmap visualisation, the hospitals will be able to observe the most commonly discussed topics in the patient feedback attributed to their particular institution.

5.4.2 Bigram Model

In the unigram model, 'single' words were identified that were most likely to belong to a certain theme or topic. The next step of the study included applying the LDA model using a bigram approach. As the name suggests, in this topic, two words or a 'word-pair' are identified as those that are likely to belong to a certain theme or topic.

Topic Identification

Similar to the unigram model, regarding the bigram approach, for each topic the top 25 'word-pairs' were selected, which were the words that were more likely to belong to a particular topic than others. The subsequent pair of words after 25 'word-pairs' were turning out to be less relevant to the topic. Hence, for each topic, the top 25 'word-pairs' were chosen. For the bigram model too the number of topics identified was 30. Table 5-4 shows the topics identified by the LDA model that were labelled by the medical experts.

Table 5-4: The 30 topics identified using the Bigram LDA Topic Modelling

Waiting Experience	Dissatisfied Service	Parking Infrastructure	Parental Hospital Visits	Comfortable Experience
Description of Quality Care	Nurse Experiences	Ophthalmology Visits	Hospital Recognition	Staff Comments
Emergency Call/Delivery	Patient Reviews	Ear Doctor Visits	Family Doctor	General surgeon
Appointment Setting	Positive Procedural Experience	Children Visits	Phone Interactions	Maternity

Emergency Care	Pain Management	Serious Disease and Cancer	Positive Staff Interactions	Communicating with Doctors
Hospital facilities	Knee and Hip Procedures	Surgery room	Heart Related	Dental Visits
Fractures and Minor Injuries	Elderly service department	Staff Communications	Laboratory Testing	Dissatisfied Care

In the above table, the 30 different topics obtained using the bigram LDA model are listed. It can be observed that the aspects are vastly different from each other and speak about different features of patient experience. Further, it can also be seen that many topics found using the bigram approach are similar to those identified from the unigram approach shown in Table 5-1. For instance, topics such as parking infrastructure, ear doctor visits, knee and hip procedures, and dental visits are similar to the topics identified from the unigram approach. Further, several topics identified were significantly different from those identified through the unigram approach. For instance, topics such as pain management, laboratory testing, communicating with the doctors are new topics that were identified in the bigram approach. Thus, through using the bigram approach novel themes and topics were identified from the patient feedback database that did not arise from the unigram method. These results can be considered as additional knowledge for automated understanding of patient feedback.

To illustrate the topic identification further, a list of the 25 word-pairs for three topics is shown in Table 5-5 below (Other topics are not shown here due to space limitations). Looking at the 25 word-pairs for each topic, reveals that many words are related to a given topic and thus, were grouped into a single topic by the LDA model.

Table 5-5: An illustration of the top 25 word pairs for each of three topics identified by the bigram LDA model

Waiting Experience	Dissatisfied Service	Parking Infrastructure
waiting_room	mental_health	car_park
waiting_area	member_staff	car_parking
waiting_time	made_feel	appointment_time
waiting_hours	staff_rude	main_entrance
wait_hours	extremely_rude	running_late
long_wait	dont_care	clinic_running
people_waiting	staff_member	parking_space
told_wait	customer_service	parking_charges
triage_nurse	formal_complaint	minutes_late
waited_hours	family_member	blue_badge
half_hours	health_issues	waiting_area
hour_wait	health_care	car_parks
half_hour	eye_contact	parking_ticket
left_waiting	rude_unhelpful	find_parking
hours_waiting	left_feeling	half_hour
back_waiting	people_skills	disabled_parking
sat_waiting	crisis_team	minutes_appointment
waited_hour	staff_dont	back_car
wait_waiting	treat_people	member_staff
reception_desk	members_staff	appointment_times
hours_doctor	wasting_time	main_reception
waiting_times	treat_patients	hospital_car
patients_waiting	reception_staff	park_car
reception_staff	health_problems	late_appointment
waiting_hour	care_patients	hour_late

In the above table, it can be seen how word-, such as ‘waiting_room’, ‘waiting_area’ and ‘waiting_time’ belong to a common theme and thus, were grouped under a single topic by the bigram LDA model. Similarly, in the parking infrastructure topic, it can be observed that word-pairs, such as ‘car_park’, ‘parking_space’, ‘parking_charges’, are more closely related to each other and hence, are grouped into this topic. The word-pairs identified by the LDA model for each topic were then reviewed and provided a label by the medical experts. In sum, this has illustrated how relevant word-pairs are identified and grouped by the LDA model, which can then be used to label a topic with an appropriate theme.

Mean Sentiment Score of Topics

Similar to the approach followed in the unigram model, the next logical step after identifying the topics is to calculate the mean sentiment for each topic`s distribution over the database. The *mean sentiment score* for the topic was calculated using the procedure mentioned in Subsection 5.3.4 gives an overall indication of the patient experience for the given topic across the dataset. Table 5-6 lists the 30 topics along with the *mean sentiment score.*

Table 5-6: Mean sentiment scores for the 30 topics identified by the bigram LDA model

Topic Name	*Mean Sentiment*	*Topic Name*	*Mean Sentiment*	*Topic Name*	*Mean Sentiment*
Waiting Experience	2.172314	**Appointment Setting**	1.754075	**Hospital facilities**	3.952548
Dissatisfied Service	1.721997	**Positive Procedural Experience**	4.621491	**Knee and Hip Procedures**	4.659152
Parking Infrastructure	2.507555	**Children Visits**	3.758021	**Surgery room**	4.485739
Parental Hospital Visits	3.014896	**Phone Interactions**	1.665391	**Heart Related**	4.566237
Comfortable Experience	4.588028	**Maternity**	3.831199	**Dental Visits**	2.535968
Description of Quality Care	4.651438	**Emergency Care**	4.362342	**Fractures and Minor Injuries**	4.194191
Nurse Experiences	1.866747	**Pain Management**	2.175537	**Staff Communications**	4.555059

Ophthalmology Visits	2.67265	**Serious Disease and Cancer**	4.26584	**Elderly service department**	4.500722
Hospital Recognition	2.501927	**Positive Staff Interactions**	4.599234	**Laboratory Testing**	2.882353
Staff Comments	4.33223	**Communicating with Doctors**	1.791641	**Dissatisfied Care**	2.036122

In the above table, it can be observed that the topics have a varied range of sentiment scores. For instance, ophthalmology visits have a low score of 2.67, whereas fractures and minor injuries along with the hip and knee department have scores above 4 scores.

Performance Validation with the Bigram Model

Similar to the performance validation of the unigram model, that for the bigram model was carried out. The *mean sentiment score* of the SVM for the comments belonging to the bigram topic model was used to predict the *mean predicted score* for each bigram topic. Figure 5-5 shows the plot for the normalised scores of the mean sentiment and the predicted sentiment.

It can be seen that when compared to the unigram performance demonstrated in Figure 5-3, the *predicted sentiment* by the bigram model is not as close to the *mean sentiment*. Importantly, the sentiment predicted by the SVM model is based on unigrams and hence, the topics identified by the unigram model are more likely to have closer *predicted sentiment* scores to this model than the bigram model.

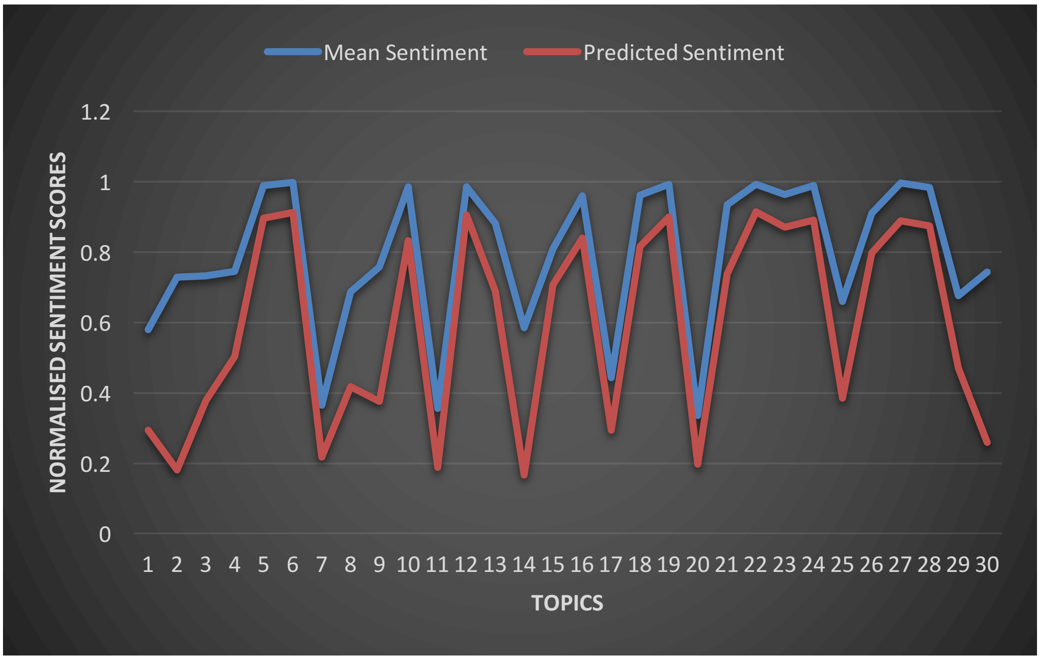
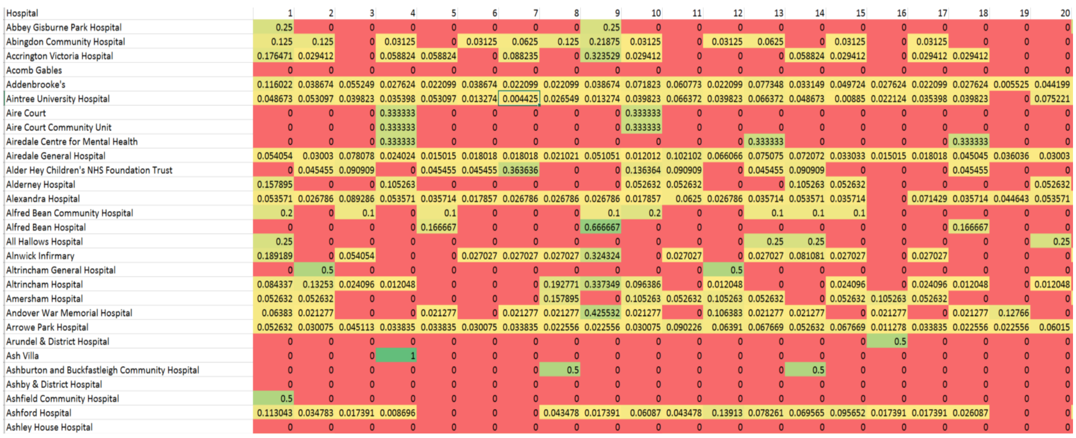


Figure 5-5: Meant sentiment scores vs the predicted sentiment scores for the 30 topics identified from the bigram model

Hospital Specific Heatmap Distribution of Topics

Similar to the unigram model, a Heatmap distribution of the topics for the bigram model is shown in the Figure 5-6. Using this Heatmap visualisation, the hospitals will be able to observe the most commonly discussed topic about their institution in the patient feedback.



Hospital	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Abbey Gisburne Park Hospital	0.25	0	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0	0	0	0	0
Abingdon Community Hospital	0.125	0.125	0	0.03125	0	0.03125	0.0625	0.125	0.21875	0.03125	0	0.03125	0.0625	0	0.03125	0	0.03125	0	0	0
Accrington Victoria Hospital	0.176471	0.029412	0	0.058824	0.058824	0	0.088235	0	0.323529	0.029412	0	0	0	0.058824	0.029412	0	0.029412	0.029412	0	0
Acomb Gables	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Addenbrooke's	0.116022	0.038674	0.055249	0.027624	0.022099	0.038674	0.022099	0.022099	0.038674	0.071823	0.060773	0.022099	0.077348	0.033149	0.049724	0.027624	0.022099	0.027624	0.005525	0.044199
Aintree University Hospital	0.048673	0.053097	0.039823	0.035398	0.053097	0.013274	0.004425	0.026549	0.013274	0.039823	0.066372	0.039823	0.066372	0.048673	0.00885	0.022124	0.035398	0.039823	0	0.075221
Aire Court	0	0	0	0.333333	0	0	0	0	0	0.333333	0	0	0	0	0	0	0	0	0	0
Aire Court Community Unit	0	0	0	0.333333	0	0	0	0	0	0.333333	0	0	0	0	0	0	0	0	0	0
Airedale Centre for Mental Health	0	0	0	0.333333	0	0	0	0	0	0	0	0	0.333333	0	0	0	0	0.333333	0	0
Airedale General Hospital	0.054054	0.03003	0.078078	0.024024	0.015015	0.018018	0.018018	0.021021	0.051051	0.012012	0.102102	0.066066	0.075075	0.072072	0.033033	0.015015	0.018018	0.045045	0.036036	0.03003
Alder Hey Children's NHS Foundation Trust	0	0.045455	0.090909	0	0.045455	0.045455	0.363636	0	0	0.136364	0.090909	0	0.045455	0.090909	0	0	0	0.045455	0	0
Alderney Hospital	0.157895	0	0	0.105263	0	0	0	0	0	0.052632	0.052632	0	0	0.105263	0.052632	0	0	0	0	0.052632
Alexandra Hospital	0.053571	0.026786	0.089286	0.053571	0.035714	0.017857	0.026786	0.026786	0.026786	0.017857	0.0625	0.026786	0.035714	0.053571	0.035714	0	0.071429	0.035714	0.044643	0.053571
Alfred Bean Community Hospital	0.2	0	0.1	0	0.1	0	0	0	0.1	0.2	0	0	0.1	0.1	0.1	0	0	0	0	0
Alfred Bean Hospital	0	0	0	0	0.166667	0	0	0	0.666667	0	0	0	0	0	0	0	0	0.166667	0	0
All Hallows Hospital	0.25	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0	0	0	0	0	0.25
Alnwick Infirmary	0.189189	0	0.054054	0	0	0.027027	0.027027	0.027027	0.324324	0	0.027027	0	0.027027	0.081081	0.027027	0	0.027027	0	0	0
Altrincham General Hospital	0	0.5	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
Altrincham Hospital	0.084337	0.13253	0.024096	0.012048	0	0	0	0.192771	0.337349	0.096386	0	0.012048	0	0	0.024096	0	0.024096	0.012048	0	0.012048
Amersham Hospital	0.052632	0.052632	0	0	0	0	0	0.157895	0	0.105263	0.052632	0.105263	0.052632	0	0.052632	0.105263	0.052632	0	0	0
Andover War Memorial Hospital	0.06383	0.021277	0	0	0.021277	0	0.021277	0.021277	0.425532	0.021277	0	0.106383	0.021277	0.021277	0	0.021277	0	0.021277	0.12766	0
Arrowe Park Hospital	0.052632	0.030075	0.045113	0.033835	0.033835	0.030075	0.033835	0.022556	0.022556	0.030075	0.090226	0.06391	0.067669	0.052632	0.067669	0.011278	0.033835	0.022556	0.022556	0.06015
Arundel & District Hospital	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0
Ash Villa	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ashburton and Buckfastleigh Community Hospital	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0.5	0	0	0	0	0	0
Ashby & District Hospital	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ashfield Community Hospital	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ashford Hospital	0.113043	0.034783	0.017391	0.008696	0	0	0	0.043478	0.017391	0.06087	0.043478	0.13913	0.078261	0.069565	0.095652	0.017391	0.017391	0.026087	0	0
Ashley House Hospital	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5-6: A screenshot of the Heatmap visualisation of the bigram topic occurrence for each hospital covered in the NHS patient feedback database

5.5 Topic Model Visualisation

In the previous sections, identifying topics using unigram and bigram modelling was covered. After manual inspection, 30 topics from each model were identified as were the top 25 relevant terms for each topic. As there are a considerable number of topics, it would be difficult or inconvenient for the users to manually view each topic label and the relation between the topics and their terms. Thus, a visual representation of the topics, its terms, the frequency of occurrence of the terms would enable the users to analyse the topics identified visually and in an easy and interactive manner. To facilitate visualisation of the topics, the "LDAvis" method of from the LDA models presented in (Sievert and Shirley, 2014; Chuang, Manning and Heer, 2012) was utilised. The LDAvis provides an interactive visualisation of the identified topics from the LDA model, thereby enabling a better understanding of the topics identified in terms of their meaning, the relationship between them, and the prevalence of each. A screenshot of the first interface of the topic visualisation is shown in Figure 5-7.

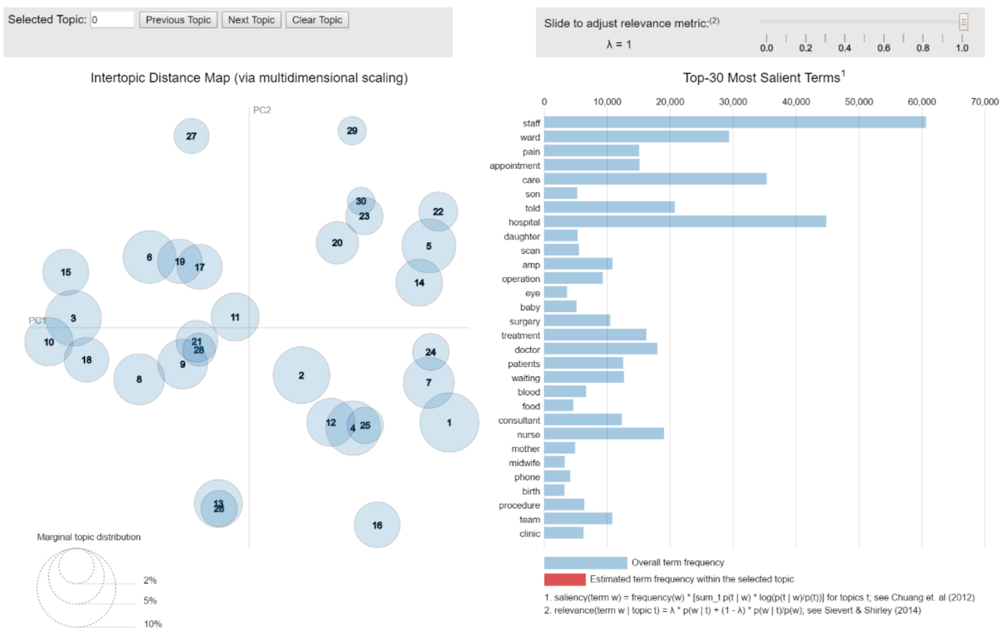


Figure 5-7: The layout of LDAvis, with the global topic view on the left and the term bar charts (with the top 30 terms) on the right

5.5.1 Topic Specificity

In the above figure, several aspects of the LDAvis visualisation method can be seen. On the left side of the figure, the 30 topics identified from the unigram model are plotted as circles. These ‘topic circles’ represent the distribution of the topics across the patient feedback database. The map uses dimensionality reduction techniques such as Principal Component Analysis (PCA) to map the closeness of the topics to each other. The x and y axes are the principal component values that enables to map the topic distribution on to a chart. Further, the size of the circle for each topic is an indication of the occurrence or distribution of it across the database. A large circle diameter indicates that that particular topic has a greater occurrence in the patient feedback database, which is further illustrated in the figure below.

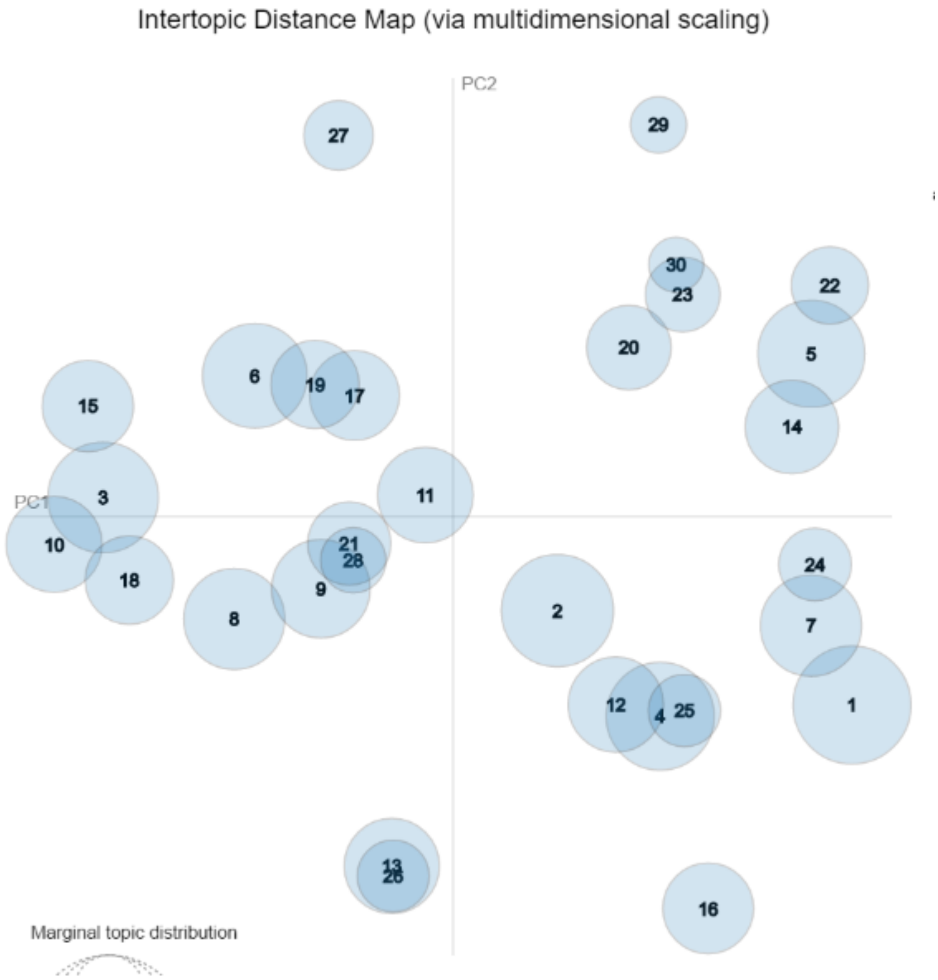


Figure 5-8: Illustration of each topic's distribution across the database

In the Figure 5-8, it can be observed that the circles belonging to topic numbers 1, 2 and 3 have relatively larger diameters when compared to those pertaining to topic numbers, 26, 27, 28, and 29. This is an indication that the topics such as topic 1, 2 and 3, as do some others, have a larger distribution across the database. Reconsidering Figure 5-7 again, the right part of the figure lists the top 30 most salient terms from the NHS patient feedback database. This implies that the terms listed in the right panel occur more frequently across the database when compared to others. A zoomed in version of the right panel is shown in the figure below.

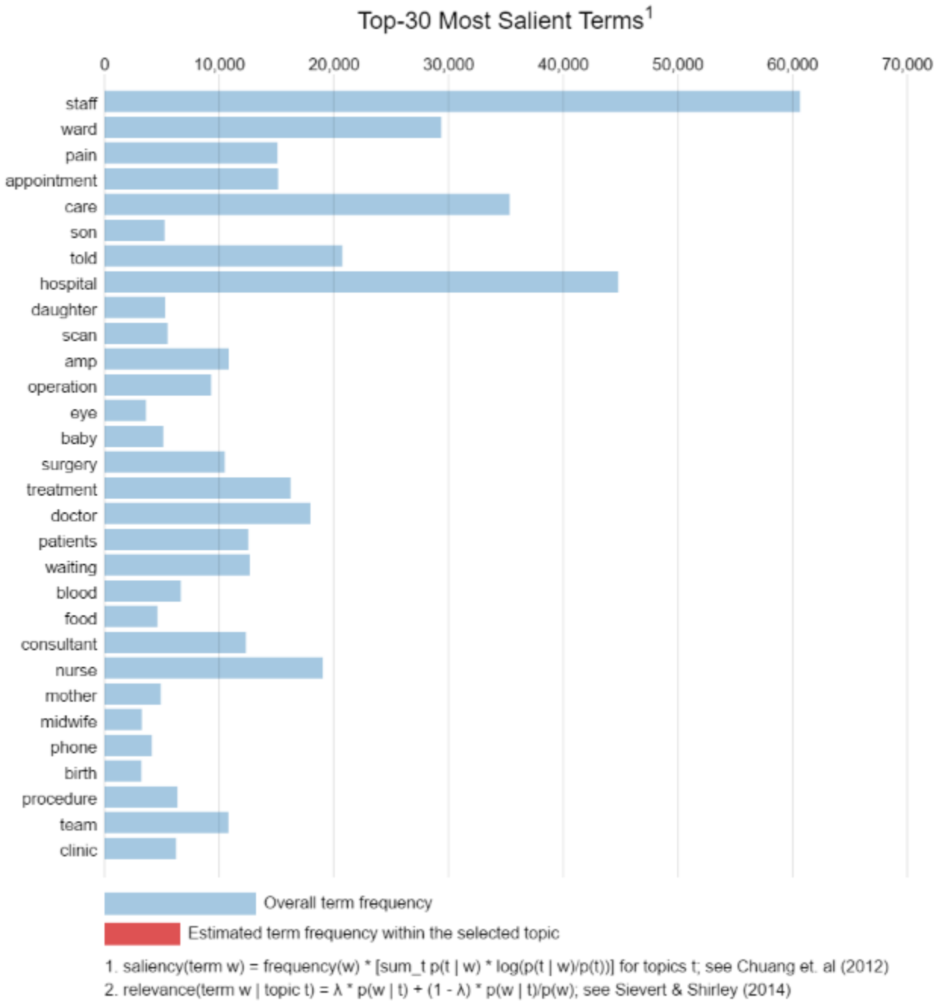


Figure 5-9: An illustration of the top 30 most salient terms from the NHS patient feedback database

Figure 5-9 illustrates the list of the top 30 most salient terms in LDAvis model. The saliency of the terms is calculated using the saliency formula proposed by (Chuang, Manning and Heer, 2012), the saliency of a word is computed as shown below:

$$Saliency(w) = P(w) * \sum_{T} P(T|w) \log \left(\frac{P(T|w)}{P(T)}\right)$$

Where, $P(w)$ is the frequency of occurrence of a word $w, P(T|w)$ is the conditional probability: the likelihood that observed word w was generated by latent topic T. Further, in the Figure 5-9, the blue horizontal bars indicate the overall frequency of occurrence of the corresponding term across the database. The frequency is calculated in terms of its number of total occurrences in the database. For instance, the term 'staff' has the highest frequency with a total amount of over 60,000 occurrences in the entire database.

The visualisation is interactive and when the user hovers over a term, then the 'topic circles' over the left panel of the screen are activated, i.e. those topics in which the chosen term has occurrences are activated. An illustration of this feature is presented in the figures below.

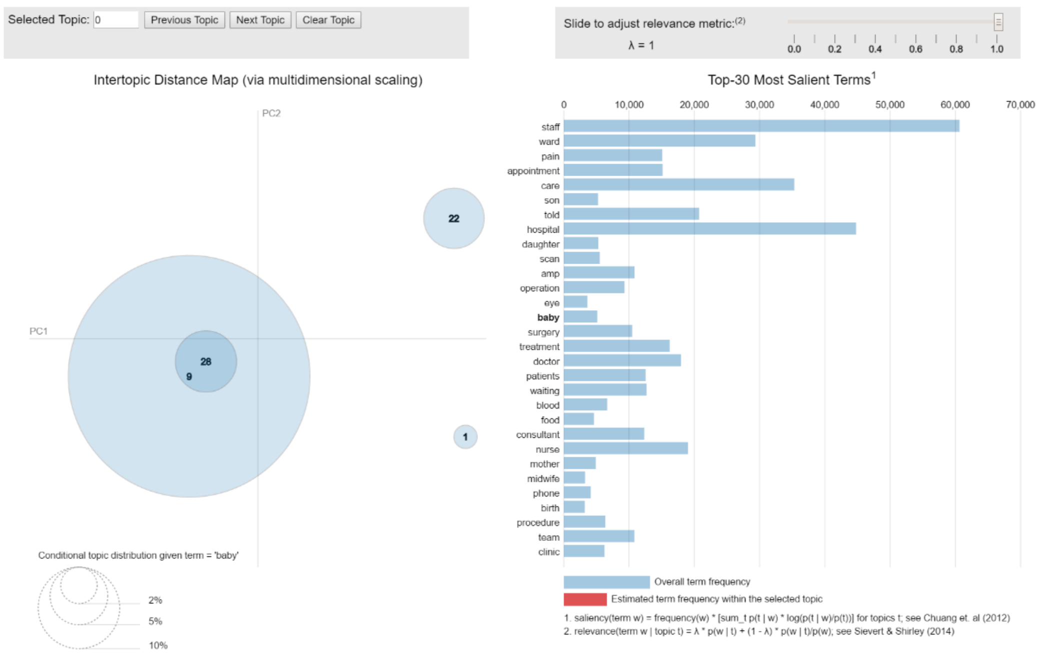


Figure 5-10: A screenshot illustrating the interactivity of the LDAvis topic visualisation

In Figure 5-10, the user hovers over the term 'baby' in the left panel and those topics for which this term has a high occurrence are activated. Further, the dimensions of the circle also change. For instance, topic 9, which pertains to "maternity department", has, as to be expected, higher occurrences of the term 'baby'. Whereas topics 1, 22, and 26 has a relatively far fewer occurrences of the term. To illustrate this feature further, in Figure 5-11 an illustration from hovering over the term 'staff' is shown. It can be observed that the term 'staff' has occurrences across most topics and since it is a generic term this is to be expected. In sum, this feature helps the users to visualise interactively the most frequently occurring terms related to patient experience and to understand its link to individual topics.

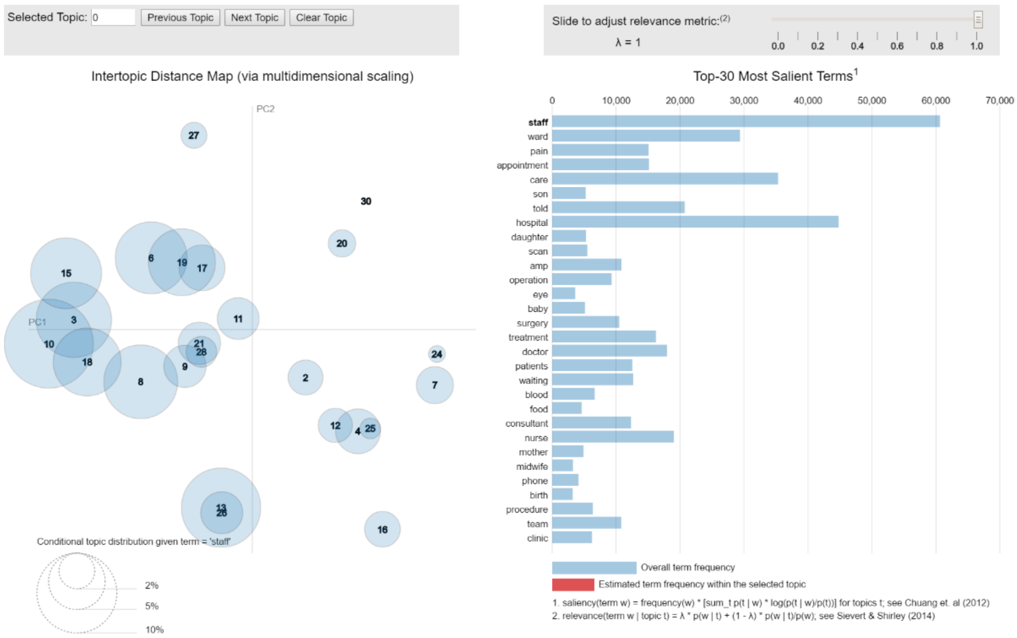


Figure 5-11: A screenshot illustrating the interactivity of the LDAvis topic visualisation

5.5.2 Topic Relevance

The next feature in LDAvis is illustrating the relevance of each topic, which is computed by the equation below and is proposed by Sievert et al. (Sievert and Shirley, 2014).

$$Relevance(w|T) = \lambda * p(w|T) + (1 - \lambda) * \frac{p(w|T)}{p(w)}$$

The relevance of a word w for a topic T is calculated using the above equation, which considers the probability of the occurrence of a word for a given topic and the frequency of its occurrence. An important term in the above equation is the 'lambda (λ)' term, which acts as a weight for calculating the relevance of the word for the topic. Setting the $\lambda = 1$, lists the terms in the decreasing rank of their relevance to the topic and when $\lambda = 0$, the terms belonging to a given topic are listed in terms of the least probability of belonging to that particular topic. This is illustrated in Figure 5-12.

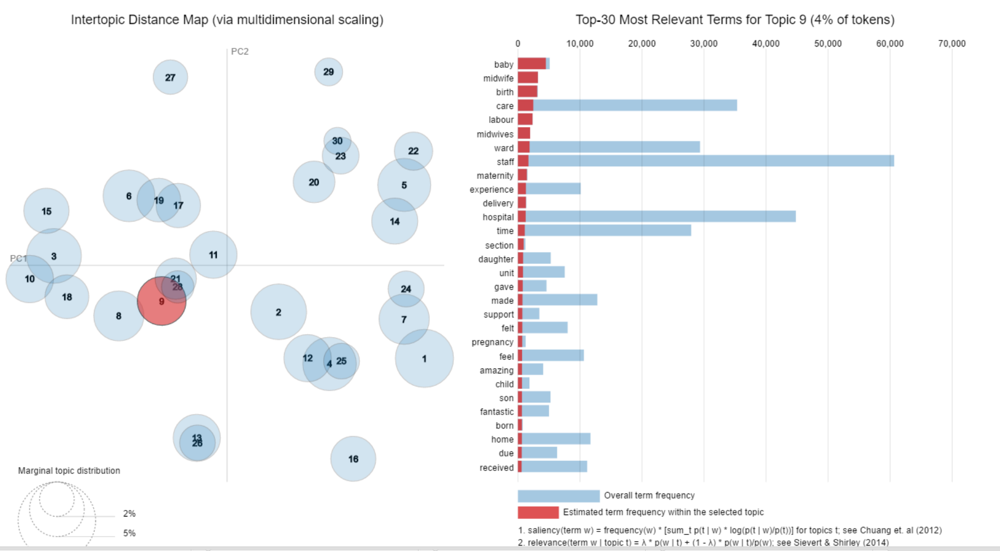


Figure 5-12: Visualisation of the topics, with topic 9 (maternity department) selected on the left with λ=1

In the above figure, topic 9 is selected which pertains to the "maternity department". The right side displays the most frequently occurring terms for this topic (red bar) overlapped on the overall term frequency of the terms (blue bar) for the entire database. In addition, the weight $\lambda = 1$, indicates the most relevant terms belonging to the "maternity department" topic is listed in the right panel and includes terms, such as baby, midwife, and birth, among others.

To illustrate further the effect of the weighting term λ, the value of the weight is set to $\lambda = 0$. It can be seen that in the visualisation a slider is provided on the top of the screen, where the user can set the λ term to a range between 0 and 1. When $\lambda = 0$, the least relevant terms in relation to the "maternity department" topic are listed in the right-side panel. The terms listed in this case include those such as midwives, labour, breastfeed, and others. It can be concluded that because some terms belonging to the theme relating to 'maternity' occur across the database, even though they do pertain to the topic “maternity department”, they are considered to be less relevant to this particular topic than other terms.

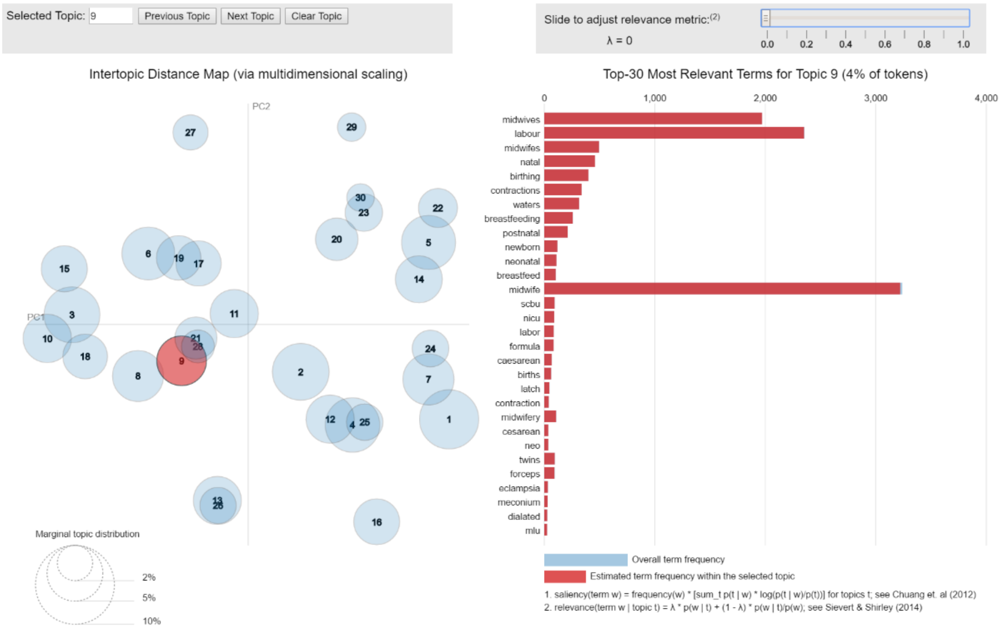


Figure 5-13: Visualisation of the topics, with topic 9 (maternity department) selected on the left with λ=0

The right side displays the most frequently occurring terms for topic 9 (red bar) overlapped on the overall term frequency of the terms (blue bar) for the entire database. Thus, this feature of the LDAvis visualisation enables visualising the most relevant terms and the least relevant terms pertaining to a topic. Further, the users are able to set different weights to the relevance factor and thus, obtain a more detailed understanding of each topic and its relationship with different types of words. For instance, they could find the most relevant terms for a given topic and associate those terms with the sentiment of the topic. This would give them an estimation of aspects of their service that are driving the sentiment in the feedback.

5.5.3 Multidimensional Scaling (Inter-topic Distance Map)

In all the above screenshots illustrating the LDAvis method of visualisation of the identified topics, it can be seen that the topics are distributed across four coordinates in the left panel of the screen. This distribution illustrates the inter-topic distance between different words. In other words, it shows the relationship between the different topics. The topic occurs in clusters, which means that these are likely to talk about similar themes and thus, overlap with each other and to have a shorter distance between them. When two topics are entirely different from each other, then the inter-topic distance between is higher and hence, the greater the distant between one another on the inter-topic distance map. This is illustrated in the figure below (reproduced from Figure 5-8).

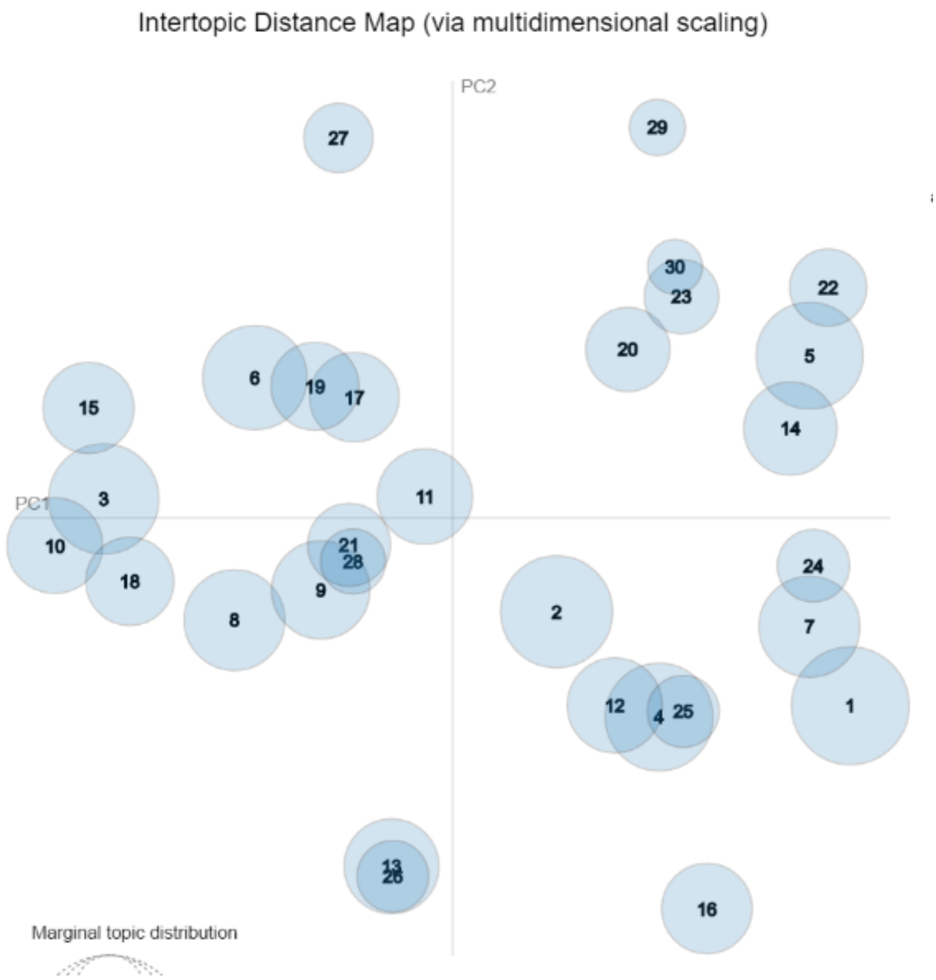


Figure 5-14: Illustration of each topic's distribution across the database

As can be seen in Figure 5-14, the topics are distributed across the four coordinates. The distribution is based on the distance between the topics. The distance between the topics is computed by using a multidimensional scaling technique, namely Principal Component Analysis (PCA), which enables the projection of inter-topic distances onto two dimensions. That is, the multidimensional scaling technique is a dimension reducing technique that aids in visualising high dimensional data. Further, multidimensional scaling also allows for computing the distance matrix between different topics and is widely implemented in information analysis.

In Figure 5-14, it can be observed that there are clusters of topics that overlap. For instance, the inter-topic distance between topic 9 and topic 28 is very close. These topics correspond to the "maternity department" and "paediatric visits". Since these two topics relate have similar themes in that they relate to childcare, they are more likely to have the occurrence of similar terms and hence, have lesser inter-topic distance. Similarly, there is a cluster for topic 13 and topic 26. These two topics correspond to "discharge lounge" and "admission ward". Because these two topics relate to the theme of a patient entering and leaving the hospital, they are likely to have similar terms and hence, lower inter-topic distance. This approach would provide the users with an option to examine a cluster of topics and then analyse the terms related to the topic clusters. Thus the users are able to segregate topics and also at the same time be able to analyse the topics and its related terms as a cluster.

Dimensionality Reduction

The multidimensional scaling methods, especially PCA, enable reducing the dimensionality of high dimensional data and thus represent the data with the more prominent components, which are also known as the principal components. Using the LDAvis method, it is possible to reduce further the topics by eliminating those that are similar or those that tend to overlap with each other. This type of topic reduction facilitates the identification of more distinct topics from the database and involves merging those containing many similar terms.

The dimensionality reduction is illustrated in the figure below, where the topic visualisation is carried out for the bigram model, with all the 30 topics identified by the LDA model being shown. In the next figure, dimensionality reduction is applied, which reduces the number of topics from 30 to 22 for the visualisation. In Figure 5-15, it can be seen that the inter-topic distance between the 30 bigram topics is not very high and there are several instances of overlap between topics, such as topics 20 and 23, topic 25 and 26, amongst others.

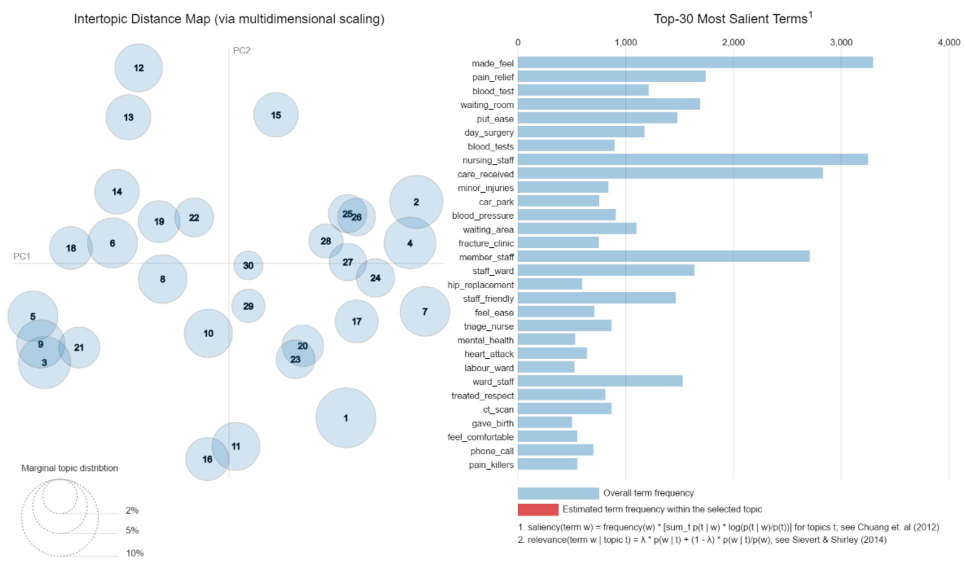


Figure 5-15: A screenshot of LDAvis visualisation of the 30 topics identified using the bigram LDA model

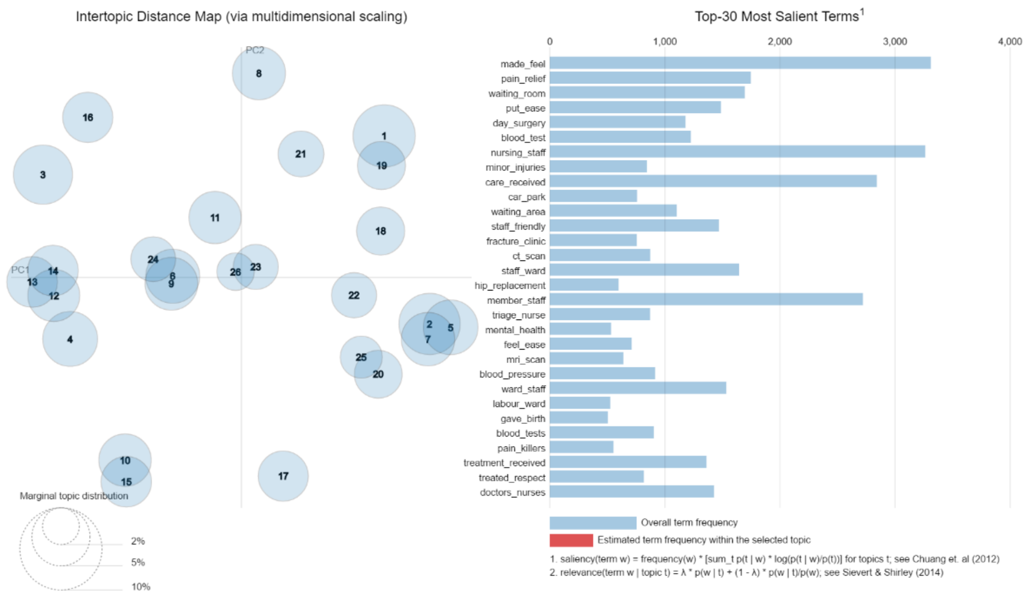


Figure 5-16: A screenshot of LDAvis visualisation after dimensionality reduction of the 30 bigram topics to only 22 topics

In Figure 5-16, the dimensionality reduction has reduced the number of bigram topics from 30 to 22 topics. This reduction reduces the topics that tend to overlap and have shorter inter-topic distance. In the distribution of these two 22 topics, it can be noted that there is an overall higher inter-topic distance between the topics when compared with the 30 topics shown in Figure 5-15. Thus, dimensionality reduction enables the user to identify more distinct and unique topics. Consequently, this results in better representation of the topics in the database because the overlap between the topics is lesser and more distinct. When similar themes are represented in large numbers rather than it may overwhelm the analysts and make analysis more difficult. Thus, being a smaller number and unique would make the topic analysis easier and more engaging.

5.6 Human Judgement Survey

To evaluate further the performance of unigram and bigram Topic Modelling in identifying “sensible” topics, a human judgement survey was conducted. The survey participants were NHS staff, who were asked to analyse the words identified for each topic from both the unigram and bigram modelling and to rate their score depending to what extent the words pertaining to a given topic was sensible. If they considered that, the words that were part of a given topic were highly sensible then they scored them as 5 or else they provided a score of 1, if they were not deemed sensible. Five participants in the survey provided their responses. A screenshot of part of the survey questionnaire is shown in Figure 5-17. It can be observed in the screenshot that for each topic, the words belonging to the topic are listed for the users. The staff reviewed the suitability of the words for each topic and rated whether the group of these words were sensible for a single topic. This survey was done to cross validate the terms identified for the labelled topics and to verify whether the terms and the topics labelled would be meaningful from the perspective of the clinical staff. Similarly, the questionnaire also included the word pairs for the bigram topics.

1. Unigram Sets :

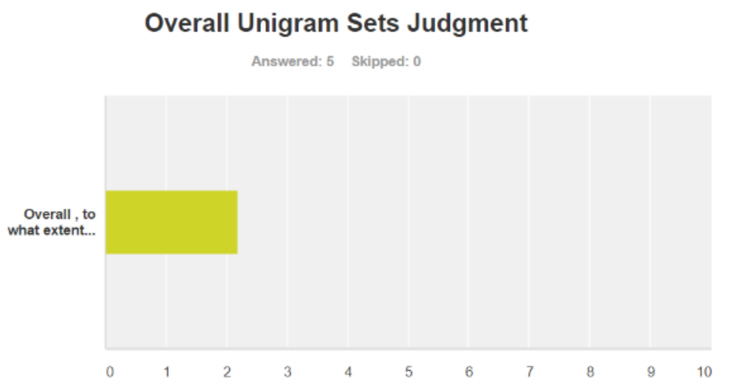
	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
1- {birth, baby, midwives, midwife, labour, maternity, care, delivery, ward, staff, pregnancy, amazing, section, experience, daughter, support, fantastic, son, child, born, natal, felt, antenatal, unit, recommend}					
2- {pain, told, doctor, antibiotics, nurse, hours, relief, home, infection, gp, severe, medication, back, blood, asked, night, dr, ambulance, painkillers, didnt, morphine, chest, agony, worse, days }					
3- {im, dentist, dont, ive, dental, tooth, didnt, told, pain, feel, teeth, life, mental, people, back, health, wisdom, wasnt, id, worse, wrong, ill, bad, years, hospital}					
4- {operation, staff, replacement, hip, knee, care, excellent, surgery, surgeon, op, ward, treatment, hospital, consultant, recommend, team, nursing, stay, hall, received, food, recovery, pre, physio, post}					
5- {blood, scan, test, tests, results, told, doctor, ct, gp, ultrasound, pregnancy, pregnant, mri, nurse, ecg, weeks, bleeding, miscarriage, pressure, asked, bloods, scans, wait, back, appointment }					

Figure 5-17: A screenshot of the survey questionnaire

The responses were obtained on the scale of how rational they thought the words were for each topic. They were asked to decide the rational based on their opinion about the relevance of the terms for a given topic title, i.e. if they find that the terms listed for each topic belong to the topic then they rate it higher and so on. A screenshot of the responses is provided in Figure 5-18. Complete results are provided in the Appendix A. It can be seen that most responses were that the words were rational for each topic. A summary of the results for all topics was obtained by asking the staff "Overall, to what extent would you agree/disagree that the Unigram Sets have been classified sensibly?" The responses obtained are shown in Figure 5-19. 4 out of 5 participants agreed that the unigram topics were sensible and one participant neither agreed nor disagreed.

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)	Total	Weighted Average
1- {birth, baby, midwives, midwife, labour, maternity, care, delivery, ward, staff, pregnancy, amazing, section, experience, daughter, support, fantastic, son, child, born, natal, felt, antenatal, unit, recommend}	20.00% 1	60.00% 3	20.00% 1	0.00% 0	0.00% 0	5	2.00
2-{pain, told, doctor, antibiotics, nurse, hours, relief, home, infection, gp, severe, medication, back, blood, asked, night, dr, ambulance, painkillers, didnt, morphine, chest, agony, worse, days }	0.00% 0	40.00% 2	60.00% 3	0.00% 0	0.00% 0	5	2.60
3- {im, dentist, dont, ive, dental, tooth, didnt, told, pain, feel, teeth, life, mental, people, back, health, wisdom, wasnt, id, worse, wrong, ill, bad, years, hospital}	0.00% 0	60.00% 3	40.00% 2	0.00% 0	0.00% 0	5	2.40
4- {operation, staff, replacement, hip, knee, care, excellent, surgery, surgeon, op, ward, treatment, hospital, consultant, recommend, team, nursing, stay, hall, received, food, recovery, pre, physio, post}	0.00% 0	60.00% 3	40.00% 2	0.00% 0	0.00% 0	5	2.40
5- {blood, scan, test, tests, results, told, doctor, ct, gp, ultrasound, pregnancy, pregnant, mri, nurse, ecg, weeks, bleeding, miscarriage, pressure, asked, bloods, scans, wait, back, appointment }	20.00% 1	80.00% 4	0.00% 0	0.00% 0	0.00% 0	5	1.80
6- {food, ward, staff, room, stay, tea, hot, clean, meals, patients, eat, good, meal, menu, choice, night, coffee, drinks, tv, offered, drink, cup, water, breakfast, bed}	20.00% 1	60.00% 3	20.00% 1	0.00% 0	0.00% 0	5	2.00
7- {waiting, told, wait, nurse, room, doctor, hours, reception, asked, receptionist, rude, hour, waited, minutes, patients, people, area, desk, called, appointment, pm, quot, sat, didnt, triage }	0.00% 0	80.00% 4	20.00% 1	0.00% 0	0.00% 0	5	2.20
8- {eye, clinic, eyes, vision, cataract, drops, sight, consultant, treatment, operation, doctor, optician, moorfields, appointment, laser, problem, surgery, department, retina, condition, ophthalmology, glaucoma, left, glasses, prescription }	25.00% 1	75.00% 3	0.00% 0	0.00% 0	0.00% 0	4	1.75
9- {procedure, ease, staff, made, feel, explained, nurse, friendly, lovely, nervous, endoscopy, felt, professional, doctor, experience, put, comfortable, kind, reassuring, colonoscopy, caring, questions, team, nurses, appointment}	0.00% 0	80.00% 4	20.00% 1	0.00% 0	0.00% 0	5	2.20

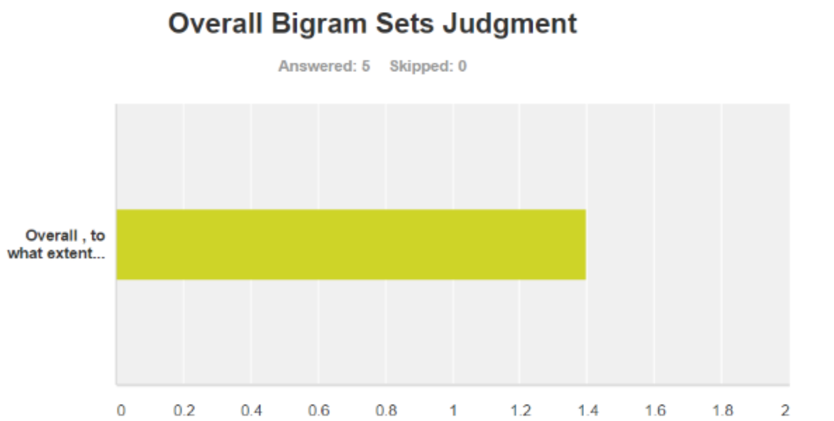
Figure 5-18: A screenshot of responses obtained for the unigram topics



	Strongly agree	Agree	Neither agree nor disagree	Disagree	Stronglydisagree	Total	Weighted Average
Overall , to what extent would you agree/disagree thattheUnigram Sets has been classifiedsensibly ?	0.00% 0	80.00% 4	20.00% 1	0.00% 0	0.00% 0	5	2.20

Figure 5-19: A screenshot of participant responses towards the overall unigram judgement

The same questions were asked for the bigram topics and the responses obtained for this method were different from the unigram method. When asked how strongly they agreed with the topics identified by the bigram method, 3 of the participants strongly agreed with the topics identified and 2 agreed that they made sense, as shown in Figure 5-20. That is, the human judgement survey showed that the NHS staff found the bigram Topic Modelling to be better performing in terms of identifying more sensible topics than the unigram method.



	Strongly agree	Agree	Neither agree nor disagree	Disagree	Stronglydisagree	Total	Weighted Average
Overall , to what extent would you agree/disagree thatthe Bigram Sets has been classifiedsensibly ?	60.00% 3	40.00% 2	0.00% 0	0.00% 0	0.00% 0	5	1.40

Figure 5-20: A screenshot of participant responses towards the overall bigram judgement

5.7 Discussion

In this chapter, the study presented showed that Topic Modelling models could be applied to identify topics and themes from the patient database automatically. The outcomes of this study indicate that a better understanding of commonly occurring themes and topics from the patient feedback can be identified. In the study, both unigram and bigram models were used to identify 30 different topics from each model. The LDA approach provided a list of 'words' that are most likely to belong to a particular group. The word list was then manually analysed and provided a label. That is, three NHS doctors were asked to provide a label to each group of listed words. The doctors drew up possible word list labels individual and then, came to a consensus through a group discussion. The process of involving doctors to provide label suggestions and then validating the topic models ensured that the topics identified were cross verified by clinicians as they are better equipped to decide about then. One limitation of the study is that the validation of the topic modelling process can be done in a qualitative approach (i.e. use of expert survey) and to the best of our knowledge there are no quantitative method to measure the performance of the topic modelling approach.

A significant result of this study is merging the Sentiment Analysis aspect of the SVM model with the Topic Modelling approach. In the previous chapter, the performance of SVM in identifying the sentiment of patient experience from the database was shown. The SA outcomes of the SVM model were then used to predict the sentiment of the topics identified from the Topic Modelling. To the best of this researcher's knowledge, this is a novel contribution in this study, as there is no literature that has applied this approach to predict patient sentiment from the topics identified in the comments.

The possibility of predicting topic sentiment is a research matter that needs to be further explored. The results of the study, as demonstrated in Subsection 5.4.1 and Figure 5-3, have shown that predicting topic sentiment from the comment sentiment predicted by the SVM model can lead to a higher level of accuracy. Finally, the LDAvis visualisation method of visualising Topic Modelling outcomes have been demonstrated as being an effective way for understanding and further analysing the topics of patient experience. In particular, the LDAvis approach reduces the effort of manually looking through the tables of word lists to get an understanding of the topics. Moreover, the relevance and specificity features of the LDAvis method enable a deeper, flexible and quick way of analysing the topics identified.

To conclude, the learning outcome of this study is that Topic Modelling methods can be very useful tools for identifying topics hidden in patient feedback about the hospital they have attended. Further, merging Sentiment Analysis methods with Topic Modelling can also be a tool in predicting the sentiment of the topic from the patient feedback. The outcomes of this study are used in the next iteration of the study, where patient feedback further analysed using the topics identified in and applying Dependency Parsing on each feedback item to estimate the rationale behind the patient sentiment for a given topic.

5.8 Summary

This study has shown that Topic modelling methods can be a useful tool to identify topics from a patient feedback dataset and that the topics identified were relevant to the users. Further, the sentiment behind each topic was estimated by applying the outcomes of the first iteration of the study. A topic model visualisation method was also presented that allows the user to visualise the topics identified from the dataset, its distribution in the database, and the relation between the topics and the terms associated with each topic.

Chapter 6: Dependency Parsing to Identify the Rationale of Patient Sentiment

6.1 Overview

The analysis of patient experience data was taken forward from just binary sentiment classification of a review to the identification of topics and the associated sentiment of each topic in the review. The study is further extended by addressing the question "what are the possible reasons behind a patient's sentiment for a particular topic in a given review?". This chapter is a continuation of the work on Sentiment Analysis of the patient experience. The study presented here explores the Natural Language Processing (NLP) technique of Dependency Parsing (DP) for facilitating an automated process to interpret the patient experience data further in relation to finding the potential reasons behind the sentiment of the patient.

The chapter is structured as follows. Section 6.2 presents a brief overview of the NLP Dependency Parsing concept. The materials and methods used for this iteration are presented in Section 6.3 followed by results analysis in Section 6.4. The discussions of the findings of the study are discussed in Section 6.5.

6.2 NLP Dependency Parsing

From the previous two iterations, it has been made possible to understand the topics potentially discussed by the patients about their experience in the comments as well as identifying the sentiment present in the topic being discussed. Building on from the previous iterations, a method for identifying the reason behind a particular sentiment for a topic in a patient comment is applied. Specifically, the focus is on automatically finding the reason for why a patient is happy or unhappy about a particular service provided by the hospital. Identifying the reason behind the sentiment makes the process of sentiment identification more robust in that it will deliver more information to the hospital that can be used to enhance the services provided to patients.

To achieve the aims of this study, the Dependency Parsing (DP) approach of NLP is applied as a method to identify the reason behind a patient's sentiment. Dependency Parsing is a method where the grammatical structure of a sentence is analysed to identify the relationship between different words in the sentence. The principle behind the "dependency" is that a sentence is such that the syntactic structure has binary asymmetrical relations between the words in it. In NLP Dependency Parsing, a dependency relation is established in a sentence by defining entities called the *head* and *dependent* (Nivre and McDonald, 2008). Then, based on the relationship between the words in terms of grammatical structure, it identifies the subject, object, verb, nouns, and other syntaxes in the sentence. Based on the identified grammatical structure, parse trees termed dependency trees, are created to analyse the sentence structure. The dependency trees then act as decision trees to analyse and make decisions about the sentence structure (Koo, Carreras Pérez and Collins, 2008).

In Figure 6-1, an illustration of the dependency in a sentence is given. It can be observed that there is a dependency between the different "parts of speech". For instance, in this case, the word "is" can be considered as the head of the sentence, which has dependents, i.e. the noun "example" and the pronoun "this". Similarly, these two dependents can further have dependents in the sentence as illustrated in the graph. Identifying such a head-dependent pair enables the identification of the different links between the parts of speech of the sentence in a given pair. The benefit of this would be that the sentence can be further broken down into noun phrase, verb phrase, adjective phrases and automatically identify the dependency between these phrases within the sentence.

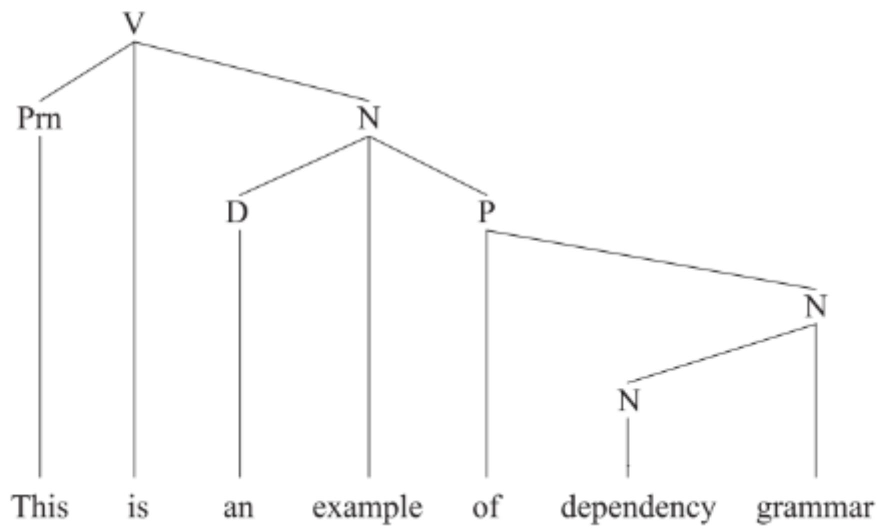


Figure 6-1: A *dependency tree* describing the head (is) and its dependants in a sentence (Sievert and Shirley, 2014).

After identifying the different parts of speech in a sentence, as seen in the above diagram, then the different dependencies within it can be identified. This can be further utilised to gain knowledge of the context of the sentence. For the study, the DP approach was used to identify the "parts of speech" in a given sentence and the relationship between the parts to further use them as a means to estimate the possible reason behind a particular sentiment in the sentence. In the next section, the approach followed to achieve this is described.

6.3 Materials and Methods

Two approaches of NLD Dependency Parsing were undertaken. In the first approach, the Apache openNLP Dependency Parsing method was applied, whilst in second, the Stanford NLP Dependency Parsing was utilised. Two different approaches were used and tested to get the best results for summarising the reviews. Both the methods are discussed below.

6.3.1 Apache openNLP approach

The Apache openNLP is machine learning based toolkit for natural language processing. It provides various language-processing functionalities, such as sentence segmentation, parts-of-speech tagging, parsing and co-reference resolution (Foundation, 2016). For this research, the parts-of-speech (PoS) tagging functionality was used to identify the (PoS) in the sentence. The steps followed while implementing this method are illustrated below.

The first step was to load the PoS dictionary and the sentiment dictionary into the R environment. The PoS dictionary available in the openNLP source identifies the parts of speech in a sentence and tags them accordingly. Prior to identifying the PoS, the openNLP model generates tokens from the sentences. Tokens are words, punctuation and numbers in the text that are separated from the algorithm for further processing. Once the tokens are identified, the PoS tagger from the openNLP model marks the tokens by the identifying the word type corresponding to the token and considers the context of the token. The PoS tagger decides the type of the word or PoS for each tag based on a probabilistic model, which analyses the context of the token and then chooses a PoS tag. The implementation of the PoS dictionary is shown below.

```
## Read POS ditionary
pos_dict <- read.csv("EnglishDictionaryPOS.csv")

## Read Sentiment Dictionary
sent_dict <- read.csv("Dictionary-11-4-v2.csv")
```

Along with the PoS tag dictionary, the sentiment dictionary was also loaded, which are the Sentiment Analysis outcomes derived from the study presented in Chapter 4: The sentiments for each review were then associated with the tags identified and this is discussed later. Before processing the patient feedback text corpus for PoS tagging, the corpus was pre-processed or “cleaned” to remove numbers, whitespaces, and punctuation. The clean corpus was then rearranged to generate a unigram term document matrix.

On the clean and rearranged corpus, the PoS tagging function of openNLP was applied that performs the "part of speech" tagging of the corpus. The PoS tagging process leads to the extraction the nouns and adjectives from the documents. The nouns were ranked based on the frequency of their occurrence in the review. The rationale behind this is that higher frequency nouns are more likely to be the "subject" that is being discussed by the patient in their feedback. Further, along with the nouns, the adjectives were also extracted from the reviews. The adjectives were extracted because they can be the indicators of the reason behind the sentiment. For instance, a positive adjective such as 'wonderful' would be an indication of a positive sentiment.

In the next step of the NLP process, the sentiment scores obtained from the previous SA study were utilised. Using the positive and negative sentiment dictionary, the adjectives were ranked according to the strongest sentiment. Thus, noun and adjective pairs ranked according to the frequency of their occurrence and strength of their sentiment were materialised. Using the extracted nouns and adjectives, one noun-adjective pair was obtained for each review, this noun being the most frequent noun regarding the subject being discussed and the strongest adjective representing sentiment, either positive or negative, thereby identifying the possible reasons behind this sentiment. A screenshot of extracted noun and adjective pairs from *RStudio* IDE is shown in Figure 6-2.

Comments review	all_pairs	open_nlp_pair
I would like to thank the	touch->personal ; experience->interesting ; experience->informative	staff,enjoyable
I would just like to thank	treatment->further	hospital
If the above was not bad	hand->left ; time->first ; reaction->allergic ; line->subject ; staff->medical ; teenager->tall ; expe	staff,amazing
The receptionists/clerica	staff->receptionists/clerical ; muddle->total	staff,helpful
My consultant and the re	consultant->brilliant ; team->supurb	consultant,brilliant
The nursing and support	matter->difficult	staff,excellent
On the whole, I cannot re	project->new ; environment->clean ; environment->sterile	hospital,clean
i came in with a foot inju	day->next ; 7month->broken ; service->good	day,good
When someone you kno	nature->human ; idea->good	hospital,good
I was given pain relief wl	visit->second	pain,second
From being admitted int	staff->ongoing	staff,excellent
My mother was addmitte	bed->available ; hospital!Eventually->whole ; face->clear ; face->english ; afternoon->late	staff,good
I went for my 12 week sc	story->long ; pain->much ; ward->dark ; week->last ; week->few	hospital,rude
This is my first experienc	experience->first ; number->large ; basis->daily ; problem->minor	hospital,minor
Massive praise to the sta	praise->massive ; Unit->short ; kind->efficient ; hand->safe	staff,efficient
i was told by my doctors	relief->strong	hospital,happy
Marvelous surgical team	team->surgical	teams,marvelous
The hospital offers the s	effect->side ; operation->previous	hospital,previous
It may be due to the age	patient->other ; difficulty->personal ; care->person-centred	staff,due
The phone system is just	hospital->other	hospital,stupid
Thank you to everyone a	work->good	work,good
Having a local hospital w	hospital->local	hospital,important
A very negative experien	experience->negative	experience,negative
I am very cross at the wa	hospital->good	staff,good
Thank you to all the staff	job->excellent	staff,excellent
A comment about hospit	charge->additional	hospital,additional
After the procedure I wa	patient->other ; patient->male ; patient->female ; patient->poor ; patient->high ; procedure->e	patients,poor

Figure 6-2: Extracted noun and adjective pairs using OpenNLP approach

An example to illustrate this process is presented below. One comment from the patient feedback database was a follow:

> "*Yesterday, Sunday February 7th I was seen in the Hillingdon Hospital emergency room. I had cut my hand open and it appeared to me that I would need medical assistance because I am not a trained doctor or nurse. I thought I would need stitched because the cut was very deep but the nurse, who I might add was very very rude to me, said it was a superficial cut. While the nurse was "helping" me, the nurse was rude and was asking me why I was crying when the nurse was bandaging and cleaning my cut. The nurse also made a comment to me about how I expect to have children if I can't handle this pain. That was much uncalled for and had nothing to do with anything. I found the nurse to be very rude and not very professional at all. I felt worse when I left than when I came in. I understand the nurse has to do their job and clean my wound and that it may hurt. The nurse did not give me any recommendations for pain relief in terms of medicine or anything. They just bandaged me/made me feel worse and then made me and my husband leave. To say the least I was very disappointed in the level of treatment I received. I don't even think I*

> *can express it enough. Now there probably isn't anything that can be done, but by the way they treated me, a young woman, I imagine they have treated others badly as well. I just want it to be known that I was not treated with any respect and maybe some kindness training for the nurses should be given. No human being would have treated me the way they did*."

When the adjective-noun pairs are extracted, the Dependency Parsing outcomes provide the following words:

"*medical, assistance, trained, doctor, superficial, cut, young woman*"

Next, the nouns and the adjectives are ranked according to their occurring frequency and sentiment strength. Thus, the noun-adjective pairs would be:

Noun-adjectives pair: *nurse-rude, doctor-trained, cut-superficial, children-assistance*

The sentiment scores are used to label the sentiment of each word in the review and the most frequently occurring ones are extracted to get a word pair that can potentially describe "why" a particular sentiment is given to a review. The noun and adjective pairs extracted using the openNLP method enable identification of the reason behind a particular sentiment in a review. Despite the openNLP method providing the noun-adjective pair that can be used for further analysis, there is no guarantee that it is directly linked. At times, the adjective linked to a noun might not belong to the linked noun but may belong to another subject or noun in the sentence. For example, from the sentence "*the nurse was not as wonderful as the doctor"*, the adjective may get wrongly associated to nurse instead of or along with the doctor. Hence, the adjective extracted might not necessarily describe the exact noun when the sentence is parsed by the tool. Further, the openNLP method does not take into consideration negation words, such as "not" that would reverse the sentiment polarity and hence, provide inaccurate estimation in some instances.

6.3.2 Stanford coreNLP Approach

For a deeper understanding and automated analysis of the patient feedback, the Stanford university`s coreNLP suite utilised for the study. This is a suite of open-source NLP tools provided by Stanford University. It has a host of functionalities including base forms of words, their parts of speech, their identities, such as whether they are names of an organisation or country and it marks up the structure of sentences in terms of phrases and dependencies, among other things. An important feature provided in the coreNLP suite is the Dependency Parsing aspect. It can identify which noun phrases refer to the same entities, the association between them and indicate the sentiment referred to by it. The coreNLP suite was applied to the patient feedback database to get a deeper understanding of the link between the adjective and the noun that was reliable. Similar to the process followed for openNLP implementation, cleaning of the corpus was first performed. The coreNLP was also implemented in an R environment.

The first step was defining the function to fetch noun-adjective pairs. The function was then applied to the corpus to get the best noun-adjective pair, as shown below. Once the noun and adjectives had been extracted, using the coreNLP method, as with the openNLP process, the sentiment scores obtained from the SoA Sentiment Analysis method presented in Chapter 4: were used to find the sentiment of each adjective. Next, based on the sentiment score, the adjectives were ranked according to the strongest sentiment. Thus, by using the coreNLP approach pairs of noun-adjective were extracted in which the adjectives are linked to the noun or subject in the text. These noun-adjective pairs can be used to summarise the review and to identify the rationale or reason behind a patient`s sentiment for a particular service provided by a hospital. A screenshot of extracted noun and adjective pairs from *RStudio* IDE is shown in Figure 6-3

text	all_pairs	core_nlp_pair
I recently spent 3 (	trouble->much ; Crowther->fabulous	staff,fabulous
I had an accident a	mum->single ; life->daily ; pain->much ; day->next ; day->next ; health->mental ; medication->otl	hospital,worse
the nurses at the t	visit->recent ; assessment->pre-op ; month->past ; department->gynicological	nurse,wonderful
thanks to the care	problem->correct ; study->full ; cardiologist->cosultant	care,correct
I was in the MAU V	night->last ; nurse->same ; nurse->friendly	staff,friendly
I cant thank labour	boy->little ; thing->only ; layout->new ; mum->first ; night->first ; coz->nice ; thing->red	time,fantastic
I was diagnosed w	condition->rare ; year->last ; experience->whole	staff,helpful
My mother was tra	person->fantastic ; year->many ; year->more	staff,fantastic
An absolute disgra	disgrace->absolute ; pain->excruciating ; attitude->same ; ward->particular	staff,appalling
I visited southtyne	complaint->simple ; line->front ; complaint->simple ; reception->main ; complaint->simple ; time	hospital,dirty
I was recently refe	unit->urgent ; assessment->whole ; route->best ; assessment->psychiatric ; home->own ; damage	time,urgent
My dad was in Her	month->half ; admiration->great ; person->whole ; symptom->medical	staff,friendly
After a chaotic adr	admittance->chaotic ; afternoon->mid ; surgery->elective ; removal->partial ; awareness->first ; w	staff,clean
My mother was tal	treatment->obvious ; dehydration->whitby?continued ; situation->avoidable	staff,better
The attention, trea	Assessment->medical ; unit->cardiac	care,excellent

Figure 6-3: Extracted noun and adjective pairs using CoreNLP approach

In the next section, the results obtained by using the above two Dependency Parsing models are discussed.

6.4 Results Analysis

As indicated in the previous section, the outcomes of using Dependency Parsing with the above two NLP methods is extracting noun-adjective pairs. It was also mentioned how the sentiment scores obtained from the SoA Sentiment Analysis models were applied to rank the adjectives and associate them with the nouns or subjects occurring in the patient feedback.

Further, the outcomes obtained from the Topic Modelling study were used to associate the noun-adjective pair with the topic identified from the Topic Modelling methods. For this study, the topics identified from the 2-gram study were included as the topics identified from that model provided a better topical representation of the review topic.

The Dependency Parsing was applied on the entire *NHS choices* patient feedback database consisting of 76,151 comments. For each patient comment, the noun-adjective pair for both the models was identified and the pairs were associated to the topics identified from the 2-gram Topic Modelling approach. It is worth mentioning that, this experiment process has taken about 3 days on 8 core i7-4970 server to be perform successfully. In the following sections, the outcomes of both, openNLP and coreNLP approaches, are provided.

6.4.1 openNLP Approach Results

The Table 6-1 below shows the noun-adjective pairs extracted for each comment. Results are shown for only 20 comments. Similar results were obtained for all the comments present in the database. The noun and adjective pairs ranked according to the frequency of their occurrence and strength of their sentiment were materialised.

Table 6-1: The noun-adj pair extracted using for each comment from the patient feedback database using the openNLP method.

Comment number	*Topic number*	*Topic Name*	*Noun-adj pair*
1	12	Positive procedural experience	staff, enjoyable
2	23	Surgery room	hospital, clean
3	13	Children visits	staff, amazing
6	21	Staff communication	nurse, professional
7	4	Parental hospital visits	staff, helpful
8	15	Maternity	consultant, brilliant
12	6	Description of quality of care	staff, excellent

13	18	Serious disease and cancer	hospital, clean
14	26	Fractures and minor injuries	day, good
15	28	Staff ward	staff, excellent
16	7	Nurse experiences	staff, good
17	7	Nurse experiences	hospital, rude
18	10	Staff comments	hospital, minor
19	6	Description of quality of care	staff, efficient

In the above table, it can be seen that the noun-adjective pairs obtained from the comment may enable understanding of the reason behind a particular piece of feedback of the patient. For instance, for comment 8, the comment is as below:

"…My consultant and the rest of the maternity team deserve medals. He is a brilliant consultant with a superb team under him.…"

The noun-adjective pair extracted is "consultant, brilliant". The noun-adjective pair were extracted by combining them according to the frequency of their occurrence and strength of their sentiment.

As another example, comment 1 is presented below:

"*…I would like to thank the staff (doctor & nurses) who carried out my procedure for their care, and also the personal touches that made the experience interesting and informative as well as, while not enjoyable, something that I would have no fear of having to undergo again…*"

The noun-adjective pair extracted is “staff, enjoyable”. As can be seen, whilst “enjoyable” is chosen as the adjective representing the strongest sentiment based on the SA dictionary, “staff” is extracted owing to the frequency of its occurrence in the dataset. It should be noted that the subject being referred to as “enjoyable” is about the experience and not the “staff”. Further, it can also be clearly noted that the “enjoyable” has a negation “not” which the openNLP method has not considered which is a drawback. Thus, these drawbacks of the openNLP implementation approach, as there is no guarantee that the adjective extracted refers to the correct noun.

To overcome this performance limitation of the openNLP approach, the Stanford`s coreNLP is applied.

6.4.2 coreNLP Method

For extraction of more suitable noun-adjective pair, the coreNLP suite is used that not just parses the sentence for the adjective but also connects the adjective to the right subject (noun) to which the adjective refers to in the review. The noun-adjective pair shown here is the pair, which has the strongest sentiment score based on the sentiment scores obtained from the SoA model. The outcomes of 15 reviews randomly chosen from the corpus are shown in Table 6-2.

Table 6-2: The noun-adj pair extracted using for each comment from the patient feedback database using the coreNLP method.

Comment number	Topic number	Topic Name	Noun-adj pair
1	12	Positive procedural experience	touches, interesting
2	15	Maternity	consultant, brilliant

3	6	Description of quality of care	opinion, excellent
4	26	Fractures and minor injuries	doctor, good
5	28	Staff ward	registrars, excellent
6	7	Nurse experiences	disappointment, admitted
7	7	Nurse experiences	nurse, professional
8	10	Staff comments	problems, large
9	6	Description of quality of care	staff, massive
10	3	Parking infrastructure	parking, difficult
11	28	Staff ward	food, excellent
12	23	Surgery	team, marvellous
13	14	Phone interactions	response, stupid
14	10	Staff comments	players, excellent
15	10	Staff comments	hospital, professional

If the noun-adjective pair extracted from the coreNLP method is compared with the openNLP method listed in Table 6-1, it can be noted that the pairs are different on several occasions. For instance, the noun-adjective pair extracted for comment 1 from the coreNLP pair is “touches, interesting” as opposed to “staff, enjoyable” extracted by the openNLP method. Similar pairs of noun-adjective pairs extracted using the coreNLP method provides a summarisation of the patient review when compared to openNLP method.

For example, the comment 3 reads:

"*…The nursing and support staff were in my opinion, excellent. They dealt with difficult matters with sensitivity and as a team…*"

The noun-adjective pair extracted by coreNLP method is "opinion, excellent". This pair provides a good summary of the review indicating that the opinion has been excellent by the patient.

6.4.3 Performance Validation

In both the above sections, it was shown that the noun-adjective pair extraction was carried out by two processes. A challenging question is; how can the performance of the Dependency Parsing methods be validated?

Effectively, both the Dependency Parsing methods have enabled us to summarise a review via a noun-adjective pair. Summarisation is a way of taking a long utterance of sentence and represents it in a very short utterance. In the current study, the summarisation pertains to a pair of words that potentially describe the strongest sentiment present in the review and the subject to which the sentiment refers to.

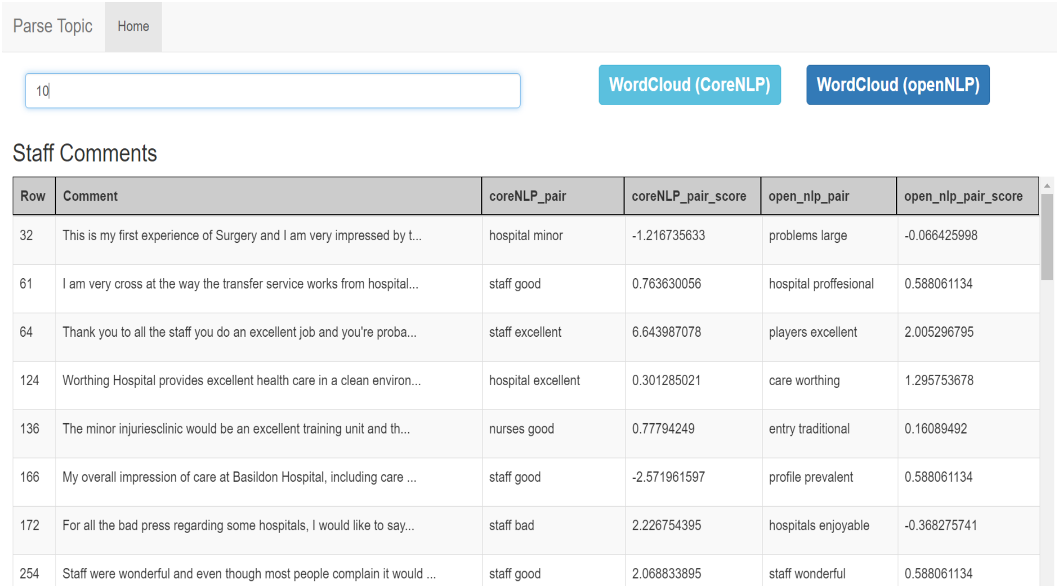
However, a challenge in summarisation is the validation of it as there is no objective method of validating such a summarisation. The best available method is human evaluation. The process of human evaluation involves asking people preferably relevant to the summarisation data type to verify whether they agree with the summarisation provided by the Dependency Parsing methods used in the study.

For human evaluation, the validation was undertaken by medical experts, the same group who evaluated the performance of the Topic Modelling methods. They were asked to check which of the two methods provided a better summarisation. They were asked to read the patient comments and the extracted noun-adjective pairs from both the tools. Each participant randomly chosen patient comments and checked the pairs extracted for around 35-40 patient comments. They were asked to state their opinion as if to which pair extracted gave a better summarisation of the patient comment. The experts agreed that the summarisation provided by the coreNLP method is more appropriate than the openNLP method. Thus, Stanford`s coreNLP approach is more suited for achieving a summarisation of patient reviews in terms of answering "why does the patient carry a particular sentiment in a given review?".

6.4.4 Visualisation of Dependency Parsing

For large databases, as in Table 6-1, Table 6-2, Figure 6-2 and Figure 6-3, listing the pairs in table would make analysis difficult, boring and less interactive to the users. To address this, a web API is developed which allows users to interactively visualise various comments in which a particular topic appear and it interactively displays the noun-adjective pair to the users to obtain a summarisation of the comment. The web API can be accessed at http://67.222.33.111/bahja/#!/ .

The web API for visualisation of Dependency Parsing is visualised in the following screenshots.



Row	Comment	coreNLP_pair	coreNLP_pair_score	open_nlp_pair	open_nlp_pair_score
32	This is my first experience of Surgery and I am very impressed by t...	hospital minor	-1.216735633	problems large	-0.066425998
61	I am very cross at the way the transfer service works from hospital...	staff good	0.763630056	hospital proffesional	0.588061134
64	Thank you to all the staff you do an excellent job and you're proba...	staff excellent	6.643987078	players excellent	2.005296795
124	Worthing Hospital provides excellent health care in a clean environ...	hospital excellent	0.301285021	care worthing	1.295753678
136	The minor injuriesclinic would be an excellent training unit and th...	nurses good	0.77794249	entry traditional	0.16089492
166	My overall impression of care at Basildon Hospital, including care ...	staff good	-2.571961597	profile prevalent	0.588061134
172	For all the bad press regarding some hospitals, I would like to say...	staff bad	2.226754395	hospitals enjoyable	-0.368275741
254	Staff were wonderful and even though most people complain it would ...	staff good	2.068833895	staff wonderful	0.588061134

Figure 6-4: A screenshot of the webpage visualising Dependency Parsing for topic 10.

In the above screenshot, it can be observed that when a particular topic is chosen, for instance, topic number 10 as shown above, all the comments in which this topic is discussed are listed on the webpage. Further, the user can click on a comment to visualise the noun-adjective pair extracted from both the openNLP and coreNLP methods. This would enable the users to quickly skim through databases to find comments on a particular topic and then obtain a quick summary of the comment, thereby obtaining an automated analysis and understanding of patient feedback and experience.

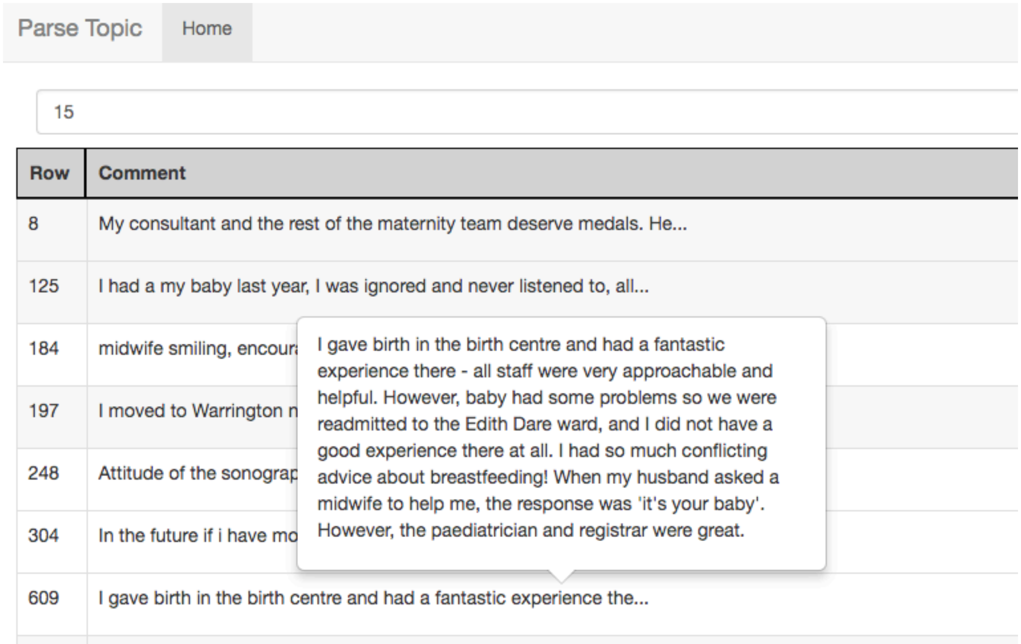


Figure 6-5: A screenshot of the webpage visualising Dependency Parsing for topic 15.

Similar to Figure 6-4, the above screenshot illustrates the API visualisation for topic number 15. Thus, this web API is an excellent tool for hospital staff to understand patient feedback in a quick and effective way.

The noun-adjective pair displayed in the above screenshots is still limited information for analysis purposes. Another method of visualisation was developed for this study to better visualise the patient feedback. Based on the SoA scores that explained earlier, functionality that is added in this visualisation API is the Word cloud generator. Word cloud is a visualisation technique, which allows for a more valuable way to communication imperative information/data at a simple glance. Word clouds work in a simple way, the more a specific or significant a term appears within a source of data, the bigger and bolder the term appears within the Word cloud (different colours could also be applied). In simple terms a Wordcloud us a visual representation of text data. When the Word cloud option is chosen for the coreNLP or openNLP method, the user can graphically visualise the top 20 most positive or negative scoring word pairs for that particular topic from all the reviews. This type of visualisation allows the user to get an overview of what are the factors that are leading to the positive or negative sentiment amongst patients for a given topic. The screenshot of the Word cloud for both positive and negative scoring for the topic 25 is shown in Figure 6-6 and Figure 6-7

Figure 6-6: Screenshot of wordcloud visualisation of positive sentiment for topic 25 using the coreNLP method.

Figure 6-7: Screenshot of wordcloud visualisation of negative sentiment for topic 25 using the openNLP method.

The logic of generating the wordcloud visualisation using the MySQL procedure is presented in Appendix B.

6.5 Discussion

In this chapter, the third iteration of the research study on patient experience was presented. The study combined the outcomes from the studies presented in Chapter 4: to apply Dependency Parsing methods on the patient feedback database to obtain an automated summary of the patient reviews.

In this study, two approaches of Dependency Parsing were explored, one using the openNLP suite and the coreNLP suite. Using these two suits each patient review was summarised or represented by a noun-adjective pair. The rationale behind this approach is to explore whether the noun-adjective pair can summarise a review`s sentiment by answering the reason behind that sentiment. The noun would pertain to the subject to which the patient refers to in the comment and the adjective would be a reflection of the sentiment of the patient.

The task of representing a review by a noun-adjective pair is challenging for multiple reasons. The main challenge is that a single review normally would refer to different subjects (nouns) and each noun may be associated with multiple adjectives. Thus, in order to overcome this challenge, the nouns in a review were ranked according to the frequency of their occurrence in the patient feedback, with as assumption that more frequently occurring nouns would more likely be the central topic of discussion in the review. Further, the Sentiment Analysis outcomes using the SoA approach presented in Chapter 4: were used to find the strongest sentiment of the existing adjectives in a review and then use it to summarise the comment.

Another challenge with Dependency Parsing is to validate the performance of the Dependency Parsing method. Since there are no objective methods available, subjective validation was deployed, where experts from NHS reviewed the summarisation obtained from both the methods and rated which method gives a better one. Apart from human validation, there are no other available methods to validate the performance of Dependency Parsing, which is a limitation of the study.

Further, since there is no objective validation of the performance of parsing methods, it is not possible to obtain the accuracy of the method. Regarding this, it can be said that on some occasions, the noun-adjective pair extracted from a particular review may not provide an accurate summarisation of the review. However, from the human validation, it was noted that the present summarisations obtained using the coreNLP method provide a satisfactory performance regarding this.

Despite the challenges and limitations, this study has enabled exploration of the concept of implementing Dependency Parsing for an automated analysis of patient feedback to find the reason behind a particular sentiment in a review. Further research on this topic in the context of patient experience could facilitate a more detailed automated analysis of patient reviews that could be used to enhance healthcare services.

6.6 Summary

In this chapter, the Dependency Parsing method was applied on the patient feedback to obtain a summary of individual patient feedback. The outcomes of the study showed that a "noun-adjective" pair based summary method could provide an overview of patient feedback and identify the possible reason behind a particular sentiment in this feedback. Further, the visualisation module API presented in this chapter could assist clinicians in obtaining the summary of each patient feedback in an interactive and engaging method.

Chapter 7: Conclusion

7.1 Summary of the Work

The research findings presented in this thesis aimed at providing a framework for an automated analysis of a patient feedback database that identifies the patient sentiment in that feedback as well as the topic and themes associated with it. The research process of the research involved five main stages that included a detailed literature review, designing the research methodology, identifying the patient sentiment from their feedback, extracting topics from the patient feedback, and estimating the rationale behind the patient sentiment for a given topic. In this section, an overview of this thesis and the findings presented in it are provided.

Chapter 2 provided a detailed literature review relevant to the research area. This part of the research can be equated with the **awareness of the problem** aspect of the Design Science Research (DSR) methodology followed in the chapter. This chapter covered the existing research in exploring natural language processing methods for analysing patient experience data. The patient experience definition and its associated features were also discussed in this chapter.

The findings of the literature review were that there was a very limited work in applying text mining and language processing methods to patient experience data. This was an important research gap that had to be addressed and thus, became a motivation for the research. In order to operationalise the patient experience analysis, the patient experience was equated with the user experience field of research. This is because, essentially, patient experience is users providing their feedback about the healthcare service they have received, which as such is asking to the definition of user experience, i.e. capturing user feedback about a product or service.

The technologies and natural language processing methods that have extensively used in user experience research were studied and three main natural language processing methods were identified for the research. This part of the research pertains to the **suggestion** aspect of the DSR methodology, as described in Chapter 3, where potential methods as solutions to address the research were identified and shortlisted

The first NLP method that was identified is the Sentiment Analysis (SA) method. The SA methods enable identifying the patient sentiment automatically from the data. This method was chosen because it would give a quick overview of the general patient perception about the healthcare service of the hospital to the staff members. This method would allow the clinicians to understand the patient satisfaction rate they achieve and thus, get an abstract idea about their performance.

The next method identified was the Topic Modelling approach. Topic Modelling would enable an automatic identification of the most commonly commented topic by the patients on the feedback form and thus, identify the topics and themes that might possibly need to more focus.

Further, the NLP Dependency Parsing method was also identified as a solution for the research. Dependency Parsing was chosen because it would allow the clinicians and researchers to identify the nouns commonly referred to in the patient feedback and the associated adjectives present in the comment. This would enable obtaining a summary of the patient feedback and thus, would facilitate a quick understanding of large patient feedback databases and hence, save time and effort for clinicians.

Chapter 3 discussed the methodology that would be applied throughout the study. The research approach selected was one that was systematic so as to obtain outcomes that were deemed reliable as well as reproducible. The methodology that was selected was Design Science Research (DSR). This methodology was discussed in much detail throughout the chapter as well as justification for its choice when compared against other similar methodologies that could have also been used to undertake this research. Initially, the chapter introduced the many different types of research approaches in relation to information systems. The approaches covered were the Positivism, Interpretive, Behavioural Science and Design Science.

The chapter went on to discuss the DSR paradigm, where a majority of the work has considered the different stages of the DSR cycle. This also allowed for a comparison of the different DSR models in terms of the similarities and differences. Specifically, the models compared were those of Vaishnavi and Kuechler, Hevner and Pfeffer. Next, a DSR based framework for patient experience analysis was presented. This part of the chapter was divided into three sections, each presenting a different iteration of the process. The first iteration pertained to the Sentiment Analysis of patient feedback data, whilst the second iteration was about Topic Identification from patient feedback and the third and final iteration referred to rationale identification of patient sentiment.

Chapter 4 was the first iteration of the research that explored the Sentiment Analysis approach for identifying patient sentiment from the NHS patient feedback database. In this chapter, a detailed discussion of the SA approach was provided. From the research process, three SA methods were finalised to explore in the study. The SA models listed were the SVM method, Naïve Bayes approach, and the Strength of Association methods. These three methods were chosen because of their extensive application in user experience studies in the literature and their proven performance in prediction.

The three SA models used in the research were essential prediction models that were trained to predict the patient sentiment. To train the models, the NHS choices dataset was divided into training, test, and validation datasets. The models were then trained on the training dataset, which included the patient comments and for each comment the ratings provided by that patient for the hospital. The ratings were provided as a sentiment score and were mapped to a scale of 1 to 5. The number 3 was considered to be a neutral rating and was discarded in the study and the scores of 1 and 2 were considered to be a negative comment, whereas scores of 4 or 5 were taken as being a positive comment. The comments and their associated score were then fed to the prediction models to train them on the patient sentiment.

The trained sentiment prediction models were then tested for their performance on the test and the validation dataset. The performance analysis of the models was further done using single-fold and multi-fold approaches, where the test and validation dataset were used with different folds of combination to test the performance of the models. The performance of the SA models was undertaken using three performance metrics: prediction accuracy, sensitivity, and specificity. The outcomes of the study showed that the SVM model gave the best performance in terms of the prediction accuracy. The SVM model gave a prediction accuracy of up to 85%. The prediction accuracy indicates that 85 out of 100 times, the SVM method was able to classify correctly the patient sentiment in the comment as either positive or negative. Thus, this iteration of the research showed that SA models could be used to predict patient sentiment with a high degree of accuracy. This would enable the clinicians to obtain an overview of the hospital`s performance in terms of patient satisfaction rate.

Chapter 5 was the second iteration of the research, wherein the Topic Modelling approach was explored for patient feedback analysis. In the first iteration of the research, the sentiment of the patient in a given comment was automatically identified using the SA methods. The sentiment identification identified what is the patient sentiment in a given comment. However, this information is limited and it would be necessary for clinicians to know what are the topics discussed by the patient in the given comment.

To achieve this, the Topic Modelling approach was explored and implemented to analyse patient experience data. In the research for this iteration, the Latent Dirichlet Allocation (LDA) Topic Modelling method was identified as being the best available for user experience analysis. Hence, this method was adapted to analyse patient experience data. The LDA model was implemented with two approaches, namely, a unigram topic model and a bigram topic model. The unigram topic model identified single word topics from the patient database, and the bigram model identified two word topics. Using each of the methods, 30 topics each were identified that were those most commonly occurring in the NHS patient feedback database.

Further, the study was expanded by using a novel method were the sentiment for each comment identified in the first iteration of the research was associated to the topic identified by the LDA method. Thus, this method not only enabled automatic identification of topics by Topic Modelling methods, but also provided a methodology for identifying the sentiment of the topic, thereby providing more detailed information about the patient feedback to the clinicians.

Additionally, a topic visualisation method using the LDAvis approach was implemented to visualise the topics as bubbles on Cartesian coordinates, where detailed information of the topic distribution in the database is provided. This visualisation enables an easy and quick analysis of patient feedback and could be helpful in identifying the problem areas through the topic occurrence and the associated sentiments identified.

Chapter 6 was the third and final iteration of the research. In this chapter, the Dependency Parsing method was explored for eliciting patient experience. The Dependency Parsing (DP) method aims at answering the “why” of the patient experience analysis process. In other words, it aims to identify, why the patient provided a given comment on the hospital.

The approach of Dependency Parsing is to perform a sentence level analysis of the patient feedback. From this analysis, the DP method identifies the frequently occurring nouns in the feedback and identifies that as the most probable subject being discussed in the comment. Further, the DP method also identifies the associated “adjectives” with the noun and thus, provides it as the associated sentiment to the noun. This gives a ‘noun-adjective’ pair to the clinician, which facilitates an automated answering of why the patient provided a positive or negative sentiment about a given topic in relation to a given hospital.

The DP methods were applied with two approaches. One was the Apache openNLP method that provides a noun-adjective pairs identified from the patient comment. The second method was the coreNLP method that was developed by Stanford University to identify the noun-adjective pairs. Thus, extracting noun-adjective pairs would provide a quick overview of the patient comment and help identify the frequently occurring subject and the associated sentiment with the subject in the patient comment. The sentiment scores obtained from the SA methods in the first iteration of the research are used to label the sentiment of each word in the review and the most frequently occurring ones are extracted as a word pair by the DP method that can potentially describe "why" a particular sentiment is given to a review.

Extracting a 'noun-adjective' pair can be associated with the phenomenon of "summarisation". The noun-adjective pairs could be considered as a summary of the patient comment in a pair of words, i.e. by extracting the possible subject (the noun) being discussed in the comment along with its associated sentiment. The DP method provides a quick summary of the patient comment. This "summary" enables the clinicians to skim quickly through large amounts of patient comments and greatly reduce their time spent, as they do not read them all in detail comment. That is, the process leads to identifying what and why a particular topic is being discussed without laborious extensive reading. A limitation, however, is that the accuracy of the 'summary' needs to be established with quantitative measures.

Further, to facilitate an easy and quick way of visualising the patient feedback database, an interactive API is provided where the clinicians or researchers can search comments in terms of topics, i.e. on the API, the user can type in a topic and the API provides a list of patient comments about the given topic. Further, the user can also find the 'noun-adjective' pair for each comment and thus, can obtain a summary of the patient feedback in each comment. This enables a quick and interactive analysis of patient feedback.

To summarise, by the end of the third iteration of this research, the sentiment scores obtained by the SA methods for patient comments and the topics identified by the Topic Modelling method were combined along with the DP method to summarise each comment that can be identified by their topics and their associated sentiment.

7.2 Contributions to Knowledge

An important aspect of a research is its contributions to knowledge. The five key ones of the research are presented below:

a. The Support Vector Model (SVM) based sentiment analysis model provides the most accurate sentiment identification from patient experience dataset when compared with SOA and NB based sentiment analysis models
b. The LDA topic modelling is a useful tool in identifying the topics from patient feedback dataset and is effective in identifying themes from the dataset
c. The NLP method could be used to obtain a summary of patient experience automatically via a *noun-adjective* pair;
d. A three-stage structured approach framework was provided to identifying sentiment in patient feedback, involving finding topics or themes in the feedback and providing a summary capturing the patient experience;
e. An interactive visualisation API for visual analysis of sentiment scores, topic distributions and summarisation.

7.2.1 A structured approach to the automated analysis of patient experience

This research has shown that an effective and structured approach to NLP methods could be used for automated analysis of patient experience data. The approach followed in this study can be observed as a multi-stage approach for patient experience analysis. It can be used to explore and identify different aspects of the patient feedback without requiring significant amounts of time or effort from the clinicians.

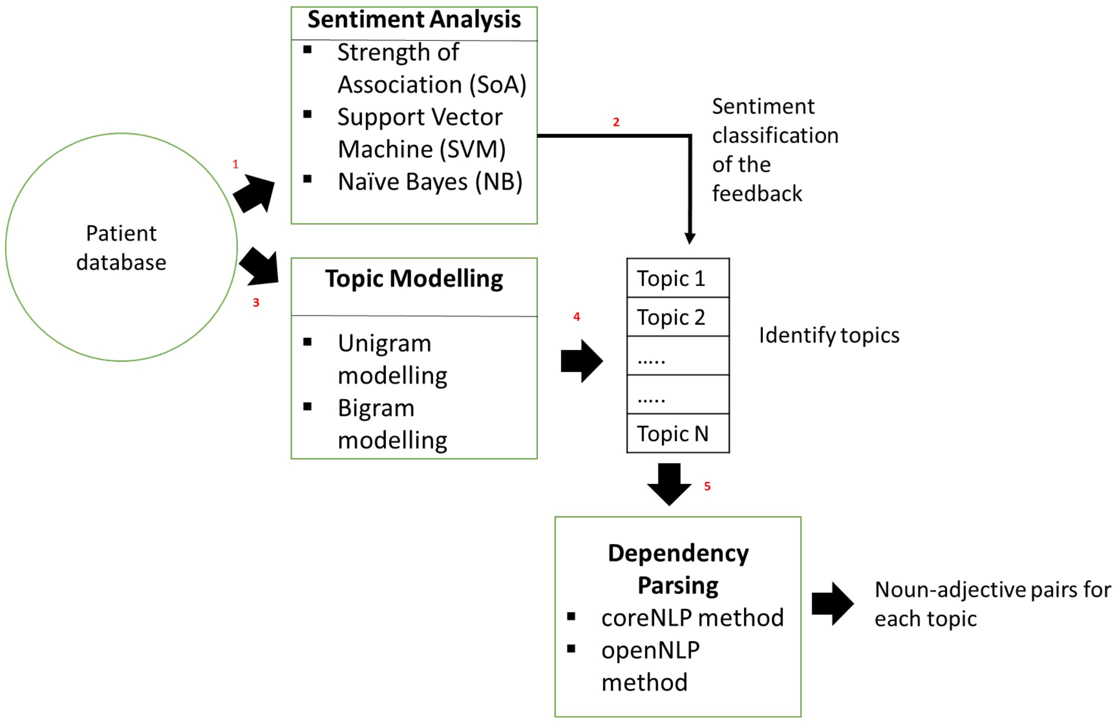


Figure 7-1: A schematic representation of the three staged approach to patient experience analysis followed in this thesis.

In Figure 7-1, the three staged approach followed for the research is illustrated. In the first stage of the structured approach deployed, the patient sentiment from the feedback database can be identified with an accuracy of 85% using the SVM approach of SA (see Section 4.4), thereby getting an understanding of the general opinion of patients about a given hospital. In the next stage, the approach involves identifying the topics being discussed in the patient feedback and provides further insight into the topic by associating it with the identified sentiment from the first stage (see Chapter 5: for details). In the final stage, the patient feedback is summarised by extracting noun-adjective pairs using Dependency Parsing methods and the sentiment scores identified from the first stage (details in Chapter 6: . It should be noted that the above-described multi-stage approach of patient experience analysis identifies several features of the patient feedback including the patient sentiment, the topic being discussed and an automated summary of the feedback to identify the reason behind a particular sentiment for a given topic.

7.2.2 Sentiment Analysis of Patient Experience

The first iteration of the research is an important contribution to knowledge. Sentiment Analysis approaches have been widely explored in the literature for computerised analysis of user experience and identifying the user sentiment for a given product or service. There are several works providing sentiment identification of topics, such as movie reviews and phone reviews. However, to the best of this researcher's knowledge very few works in the literature have applied SA approaches for analysing patient experience data.

Moreover, there is a lack of research work on computerised analysis of patient sentiment regarding the NHS database in the UK. The NHS, being one of the largest public healthcare service providers in the world, has made considerably little exploration of SA methods and other NLP methods. Thus, the study presented in chapter 4 is possibly the earliest work in applying SA models for NHS patient database, therefore acting as a stepping-stone for all future works in this field of interest.

In the SA study, the performance of three approaches to sentiment analysis – Strength of Association, Support Vector Machine and Naïve Bayes – were assessed in the context of a dataset related to patient experience (drawn from the NHS choices website). The performance of the model was further cross-validated by using a multi-fold cross validation approach for the study. The SVM approach provided the most accurate prediction (85% accuracy) in assessing patient feedback as being positive or negative (see Section 4.4). That is, in the single-fold and multi-fold study, the SVM showed the best performance, with f-scores of 0.84 and 0.78 respectively (see Section 4.4). The outcomes of this study showed that SA models, such as SVM prediction models, can be effectively used for patient sentiment identification in NHS data. That is, they can be deployed for obtaining a top-level analysis of patient experience and sentiment through easy, quick, and automated ways. The results of this study are consistent with the findings of Alemi *et al.*(2012) and Greaves *et al.*(2012c), who observed that SVM classification method shows high performance accuracy when compared to other classification methods. The findings of the study were further utilised in the second and third stages of the research, as described in the following subsection.

In summary, the exploration of SA models on patient feedback showed that the SVM based SA provides the highest accuracy in detecting the patient sentiment from the database. Thus, it is possible to get an overview of the patient sentiment from the database using SA models.

7.2.3 Topic Modelling to identify Topics from Patient Feedback

Another important contribution of this research is that it has demonstrated that NLP methods, such as Topic Modelling, can be applied to identify topics from a patient feedback database, which has not been the case in previous studies or works. In the study presented in Chapter 5, two different Topic Modelling methods, namely, unigram and bigram methods, were applied on the database. The outcomes of the study showed that the LDA method of Topic Modelling could identify pertinent topics from the patient feedback. From the unigram model, 30 topics were identified. Each topic had 25 ‘single’ words, and these words were identified by the unigram method based on the probability of them belonging to a single topic. The identified words for each topic were manually analysed by medical experts and they provided a topic label for each category of 25 words (see Table 5-1 for further details). A similar approach was applied using the bigram method, where each topic had 25 ‘pair of words’ and were identified by the model based on the probability of them belonging to a topic. In this case also, the identified pair of words for each topic were manually analysed by medical experts and they provided a topic label for each category (see Table 5-4 for more details). From both of these methods 30 topics were identified from the database and the NHS experts who took part in the study confirmed that the topic identification executed by the LDA method were appropriate topics and consequently, the models are able to capture the themes embedded in patients' comments.

Hence, in this part of the study it was demonstrated that Topic Modelling can identify the topics presented in a given patient feedback database and is able to provide information on the topic distribution and other related information. In the literature, to the best of this researcher's knowledge there is only one work by Purver *et al.*(2006) that used the LDA approach to capture aspects of doctor-patient communication in therapy, including predicting symptoms and therapeutic relationships. The outcomes of their study showed that LDA was effective in the latter, i.e. in predicting the therapeutic relationship between the doctor and the patients.

In summary, the study conducted using topic modelling demonstrated that the LDA method is useful in identifying the themes from the dataset that can be useful for the users for further analysis along with the help of the visualization tool developed. This research is one of few works that have investigated Topic Modelling for eliciting patient experience. The outcomes can pave the way for further fine-grained analysis of patient feedback that can enable better identification of topics and associated content from databases.

7.2.4 An approach to summarising patient feedback

Another significant contribution of this research is the approach to summarising patient feedback by exploring the NLP methods, in particular, Dependency

Parsing. The study presented in Chapter 6 applied the DP methods to generate ‘noun-adjective’ pairs that effectively gives a summary of the patient feedback.

The study involved applying and evaluating two DP methods: *openNLP* approach and the *coreNLP* method (refer Section 6.3 for details). Using these two approaches, each patient review was summarised or represented by a noun-adjective pair. The rationale behind this approach is to explore whether the noun-adjective pair can summarise a review`s sentiment by answering the reason behind that sentiment. The noun refers to the subject to which the patient is pointing to in the comment and the adjective is a reflection of the sentiment of the patient (refer Section 6.4). Further, the Sentiment Analysis outcomes using the SoA approach presented in Chapter 4: were used to find the strongest sentiment of the existing adjectives in a review and then this was to summarise the comment in either a positive or negative category. In order to analyse the performance of the NLP tools, a subjective validation study involving medical experts was carried out. The medical experts agreed that the summarisation provided by the coreNLP suite was better than that of the openNLP suite (see Section 6.4 for discussion on the results).

This is a novel contribution to knowledge as there are no works in the literature to date have explored the DP approach to provide a summary of user experience in the context of patient experience. The approach presented in this study combines the Topic Modelling and Sentiment Analysis method approaches to extract a word pair that captures the essence of the given patient comment. Further, the DP method was also evaluated by the NHS experts in a survey and they agreed that the DP method is a useful tool for obtaining a quick summary of a given patient comment.

7.2.5 An interactive visualisation interface

Another contribution of this study is the API that was presented in Chapter 5 and Chapter 6 for data visualisation. The Topic Modelling API provides an interactive interface to the clinicians, where the top 30 topics from the patient feedback database are visualised on Cartesian coordinates along with their distribution across the dataset. This visualisation provides the clinicians with detailed information about the topics and the associated keywords with the topic, thereby making the analysis more easily interpretable (see Section 5.5 for details). The visualisation provided for Topic Modelling is inspired by the LDAvis approach provided by Sievert and Shirley (2014) and Chuang, Manning and Heer (2012). That is, the API provided by the authors was adapted for this case study. Further, a visualisation API was built that visualises the outcomes of the three iterations of the study. The details of this API, as provided in Subsection 6.4.4 briefly summarised below.

The outcomes of the Dependency Parsing approach were also provided on an interactive interface. On this interface, the user can search relevant comments for any given topic identified by the LDA method and then for each comment, the user will be able to find a word pair that provides an effective summary of it. This interactive visualisation helps the clinicians to obtain an automated summary of the overall patient sentiment and opinions (screenshots of the API are provided in Subsection 6.4.4).

For instance, in the below screenshot, it can be observed that when a particular topic is chosen, for instance, topic number 25, all the comments in which this topic discussed are listed on the webpage. Further, the user can click on a comment to visualise the noun-adjective pair extracted from both the openNLP and coreNLP methods. This would enable the users to skim quickly through databases to find comments on a particular topic and then obtain a quick summary of the comment, thereby delivering automated analysis and understanding of patient feedback and experience.

Figure 7-2: A screenshot of the webpage visualising Dependency Parsing for topic 25.

In sum, much of the contribution of this study comes from the fact that little or no research has been undertaken on patient experience in the field of Sentiment Analysis and Topic Modelling. Consequently, the findings can provide future researchers a platform on which they can base their work. The lack of prior work did prove to be problematic when analysing literature, but the subsequent adoption of the DSR approach led to the opening up of a whole new topic for discussion in the field of data mining. To reiterate, the Topic Modelling API presented is also one of the novelties of this research and again, this is something that future research can build upon to improve the capture of patient experience sentiment or an entirely different domain could be probed utilising the combined approaches introduced in this thesis to provide a summary of user feedback.

7.3 Research Limitations and Future Work

7.3.1 Limitations

As with most researches, the study presented also has certain limitations, which are discussed below.

Improving the Accuracy of prediction models

The SA models have shown to be effective in predicting the patient sentiment from the database. However, the best performing model, SVM, had an accuracy of close to 86%. Even though the accuracy is good enough to be acceptable, there is still a need for significant increase in the accuracy levels to make the prediction model more reliable and robust.

Further, the model training was done by combining the patient ratings and their associated comments. However, in many cases, the patients just provided the ratings and would not make comments in the feedback form or vice versa. Such ratings were not included in the training model. Thus, this limits the patient feedback database and poses challenges in improving the accuracy of the sentiment prediction model.

Moreover, classifying a patient comment into a binary classification of positive and negative might not always be the best classification, because in many instances, a given comment can consist of both positive and negative sentiment and thus, may confuse the prediction model, resulting in it classifying inaccurately. Hence, these challenges need to be addressed.

Topic Modelling needs supervision

The LDA Topic Modelling method identifies the topics from a given patient feedback database and can be termed an automated process. However, a limitation with this approach is that the LDA method will be able to identify a “bag of words” that are most likely to belong to a topic. The bag of words then needs to be analysed by a clinician, who would then provide a label to the topic, such as maternity department, parking area, etc. This implies that there needs to be timely intervention by the clinician to recognise when new topics are identified from the recent patient feedback and then assign a label to the topic identified or categorise the topic into an older one. Thus, this makes the process semi-automated at this stage. Further, the Topic Modelling method identifies the most frequently occurring themes and hence, other topics that may be more salient than some of those identified for clinicians to be aware of could be overlooked.

Lack of Objective Measurements

An important limitation of the approaches used in the research is that there are no objective measurements to test the performance of the Topic Modelling and Dependency Parsing approaches. The Sentiment Analysis approaches can be measured for their performance via prediction accuracy and specificity measures. However, the performance of the other two approaches cannot be measured via quantitative measures.

In the current study, this limitation was overcome by implementing qualitative measurements of the performance of the Topic Modelling and Dependency Parsing methods. Experts from NHS collaborated on the research and contributed by assigning suitable labels for the topics identified by the Topic Modelling method. The qualitative survey conducted with the doctors confirmed that the LDA method was useful in identifying relevant topics from the patient feedback database. Similar qualitative measurements were conducted for the outcomes of the Dependency Parsing methods and the NHS experts were satisfied with the performance of these methods.

However, due to the nature of NLP, it is often difficult quantitatively to measure the performance of some of the text mining methods used in the research.

Model needs continuous, complex and diverse training dataset

For this research, a large database of patient feedback was used. However, for prediction models, such as Sentiment Analysis, Topic Modelling and Dependency Parsing methods, there is a need for constant supply of complex and diverse datasets for training the models and to keep them continually updated. A limitation of this approach is that as the database increases, in which case the prediction model`s performance may become further complicated and this pose challenges in relation to maintenance and supervision. This would lead to more resources being required to manage the entire process.

7.3.2 Future Work

The current work is one of the earliest works towards developing automated analysis of patient experience data. Several avenues for future work are proposed below

- Applying more advanced machine learning methods to improve the accuracy of the Sentiment Analysis models.
- Exploring other existing Topic Modelling methods from the literature, such as Latent Semantic Analysis, Explicit Semantic Analysis and Matrix Factorisation, amongst others, to identify more topics from the patient feedback database.
- Applying deep learning approaches along with Dependency Parsing methods to understand and summarise the individual reviews in a more effective and tangible way and thus, enable identifying the reason behind a particular sentiment in a particular patient comment.
- In order to make the sentiment prediction model, Topic Modelling method and the Dependency Parsing methods more effective, accurate and reliable, future works may consider collecting more data from other patient feedback resources, such as private hospital networks, social media and other resources and use them to train and develop the current models into better performing ones.
- Finding other databases of patient feedback and applying the developed models from this research to them and then, comparing the performance of all the models, including that developed in this thesis.
- Developing metrics that can be useful for a quantitative measurement of patient experience data analysis using NLP methods.
- Creating a test dataset that could be further utilized in objective evaluation of topic modelling and summary generation outcomes.

- Creating a generic API and web service that will be available to clinicians and associated staff from the NHS so that they can use the hands-on tool for analysing patient sentiment and also possibly get real time information of hospital performance regarding patient satisfaction.
- Exploring other data visualisation, such as time series, to track the progress of the hospitals in enhancing the capturing of patient sentiment and feedback. Further, exploring other visualisation options to make the patient experience analysis more interactive and easily interpretable.
- Further work to consider for future is that of perhaps representing patient experience over time in a time series. Time series can help health care professionals in a number of areas, including, but not limited to, needs assessments, service planning as well as policy development. Time series also makes it possible for professionals to predict future frequencies along with the rates of occurrence.

References

Adrover, C., Bodnar, T., Huang, Z., Telenti, A. and Salathé, M. (2015) 'Identifying Adverse Effects of HIV Drug Treatment and Associated Sentiments Using Twitter', *JMIR public health and surveillance,* 1(2).

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011) 'Sentiment analysis of twitter data', *Proceedings of the workshop on languages in social media.* Association for Computational Linguistics, 30-38.

Agarwal, B., Poria, S., Mittal, N., Gelbukh, A. and Hussain, A. (2015) 'Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach', *Cognitive Computation,* 7(4), pp. 487-499.

Alderwick, H., Robertson, R., Appleby, J., Dunn, P. and Maguire, D. (2015) *Better value in the NHS.* NHS.

Alemi, F., Torii, M., Clementz, L. and Aron, D.C. (2012) 'Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments', *Quality Management in Healthcare,* 21(1), pp. 9-19.

Aletras, N. (2014) *Interpreting Document Collections with Topic Models,* .

Andrew, S. (2014) 'Evaluation and measurement of patient experience', *Patient Experience Journal,* 1(1), pp. 28-36.

Arora, S., Ge, R. and Moitra, A. (2012) 'Learning topic models--going beyond SVD', *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on.* IEEE, 1-10.

Berger, A.L., Pietra, V.J.D. and Pietra, S.A.D. (1996) 'A maximum entropy approach to natural language processing', *Computational linguistics,* 22(1), pp. 39-71.

Bhattacherjee, A. (2012) 'Social science research: principles, methods, and practices', .

Blei, D.M. (2012) 'Probabilistic topic models', *Communications of the ACM,* 55(4), pp. 77-84.

Blei, D.M. and Lafferty, J.D. (2009) 'Topic models', *Text mining: classification, clustering, and applications,* 10(71), pp. 34.

Buchholz, S. and Marsi, E. (2006) 'CoNLL-X shared task on multilingual dependency parsing', *Proceedings of the Tenth Conference on Computational Natural Language Learning.* Association for Computational Linguistics, 149-164.

Cambria, E., Schuller, B., Liu, B., Wang, H. and Havasi, C. (2013a) '***Statistical***

approaches to concept-level sentiment analysis', *IEEE Intell. Syst.,* 28(3), pp. 6-9.

Cambria, E. (2013) 'An introduction to concept-level sentiment analysis', *Mexican International Conference on Artificial Intelligence.* Springer, 478-483.

Cambria, E., Schuller, B., Liu, B., Wang, H. and Havasi, C. (2013b) 'Knowledge-based approaches to concept-level sentiment analysis', *IEEE Intelligent Systems,* 28(2), pp. 12-14.

Chen, C.C. and Tseng, Y. (2011) 'Quality evaluation of product reviews using an information quality framework', *Decision Support Systems,* 50(4), pp. 755-768.

Chen, Z. and Liu, B. (2016) *Lifelong Machine Learning.* Morgan & Claypool Publishers.

Cheng, X., Yan, X., Lan, Y. and Guo, J. (2014) 'Btm: Topic modeling over short texts', *IEEE Transactions on Knowledge and Data Engineering,* 26(12), pp. 2928-2941.

Christidis, K. and Mentzas, G. (2013) 'A topic-based recommender system for electronic marketplace platforms', *Expert Systems with Applications,* 40(11), pp. 4370-4379.

Chuang, J., Manning, C.D. and Heer, J. (2012) 'Termite: Visualisation techniques for assessing textual topic models', *Proceedings of the International Working Conference on Advanced Visual Interfaces.* ACM, 74-77.

Cohen, L., Manion, L. and Morrison, K. (2013) *Research methods in education.* Routledge.

Cole-Lewis, H., Varghese, A., Sanders, A., Schwarz, M., Pugatch, J. and Augustson, E. (2015) 'Assessing electronic cigarette-related Tweets for sentiment and content using supervised machine learning', *Journal of medical Internet research,* 17(8), pp. e208.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M. and Taniguchi, K. (2008) 'BioCaster: detecting public health rumors with a Web-based text mining system', *Bioinformatics (Oxford, England),* 24(24), pp. 2940-2941.

Collobert, R. and Weston, J. (2008) 'A unified architecture for natural language processing: Deep neural networks with multitask learning', *Proceedings of the 25th international conference on Machine learning.* ACM, 160-167.

Coulter, A., Fitzpatrick, R. and Cornwell, J. (2009a) *Measures of patients' experience in hospital: purpose, methods and uses.* King's Fund.

Coulter, A., Locock, L., Ziebland, S. and Calabrese, J. (2014) 'Collecting data on patient experience is not enough: they must be used to improve care', *BMJ,* 348(mar26 1), pp. g2225-g2225.

Coulter, A., Fitzpatrick, R. and Cornwell, J. (2009b) 'The point of care measures of patients' experience in hospital: purpose, methods and uses', *The Kings Fund,* , pp. 1-32.

De Smet, W. and Moens, M. (2009) 'Cross-language linking of news stories on the web using interlingual topic modelling', *Proceedings of the 2nd ACM workshop on Social web search and mining.* ACM, 57-64.

Department of Health (2016) *NHS Outcomes Framework: at-a-glance*Department of Health.

Di Caro, L. and Grella, M. (2013) 'Sentiment analysis via dependency parsing', *Computer Standards & Interfaces,* 35(5), pp. 442-453.

Diao, Q., Qiu, M., Wu, C., Smola, A.J., Jiang, J. and Wang, C. (2014) 'Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars)', *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 193-202.

Duan, W., Cao, Q., Yu, Y. and Levy, S. (2013) 'Mining online user-generated content: using sentiment analysis technique to study hotel service quality', *System Sciences (HICSS), 2013 46th Hawaii International Conference on.* IEEE, 3119-3128.

Edmund, R., Jason, M., Bae, M., Ariel, C., Liu PhD, M., Sandra, S., Wilson, M. and Milstein, M. (2014) 'Factors in patients' experience of hospital care: Evidence from California, 2009–2011', *Patient Experience Journal,* 1(1), pp. 95-110.

Fang, X. and Zhan, J. (2015) 'Sentiment analysis using product review data', *Journal of Big Data,* 2(1), pp. 1.

Feinerer, I. (2015) *Introduction to the tm Package Text Mining in R.*

Feinerer, I. (2013) 'Introduction to the tm Package Text Mining in R', *http://cran.r-project.org/web/packages/tm/ vignettes/tm.pdf,* .

Feinerer, I. (2008) 'An introduction to text mining in R', *R News,* 8(2), pp. 19-22.

Feldman, R. (2013) 'Techniques and applications for sentiment analysis', *Communications of the ACM,* 56(4), pp. 82-89.

Feldman, R., Netzer, O., Peretz, A. and Rosenfeld, B. (2015) 'Utilizing text mining on online medical forums to predict label change due to adverse drug reactions', *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1779-1788.

Foundation, A. (2016) *Apache openNLP* 2015).

Gann, W., Day, J. and Zhou, S. (2014) 'Twitter analytics for insider trading fraud detection system In: Proceedings of the sencond ASE international conference on Big Data', ASE.

Ganu, G., Elhadad, N. and Marian, A. (2009) 'Beyond the Stars: Improving Rating Predictions using Review Text Content.', *WebDB.* Citeseer, 1-6.

Ganu, G., Marian, A. and Elhadad, N. (2010) *URSA-User Review Structure Analysis: Understanding Online Reviewing Trends.* DCS technical report: Citeseer.

Ghiassi, M., Skinner, J. and Zimbra, D. (2013) 'Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network', *Expert Systems with Applications,* 40(16), pp. 6266-6282.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J. and Reyes, A. (2015) 'Semeval-2015 task 11: Sentiment analysis of figurative language in twitter', *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).* , 470-478.

Greaves, F., Pape, U.J., King, D., Darzi, A., Majeed, A., Wachter, R.M. and Millett, C. (2012a) 'Associations between Web-based patient ratings and objective measures of hospital quality', *Archives of Internal Medicine,* 172(5), pp. 435-436.

Greaves, F., Pape, U.J., King, D., Darzi, A., Majeed, A., Wachter, R.M. and Millett, C. (2012b) 'Associations between Web-based patient ratings and objective measures of hospital quality', *Archives of Internal Medicine,* 172(5), pp. 435-436.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. and Donaldson, L. (2012c) 'Machine learning and sentiment analysis of unstructured free-text information about patient experience online', *The Lancet,* 380, pp. S10.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. and Donaldson, L. (2013a) 'Use of sentiment analysis for capturing patient experience from free-text comments posted online', *Journal of medical Internet research,* 15(11), pp. e239.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. and Donaldson, L. (2013b) 'Use of sentiment analysis for capturing patient experience from free-text comments posted online', *Journal of medical Internet research,* 15(11), pp. e239.

Green, N. (2011) 'Dependency parsing', *WDS 2011 Proceedings of Contributed Papers,* , pp. 137-142.

Gregor, S. and Hevner, A.R. (2013) 'Positioning and presenting design science research for maximum impact.', *MIS quarterly,* 37(2), pp. 337-355.

Group, N. (2014) *The NHS Outcomes Framework 2015/16.* NHS.

Guzman, E. and Maalej, W. (2014) 'How do users like this feature? a fine grained sentiment analysis of app reviews', *2014 IEEE 22nd international requirements engineering conference (RE).* IEEE, 153-162.

Hair, J.F. (2009) 'Multivariate data analysis', .

Hedegaard, S. and Simonsen, J.G. (2013) 'Extracting usability and user experience information from online user reviews', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2089-2098.

Hevner, A., March, S., Park, J. and Ram, S. (2004) 'Design Science in Information Systems Research', *MIS Quarterly,* 28(1), pp. pp.75-105.

Hevner, A.R. (2007) 'A three cycle view of design science research', *Scandinavian journal of information systems,* 19(2), pp. 4.

Hoeksema, J. (2010) 'Negative and Positive Polarity Items: An investigation of the interplay of lexical meaning and global conditions on expression. in LR Horn (ed.),', *The expression of negation. The expression of cognitive categories, no. 7,* De Gruyter Mouton, Berlin, pp. 187-224.

Höpken, W., Fuchs, M., Höll, G., Keil, D. and Lexhagen, M. (2013) 'Multi-dimensional data modelling for a tourism destination data warehouse', in *Information and communication technologies in tourism 2013.* Springer, pp. 157-169.

Hu, M. and Liu, B. (2004) 'Mining and summarizing customer reviews', *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 168-177.

Hu, X., Tang, J., Gao, H. and Liu, H. (2013) 'Unsupervised sentiment analysis with emotional signals', *Proceedings of the 22nd international conference on World Wide Web.* ACM, 607-618.

Jakob, N., Weber, S.H., Müller, M.C. and Gurevych, I. (2009) 'Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations', *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.* ACM, 57-64.

Jason, W., Victoria, N., Dianne, M. and Sherri, L. (2014) 'Defining Patient Experience', *Patient Experience Journal,* 1(1), pp. 7-19.

Jiang, Y. and Saxena, A. (2013) 'Discovering Different Types of Topics: Factored Topic Models.', *IJCAI.*

Joshi, M. and Penstein-Rosé, C. (2009) 'Generalizing dependency features for opinion mining', *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.* Association for Computational Linguistics, 313-316.

Kalamatianos, G., Mallis, D., Symeonidis, S. and Arampatzis, A. (2015) 'Sentiment analysis of greek tweets and hashtags using a sentiment lexicon', *Proceedings of the 19th Panhellenic Conference on Informatics.* ACM, 63-

68.

Kasanen, E., Lukka, K. and Siitonen, A. (1993) 'The Constructive Approach in Management Accounting Research', *Journal of Management Accounting Research,* 5, pp. 243-264.

Kennedy, A. and Inkpen, D. (2006) 'Sentiment classification of movie reviews using contextual valence shifters', *Computational Intelligence,* 22(2), pp. 110-125.

Kim, S., Pantel, P., Chklovski, T. and Pennacchiotti, M. (2006) 'Automatically assessing review helpfulness', *Proceedings of the 2006 Conference on empirical methods in natural language processing.* Association for Computational Linguistics, 423-430.

Kiran, B.B.V. and Kumar, M.S. (2015) 'SOCIAL SUMMARIZATION FOR E-COMMERCE SITE BY EVALUATING THE COMMENTS AND RATINGS', .

Kiritchenko, S., Zhu, X. and Mohammad, S.M. (2014) 'Sentiment analysis of short informal texts', *Journal of Artificial Intelligence Research,* 50, pp. 723-762.

Kolyshkina, I., Levin, B. and Goldsworthy, G. (2013) 'Using Social Media Data for Comparing Brand Awareness, Levels of Consumer Engagement, Public

Opinion and Sentiment for Big Four Australian Banks.', .

Koo, T., Carreras Pérez, X. and Collins, M. (2008) 'Simple semi-supervised dependency parsing', *46th Annual Meeting of the Association for Computational Linguistics.* , 595-603.

KUBRA, A., KHAN, A. and FARHA, S. (2015) 'Modelling A Probabilistic Theory for Mining The Reviews on Pharmaceuticals', .

Kuechler, W.L. and Vaishnavi, V.K. (2008) 'The emergence of design research in information systems in North America', *Journal of Design Research,* 7(1), pp. pp1-16.

Kumar, V.R. and Raghuveer, K. (2013) 'Dependency driven semantic approach to product features extraction and summarization using customer reviews', in *Advances in Computing and Information Technology.* Springer, pp. 225-238.

Lane, J.V., Hamilton, D.F., MacDonald, D.J., Ellis, C. and Howie, C.R. (2016) 'Factors that shape the patient's hospital experience and satisfaction with lower limb arthroplasty: an exploratory thematic analysis', *BMJ open,* 6(5), pp. e010871-2015-010871.

Laukkanen, T. (2006) 'Customer-perceived value of e-financial services: a means-end approach', *International journal of electronic finance,* 1(1), pp. 5-17.

LaVela, S.L. and Gallan, A. (2014) 'Evaluation and measurement of patient experience', *Patient Experience Journal,* 1(1), pp. 28-36.

Lees, C. (2011) 'Measuring the patient experience', *Nurse researcher,* 19(1), pp. 25-28.

Li, F., Wang, S., Liu, S. and Zhang, M. (2014) 'SUIT: A Supervised User-Item Based Topic Model for Sentiment Analysis.', *AAAI.* , 1636-1642.

Li, T., Zhang, Y. and Sindhwani, V. (2009) 'A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge', *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1.* Association for Computational Linguistics, 244-252.

Li, W. and McCallum, A. (2006) 'Pachinko allocation: DAG-structured mixture models of topic correlations', *Proceedings of the 23rd international conference on Machine learning.* ACM, 577-584.

Lin, C., He, Y., Everson, R. and Ruger, S. (2012) 'Weakly supervised joint sentiment-topic detection from text', *IEEE Transactions on Knowledge and Data Engineering,* 24(6), pp. 1134-1145.

Lin, J. and Kolcz, A. (2012) 'Large-scale machine learning at twitter', *Proceedings of the 2012 ACM SIGMOD International Conference on Management of*

Data. ACM, 793-804.

Liu, B. (2015) *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press.

Liu, B. (2012) 'Sentiment analysis and opinion mining', *Synthesis lectures on human language technologies,* 5(1), pp. 1-167.

Liu, B. (2010) 'Sentiment analysis and subjectivity', *Handbook of natural language processing,* 2, pp. 627-666.

Liu, B. (2007) *Web data mining: exploring hyperlinks, contents, and usage data.* Springer Science & Business Media.

Liu, B., Blasch, E., Chen, Y., Shen, D. and Chen, G. (2013) 'Scalable sentiment classification for big data analysis using Naive Bayes Classifier', *Big Data, 2013 IEEE International Conference on.* IEEE, 99-104.

Liu, D., Shih, M., Liau, C. and Lai, C. (2009) 'Mining the change of event trends for decision support in environmental scanning', *Expert Systems with Applications,* 36(2), pp. 972-984.

Liu, K., Li, W. and Guo, M. (2012) 'Emoticon Smoothed Language Models for Twitter Sentiment Analysis.', *AAAI.*

Liu, Z. (2013) *High performance latent dirichlet allocation for text mining,* .

Lukka, K. (2003) 'The constructive research approach. In: Ojala, L. & Hilmola, O.-P. (eds.) Case study research in logistics. ', *Turku School of Economics and Business Administration,* .

Manary, M.P., Boulding, W., Staelin, R. and Glickman, S.W. (2013) 'The patient experience and health outcomes', *New England Journal of Medicine,* 368(3), pp. 201-203.

Manning, C.D. and Schütze, H. (1999) *Foundations of statistical natural language processing.* MIT Press.

March, S.T. and Smith, G.F. (1995) 'Design and natural science research on information technology', *Decision Support Systems,* 15, pp. 251-266.

Maynard, D. and Greenwood, M.A. (2014) 'Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis.', *LREC.* , 4238-4243.

McDonald, R., Hannan, K., Neylon, T., Wells, M. and Reynar, J. (2007) 'Structured models for fine-to-coarse sentiment analysis', *Annual Meeting-Association For Computational Linguistics.* Citeseer, 432.

Medhat, W., Hassan, A. and Korashy, H. (2014a) 'Sentiment analysis algorithms and applications: A survey', *Ain Shams Engineering Journal,* 5(4), pp. 1093-1113.

Medhat, W., Hassan, A. and Korashy, H. (2014b) 'Sentiment analysis algorithms and applications: A survey', *Ain Shams Engineering Journal,* 5(4), pp. 1093-1113.

Melville, P., Gryc, W. and Lawrence, R.D. (2009) 'Sentiment analysis of blogs by combining lexical knowledge with text classification', *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 1275-1284.

Meyer, C. and Schwager, A. (2007) 'Understanding Customer Experience. *Harvard Business Review*', *Harvard Business Review,* (Feb 2007).

Mingers, J., Mutch, A. and Willcocks, L. (2013) 'Critical realism in information systems research', *Mis Quarterly,* 37(3), pp. 795-802.

Mohr, J.W. and Bogdanov, P. (2013) 'Introduction—Topic models: What they are and why they matter', *Poetics,* 41(6), pp. 545-569.

Mukherjee, S. and Bhattacharyya, P. (2012) 'Feature specific sentiment analysis for product reviews', *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer, 475-487.

Mullen, T. and Collier, N. (2004) 'Sentiment Analysis using Support Vector Machines with Diverse Information Sources.', *EMNLP.* , 412-418.

Myslín, M., Zhu, S., Chapman, W. and Conway, M. (2013) 'Using twitter to examine smoking behavior and perceptions of emerging tobacco products', *Journal of medical Internet research,* 15(8), pp. e174.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F. and Stoyanov, V. (2016) 'SemEval-2016 task 4: Sentiment analysis in Twitter', *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming).*

NHS Choices (2016) *Reviews and ratings - Hillingdon Hospital - NHS Choices*. Available at: http://www.nhs.uk/Services/hospitals/ReviewsAndRatings/DefaultView.aspx?id=676 (Accessed: 0815 2016).

Nivre, J. and McDonald, R.T. (2008) 'Integrating Graph-Based and Transition-Based Dependency Parsers.', *ACL.* , 950-958.

Orlikowski, W.J. and Iacono, C.S. (2001) 'Research Commentary: Desperately Seeking the 'IT' in IT Research – A Call to Theorizing the IT Artifact', *Information Systems Research,* 12(2), pp. 121-134.

Ortigosa, A., Martín, J.M. and Carro, R.M. (2014) 'Sentiment analysis in Facebook and its application to e-learning', *Computers in Human Behavior,* 31, pp. 527-541.

Pang, B. and Lee, L. (2008a) 'Opinion mining and sentiment analysis', *Foundations and trends in information retrieval,* 2(1-2), pp. 1-135.

Pang, B. and Lee, L. (2008b) 'Opinion mining and sentiment analysis', *Foundations and trends in information retrieval,* 2(1-2), pp. 1-135.

Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2008) 'A Design Science Research Methodology for Information Systems Research', *Journal of Management Information Systems,* 24(3), pp. 45-78.

Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V. and Bragge, J. (2006) 'The design science research process: a model for producing and presenting information systems research', *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006).* , 83-106.

Pei, W., Ge, T. and Chang, B. (2015) 'An effective neural network model for graph-based dependency parsing', *Proc. of ACL.*

Picker Institute (2009) *Using patient feedback.* Oxford: Picker Institute Europe.

Poria, S., Cambria, E., Howard, N., Huang, G. and Hussain, A. (2016) 'Fusing audio, visual and textual clues for sentiment analysis from multimodal content', *Neurocomputing,* 174, pp. 50-59.

Poria, S., Cambria, E., Winterstein, G. and Huang, G. (2014) 'Sentic patterns: Dependency-based rules for concept-level sentiment analysis', *Knowledge-*

Based Systems, 69, pp. 45-63.

Prakash, S. and Gupta, P. (2015) 'Introduction', in *Ear Biometrics in 2D and 3D.* Springer, pp. 1-20.

Purver, M., Griffiths, T.L., Körding, K.P. and Tenenbaum, J.B. (2006) 'Unsupervised topic modelling for multi-party spoken discourse', *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 17-24.

Rajagopal, D., Olsher, D., Cambria, E. and Kwok, K. (2013) 'Commonsense-based topic modeling', *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining.* ACM, 6.

Ramage, D., Manning, C.D. and Dumais, S. (2011) 'Partially labeled topic models for interpretable text mining', *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 457-465.

Rehurek, R. and Sojka, P. (2010) 'Software framework for topic modelling with large corpora', *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Citeseer.

Rocha, C.G.D., Formoso, C.T., Tzortzopoulos-Fazenda, P., Koskela, L. and Tezel, A. (2012) 'Design Science Research in Lean Construction: Process and

Outcomes', *20th Annual conference, international group for lean construction.* San Diego, .

Rodden, K., Hutchinson, H. and Fu, X. (2010) 'Measuring the User Experience on a Large Scale: User-Centered Metrics for Web Applications', *Proceedings of CHI 2010.* ACM Press.

Rossetti, M., Stella, F. and Zanker, M. (2016) 'Analyzing user reviews in tourism with topic models', *Information Technology & Tourism,* 16(1), pp. 5-21.

Ryan, J.O., Kaltman, E., Mateas, M. and Wardrip-Fruin, N. (2015) 'What we talk about when we talk about games: Bottom-up game studies using natural language processing', *Proc.FDG,* .

Sarvabhotla, K., Pingali, P. and Varma, V. (2011) 'Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents', *Information Retrieval,* 14(3), pp. 337-353.

Schinas, M., Papadopoulos, S., Kompatsiaris, Y. and Mitkas, P.A. (2015) 'Visual event summarization on social media using topic modelling and graph-based ranking algorithms', *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval.* ACM, 203-210.

Schwartz-Shea, P. and Yanow, D. (2013) *Interpretive research design: Concepts and processes.* Routledge.

Sievert, C. and Shirley, K.E. (2014) 'LDAvis: A method for visualising and interpreting topics', .

Singh, V., Piryani, R., Uddin, A. and Waila, P. (2013) 'Sentiment analysis of Movie reviews and Blog posts', *Advance Computing Conference (IACC), 2013 IEEE 3rd International.* IEEE, 893-898.

Somprasertsri, G. and Lalitrojwong, P. (2010) 'Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization.', *J.UCS,* 16(6), pp. 938-955.

Steyvers, M. and Griffiths, T. (2007) 'Probabilistic topic models', *Handbook of latent semantic analysis,* 427(7), pp. 424-440.

Stoyanov, V. and Cardie, C. (2006) 'Partially supervised coreference resolution for opinion summarization through structured rule learning', *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 336-344.

Su, Q., Xiang, K., Wang, H., Sun, B. and Yu, S. (2006) 'Using pointwise mutual information to identify implicit features in customer reviews', in *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead.* Springer, pp. 22-30.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) 'Lexicon-based methods for sentiment analysis', *Computational linguistics,* 37(2), pp.

267-307.

The Beryl Institute (2014) *Defining Patient Experience - The Beryl Institute - Improving the Patient Experience*. Available at: http://www.theberylinstitute.org/?page=definingpatientexp (Accessed: 11/24 2015).

Titov, I. and McDonald, R. (2008) 'Modeling online reviews with multi-grain topic models', *Proceedings of the 17th international conference on World Wide Web.* ACM, 111-120.

Toikka, A. (2016) 'Uncovering Multi-Level Governance and Policy Idea Transfer in Energy Policy Using Topic Modelling on Large Policy Corpuses', *Third ISA Forum of Sociology (July 10-14, 2016).* Isaconf.

Tullis, T. and Albert, B. (2008) *Measuring the user experience.* 1st edn. Amsterdam: Elsevier.

Turney, P. (2001) 'Mining the web for synonyms: PMI-IR versus LSA on TOEFL', .

Underwood, T. (2012) 'Topic modeling made just simple enough', *The Stone and the Shell,* 7.

Uys, J., Du Preez, N. and Uys, E. (2008) 'Leveraging unstructured information using topic modelling', *PICMET'08-2008 Portland International Conference* *on Management of Engineering & Technology.* IEEE, 955-961.

Vaishnavi, V. and Kuechler, W. (2007) *Design Science Research methods and Patterns: Innovating Information and Communication Technology.* 1st edn. New York: Taylor & Francis Group.

Vaishnavi, V.K. and Kuechler, W. (2015) *Design science research methods and patterns: innovating information and communication technology.* Crc Press.

Varghese, R. and Jayasree, M. (2013) 'Aspect based Sentiment Analysis using support vector machine classifier', *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on.* IEEE, 1581-1586.

Vinodhini, G. and Chandrasekaran, R. (2012) 'Sentiment analysis and opinion mining: a survey', *International Journal,* 2(6).

Vohra, S. and Teraiya, J.B. (2013) 'A comparative study of sentiment analysis techniques', *International conference on Research and Development in Engineering and Sciences,* 2, pp. 1-6.

Wang, C., Blei, D. and Heckerman, D. (2012) 'Continuous time dynamic topic models', *arXiv preprint arXiv:1206.3298,* .

Wang, X., Wei, F., Liu, X., Zhou, M. and Zhang, M. (2011) 'Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach', *Proceedings of the 20th ACM international conference on Information and knowledge management.* ACM, 1031-1040.

Wieringa, R. (2009) 'Design science as nested problem solving', *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology.* Philadelphia, Pennsylvania, Philadelphia, Pennsylvania: ACM.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K. and Morency, L. (2013) 'Youtube movie reviews: Sentiment analysis in an audio-visual context', *IEEE Intelligent Systems,* 28(3), pp. 46-53.

Wu, Y., Zhang, Q., Huang, X. and Wu, L. (2009) 'Phrase dependency parsing for opinion mining', *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3.* Association for Computational Linguistics, 1533-1541.

Xu, K., Liao, S.S., Li, J. and Song, Y. (2011) 'Mining comparative opinions from customer reviews for Competitive Intelligence', *Decision Support Systems,* 50(4), pp. 743-754.

Yu, L., Wu, J., Chang, P. and Chu, H. (2013) 'Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news', *Knowledge-Based Systems,* 41, pp. 89-97.

Zhai, K., Boyd-Graber, J., Asadi, N. and Alkhouja, M.L. (2012) 'Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce', *Proceedings of the 21st international conference on World Wide Web.* ACM, 879-888.

Zhang, Y. and Clark, S. (2008) 'A tale of two parsers: investigating and combining graph-based and transition-based dependency parsing using beam-search', *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 562-571.

Zhuang, L., Jing, F. and Zhu, X. (2006) 'Movie review mining and summarization', *Proceedings of the 15th ACM international conference on Information and knowledge management.* ACM, 43-50.

Appendix A

- Human Judgment Survey for the Topic Modelling results.

The human judgement survey was conducted to gather expert opinion on the topics identified by the LDA topic modelling method. For the unigram method, they were presented with the terms identified for 30 topics. They were then asked to rate if they find the terms suitable to the topics. Similarly, for the bigram set, the terms identified were presented and the experts were asked to rate according to how suitable they found the terms to be for the topics. They were also then asked to rate their overall opinion about the topics and terms identified by both unigram and bigram modelling. The survey questionnaire screenshots are provided below.

Topic Modelling Results for the NHS patients' review datasets from (Jan 2010 to May 2015)

Human Judgment Survey for Unigram topics

In this survey questions, we would like know the human judgment about our research results of the LDA topic modelling algorithm. In our research, we have applied the LDA algorithm to the NHS patients' reviews open database. Data collected covers the period between January 2010 to July 2015.
Simply, you will be presented with 25 unigrams sets, each of which represents a topic.
You will be judging the quality of each topic, in other words, to what extent is each topic sensible using a scale from 5 (Very Rational) to 1 (Very Irrational).
A “Vary Rational” topic is one that is semantically coherent, sensible and meaningful. Also the topic will be clear and can easily be labeled. For example, for the following unigrams sets (January, February, March, April, May, June, July, August, September, October, November, December) we can describe this unigram set as “Very Rational” and easily we can label this topic as “Month Names”.
If some of the topic unigrams are coherent and interpretable but others are not then the topic would be scaled as 3 which is “Average”.
Finally, if the words appear random and unrelated to each other then the topic is scaled as 2 which is “Irrational” or as 1 which is “Very Irrational”.
Your answers will be highly anonymous and confidential, and will be used for research purposes only. In accordance with ethical guidance, all answers you provide will remain entirely anonymous.
Finally, you are free to withdraw at anytime by simply closing the web browser.
For more details, please feel free to contact with the students who curries out this survey or his Supervisor.

Student name: Mohammed Bahja

Student email: mohammed.bahja@brunel.ac.uk

Supervisor name: Prof Mark Lycett

Supervisor email: mark.lycett@brunel.ac.uk

1. Unigram Sets :

1

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
1- {birth, baby, midwives, midwife, labour, maternity, care, delivery, ward, staff, pregnancy, amazing, section, experience, daughter, support, fantastic, son, child, born, natal, felt, antenatal, unit, recommend}	○	○	○	○	○
2- {pain, told, doctor, antibiotics, nurse, hours, relief, home, infection, gp, severe, medication, back, blood, asked, night, dr, ambulance, painkillers, didnt, morphine, chest, agony, worse, days }	○	○	○	○	○
3- {im, dentist, dont, ive, dental, tooth, didnt, told, pain, feel, teeth, life, mental, people, back, health, wisdom, wasnt, id, worse, wrong, ill, bad, years, hospital}	○	○	○	○	○
4- {operation, staff, replacement, hip, knee, care, excellent, surgery, surgeon, op, ward, treatment, hospital, consultant, recommend, team, nursing, stay, hall, received, food, recovery, pre, physio, post}	○	○	○	○	○
5- {blood, scan, test, tests, results, told, doctor, ct, gp, ultrasound, pregnancy, pregnant, mri, nurse, ecg, weeks, bleeding, miscarriage, pressure, asked, bloods, scans, wait, back, appointment }	○	○	○	○	○
6- {food, ward, staff, room, stay, tea, hot, clean, meals, patients, eat, good, meal, menu, choice, night, coffee, drinks, tv, offered, drink, cup, water, breakfast, bed}	○	○	○	○	○
7- {waiting, told, wait, nurse, room, doctor, hours, reception, asked, receptionist, rude, hour, waited, minutes, patients, people, area, desk, called, appointment, pm, quot, sat, didnt, triage }	○	○	○	○	○
8- {eye, clinic, eyes, vision, cataract, drops, sight, consultant, treatment, operation, doctor, optician, moorfields, appointment, laser, problem, surgery, department, retina, condition, ophthalmology, glaucoma, left, glasses, prescription }	○	○	○	○	○
9- {procedure, ease, staff, made, feel, explained, nurse, friendly, lovely, nervous, endoscopy, felt, professional, doctor, experience, put, comfortable, kind, reassuring, colonoscopy, caring, questions, team, nurses, appointment}	○	○	○	○	○
10- {son, daughter, childrens, child, children, year, parents, sons, daughters, paediatric, boy, play, nurse, ward, school, staff, amp, nurses, care, doctor, month, amazing, fantastic, parent, baby}	○	○	○	○	○
11- {amp, ambulance, staff, admitted, wife, tests, care, ward, paramedics, excellent, chest, treatment, quickly, doctor, professional, crew, arrived, stroke, night, received, husband, sunday, morning, emergency, hours}	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
12- {care, staff, team, professional, nhs, professionalism, patients, excellent, nursing, ward, patient, caring, treatment, kindness, high, received, support, medical, experience, level, impressed, positive, praise, service, outstanding}	○	○	○	○	○
13- {pain, consultant, condition, years, life, hearing, treatment, gp, months, physio, referred, symptoms, clinic, appointment, problem, diagnosis, diagnosed, specialist, mri, problems, medication, ent, ear, doctor, disease}	○	○	○	○	○
14- {care, ward, staff, mum, mother, family, respect, treated, father, dignity, admitted, dad, kindness, caring, nurses, received, excellent, compassion, nursing, wonderful, kind, stay, team, doctors, amp}	○	○	○	○	○
15- {fracture, ray, broken, pain, foot, ankle, leg, arm, knee, injury, clinic, wrist, plaster, amp, told, cast, back, shoulder, crutches, bone, accident, fractured, xray, rays, walk}	○	○	○	○	○
16- {cancer, breast, treatment, consultant, surgery, operation, diagnosed, surgeon, removed, biopsy, team, chemotherapy, bladder, bowel, lump, weeks, scan, tumour, care, clinic, urology, remove, kidney, chemo, diagnosis}	○	○	○	○	○
17- {heart, care, cardiac, attack, team, ward, staff, life, treatment, admitted, received, excellent, angiogram, husband, operation, saved, icu, transferred, pacemaker, cardiology, coronary, fitted, class, doctors, ccu}	○	○	○	○	○
18- {bed, ward, nurse, room, toilet, floor, left, patients, blood, nurses, told, night, asked, dirty, water, staff, patient, chair, clean, quot, cleaned, didnt, sick, hours, urine}	○	○	○	○	○
19- {staff, surgery, operation, ward, surgeon, theatre, op, ease, day, team, care, friendly, procedure, caring, anaesthetist, nurses, recovery, excellent, made, professional, stay, consultant, kind, nervous, feel}	○	○	○	○	○
20- {mother, mum, father, ward, dad, home, told, bed, admitted, elderly, family, care, sister, dementia, law, discharged, left, brother, staff, medication, stroke, discharge, hospital, year, days }	○	○	○	○	○
21- {appointment, told, letter, weeks, consultant, waiting, gp, months, appointments, cancelled, clinic, results, date, scan, wait, referral, phone, referred, call, week, secretary, list, received, back, booked }	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
22- {operation, op, ward, surgery, pm, told, theatre, day, bed, nurse, surgeon, home, pre, procedure, discharge, hours, waiting, recovery, morning, room, arrived, wait, discharged, back, asked }					
23- {ward, staff, nurses, stay, care, caring, amazing, admitted, friendly, fantastic, night, hard, doctors, made, work, patients, team, lovely, brilliant, nursing, great, kind, job, people, feel}					
24- {hospital, staff, care, royal, hospitals, nhs, service, years, treatment, excellent, treated, london, north, bad, local, clean, doctors, received, people, amp, ive, class, recommend west, patient}					
25- {staff,friendly,appointment,helpful,professional,clinic, efficient, department, excellent, treatment, service, impressed, consultant, attended, clean, polite, pleasant, explained reception, caring, appointments, waiting,experience,visit,received}					
26- {baby, midwife, labour, birth, told, midwives, ward, pain, delivery, hours, maternity, section, partner, contractions, room, didnt, waters, home, epidural, born, pregnant, midwifes, wasnt, child, night }					
27- {nurse, injuries, minor, amp, ray, service, friendly, injury, minutes, staff, efficient, triage, wait, unit, wound, helpful, excellent, quickly, receptionist, professional, attended, hour, reception, doctor, son}					
28- {parking, car, park, pay, entrance, hospital, people, appointment, main, disabled, find, walk, space, reception, building, area, ticket, system, patients, waiting, money, machine, minutes, charges, free}					
29- {phone, call, appointment, number, answer, telephone, ring, message, contact, rang, told, calls, answered, called, department, switchboard, service, person, quot, back, appointments, minutes, phoned, letter, line}					
30- {patients, patient, quot, staff, lack, care, poor, information, communication, medical, rude, health, experience, nhs, complaint, management, issues, treatment, people, dont, medication, nursing, mental, basic, attitude}					

4

The

Topic Modelling Results for the NHS patients' review datasets from (Jan 2010 to May 2015)

Overall **Unigram** Sets Judgment

2. Overall Unigram Sets Judgment

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Overall , to what extent would you agree/disagree that the ***Unigram*** Sets has been classified sensibly ?	○	○	○	○	○

Other (please specify)

Topic Modelling Results for the NHS patients' review datasets from (Jan 2010 to May 2015)

Human Judgment Survey for **Bigram** topics

3. Bigram Sets :

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
1- waiting_room waiting_area waiting_time waiting_hours wait_hours long_wait people_waiting told_wait triage_nurse waited_hours half_hours hour_wait half_hour left_waiting hours_waiting back_waiting sat_waiting waited_hour wait_waiting reception_desk hours_doctor waiting_times patients_waiting reception_staff waiting_hour					

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
mental_health member_staff made_feel staff_rude extremely_rude dont_care staff_member customer_service formal_complaint family_member health_issues health_care eye_contact rude_unhelpful left_feeling people_skills crisis_team staff_dont treat_people members_staff wasting_time treat_patients reception_staff health_problems 2- care_patients	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
car_park car_parking appointment_time main_entrance running_late clinic_running parking_space parking_charges minutes_late blue_badge waiting_area car_parks parking_ticket find_parking half_hour disabled_parking minutes_appointment back_car member_staff appointment_times main_reception hospital_car park_car late_appointment hour_late	○	○	○	○	○

3-

mother_admitted
mother_law
care_home
year_mother
father_admitted
elderly_mother
father_law
end_life
family_members
nursing_home
mum_admitted
year_father
admitted_ward
admitted_hospital
nursing_staff
care_mother
mother_received
stroke_unit
family_member
care_mum
spent_weeks
mother_hospital
elderly_father
staff_ward
4- moved_ward

(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
made_feel feel_ease feel_comfortable make_feel feel_relaxed feel_safe staff_made felt_safe staff_friendly making_feel safe_hands good_hands put_ease made_comfortable didnt_feel member_staff felt_comfortable staff_amazing feel_cared friendly_made answered_questions felt_looked hospital_clean staff_lovely felt_cared 5-	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
care_received					
care_attention					
excellent_care					
level_care					
standard_care					
nursing_staff					
high_standard					
treatment_received					
received_excellent					
attention_received					
speak_highly					
quality_care					
high_level	○	○	○	○	○
care_treatment					
care_provided					
staff_involved					
care_support					
treatment_care					
involved_care					
care_staff					
medical_staff					
member_staff					
nursing_care					
members_staff					
6- high_standards					

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
7- asked_nurse moved_ward told_nurse told_home nil_mouth pain_relief eat_drink back_ward night_staff ward_told nurses_station nurse_told staff_nurse food_drink admitted_ward side_room left_hours ward_pm pm_told blood_pressure nurse_asked back_bed cup_tea till_pm half_hour	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
eye_clinic eye_hospital waiting_list pre_op operation_cancelled left_eye eye_drops eye_department cataract_operation eye_unit told_operation waiting_operation waiting_time cataract_surgery operation_date eye_casualty operation_told cancelled_due op_cancelled day_operation operation_day attended_eye years_ago detached_retina 8- phone_call	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
9- local_hospital university_hospital royal_london st_georges back_hospital admitted_hospital chase_farm queen_elizabeth hospital_staff royal_infirmary london_hospital hospital_hospital middlesex_hospital st_thomas queens_hospital hospital_days hospital_years bank_holiday royal_hospital west_middlesex king_george blackburn_hospital treatment_hospital georges_hospital st_james	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)

good_work
bad_press
hospital_staff
hospital_clean
hospital_good
staff_friendly
local_hospital
recommend_hospital
nhs_staff
good_hospital
good_service
general_hospital
health_service
great_job
service_received
people_quick
nhs_hospital
negative_comments
service_staff
clean_staff
health_care
horror_stories
hard_work
service_provided
10- staff_great

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
11- follow_appointment received_letter mri_scan appointment_weeks waiting_list appointment_cancelled appointment_told told_appointment appointment_letter make_appointment appointment_made appointment_consultant weeks_appointment wait_weeks received_appointment phone_call appointment_time weeks_ago appointment_months consultants_secretary months_appointment waiting_appointment wait_months appointment_week weeks_time	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
12- put_ease made_feel put_mind explained_procedure feel_ease mind_rest endoscopy_unit member_staff cup_tea mind_ease procedure_carried procedure_explained explained_happen staff_put nursing_staff friendly_professional staff_friendly carried_procedure procedure_staff explained_detail friendly_put answered_questions lovely_nurse putting_ease staff_nurse	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
13- childrens_ward year_son year_daughter childrens_aampe daughter_admitted care_daughter son_admitted care_son childrens_ae month_son childrens_hospital daughter_received month_daughter time_daughter son_received head_injury hospital_daughter yr_son staff_childrens daughter_treated admitted_childrens son_treated son_recently ae_department told_daughter	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
14- call_back phone_call answer_phone phone_number phone_calls telephone_number make_appointment called_back leave_message ring_back left_message hearing_aid told_call answered_phone book_appointment number_call main_switchboard telephone_call put_phone contact_details received_phone call_day contact_hospital change_appointment phone_back	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
15- labour_ward gave_birth delivery_suite maternity_ward maternity_unit giving_birth give_birth post_natal birth_centre breast_feeding care_received baby_born weeks_pregnant baby_boy birth_son emergency_section baby_girl birth_baby birth_child antenatal_care natal_ward st_marys maternity_services waters_broke postnatal_ward	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
ambulance_crew					
early_hours					
ambulance_staff					
doctors_nurses					
ae_department					
accident_emergency					
aampe_department					
nursing_staff					
staff_aampe					
treatment_received					
nurses_doctors					
chest_pains					
ambulance_service	○	○	○	○	○
assessment_unit					
admitted_aampe					
ambulance_arrived					
care_received					
staff_ae					
sunday_morning					
called_ambulance					
emergency_department					
ct_scan					
aampe_staff					
admitted_ae					
16- saturday_morning					

22

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
17- pain_relief pain_killers severe_pain lot_pain back_pain pain_management extreme_pain abdominal_pain pain_told excruciating_pain ct_scan due_pain asked_pain severe_abdominal hours_pain pain_pain told_pain pain_back home_pain left_pain pain_hours great_pain pain_left crying_pain chronic_pain	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
breast_cancer breast_care treatment_received consultant_team years_ago breast_clinic specialist_nurse prostate_cancer bowel_cancer past_years diagnosed_breast royal_free cancer_treatment speak_highly day_unit waiting_times care_treatment excellent_treatment care_unit past_months side_effects charing_cross nurse_specialist treatment_care 18- royal_preston	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
staff_friendly					
friendly_helpful					
staff_helpful					
reception_staff					
friendly_professional					
extremely_helpful					
waiting_time					
staff_polite					
excellent_service					
found_staff					
staff_extremely					
service_received					
helpful_friendly					
friendly_efficient					
member_staff					
staff_efficient					
polite_helpful					
appointment_time					
staff_professional					
professional_friendly					
time_explain					
efficient_friendly					
polite_friendly					
staff_pleasant					
19- extremely_impressed					

(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
○	○	○	○	○

doctor_told
told_doctor
hours_doctor
waited_hours
told_home
nurse_told
weeks_pregnant
urine_sample
told_back
wait_hours
told_wait
triage_nurse
doctor_didnt
doctor_asked
early_pregnancy
nurse_doctor
hours_told
made_feel
doctor_doctor
blood_clot
scan_told
junior_doctor
waiting_hours
told_wrong
asked_doctor

20-

26

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
staff_amp care_amp amp_staff st_richards amp_department friendly_amp nurses_amp weeks_ago hospital_amp bishop_auckland amp_caring couple_weeks doctors_amp ward_amp time_amp amp_told caring_amp amp_made st_marks amp_nurses stoke_mandeville attended_amp amp_professional amp_dept admitted_amp	○	○	○	○	○

21-

22-	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
hip_replacement knee_replacement nursing_staff recommend_hospital care_received treatment_received total_hip total_knee start_finish nhs_patient horder_centre staff_excellent consultant_team highly_recommend replacement_operation hall_hospital post_op received_excellent goring_hall food_good care_excellent private_hospital treatment_care care_attention replacement_surgery	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
23- day_surgery surgery_unit day_case theatre_staff post_op general_anaesthetic back_ward recovery_room nursing_staff surgical_team pre_op staff_ward ward_staff day_care operating_theatre staff_day admitted_day day_patient day_operation surgery_day surgeon_anaesthetist gall_bladder day_unit surgeon_team surgery_ward	○	○	○	○	○

29

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
24- heart_attack treated_respect treated_utmost dignity_respect treated_dignity care_unit respect_dignity utmost_respect intensive_care treated_kindness staff_treated care_respect treated_care saved_life coronary_care treated_great utmost_care doctors_nurses respect_staff care_received kindness_respect nursing_staff critical_care respect_care treated_upmost	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
dental_hospital waste_time left_feeling wisdom_tooth read_notes made_feel years_ago wasting_time ent_department asked_questions medical_history referred_hospital story_short consultant_told nose_throat long_story ear_nose follow_appointment full_time junior_doctor complete_waste answer_questions side_effects cut_long 25- gp_referred	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
26- minor_injuries fracture_clinic injuries_unit triage_nurse urgent_care nurse_practitioner minor_injury care_centre walk_centre xray_department plaster_room appointment_fracture excellent_service visited_minor aampe_department suspected_broken short_wait attended_minor broken_ankle injury_unit broken_wrist half_hour staff_friendly friendly_helpful plaster_cast	○	○	○	○	○

32

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
27- staff_ward member_staff doctors_nurses nurses_doctors hard_work staff_amazing care_received ward_staff difficult_time kind_caring work_hard praise_staff hard_working special_mention staff_work fantastic_care absolutely_fantastic fantastic_job amazing_job staff_fantastic wonderful_care care_staff health_care wonderful_staff staff_absolutely					

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
staff_ward					
nursing_staff					
ward_staff					
food_good					
ward_clean					
stay_ward					
staff_friendly					
admitted_ward					
care_received					
night_staff					
spent_days					
short_stay	○	○	○	○	○
cleaning_staff					
catering_staff					
stay_hospital					
staff_helpful					
staff_caring					
day_night					
doctors_nurses					
staff_nurses					
ward_care					
spotlessly_clean					
care_assistants					
28- made_stay					

34

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
29- blood_test blood_tests blood_pressure ct_scan test_results chest_xray chest_pain chest_pains results_back blood_results taking_blood heart_attack blood_sample nurse_blood high_blood blood_samples ecg_blood results_blood told_blood hospital_blood blood_sugar low_blood blood_clot heart_rate waiting_blood	○	○	○	○	○

	(Very Rational)	(Rational)	(Average)	(Rational)	(Very Rational)
nursing_staff					
lack_communication					
patient_care					
staff_patients					
lack_care					
elderly_patients					
care_patients					
nursing_care					
communication_staff					
medical_staff					
poor_communication					
patients_relatives					
family_member	○	○	○	○	○
ward_staff					
due_lack					
infection_control					
lack_staff					
nurses_station					
basic_care					
patients_ward					
patients_left					
patients_staff					
staff_time					
patients_care					
30- level_care					

36

Topic Modelling Results for the NHS patients' review datasets from (Jan 2010 to May 2015)

Overall Human Judgment for Bigram topics

4. Overall Bigram Sets Judgment

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
Overall , to what extent would you agree/disagree that the ***Bigram*** Sets has been classified sensibly ?	○	○	○	○	○

Other (please specify)

Appendix B

The logic of generating the wordcloud visualisation using the MySQL procedure is presented below.

```
ALTER PROCEDURE [dbo].[USP_GET_TOPIC]
@Topic INT,
@pairType CHAR(3),
@isPosotive TINYINT
AS
BEGIN
SET NOCOUNT ON;
DECLARE @result TABLE(topic int, word varchar(500), Total int)
DECLARE @total DECIMAL(9,2)
DECLARE @gt INT
DECLARE @lt INT
SET @gt=-500000
SET @lt=0

IF @isPosotive = 1
BEGIN
SET @gt=-1
SET @lt=500000
END

IF @pairType ='adj'
BEGIN
INSERT INTO @result
SELECT
top 20 topic
,replace(coreNLP_pair,',',' ') word
,COUNT(*) Total
FROM ParseTopicData_V2
WHERE
coreNLP_pair<>'NA'
AND Topic=@Topic
AND coreNLP_pair_score >@gt
AND coreNLP_pair_score <@lt
GROUP BY
Topic,
```

```
replace(coreNLP_pair,',',' ')
ORDER BY
topic,
COUNT(*) DESC
END
ELSE IF @pairType ='nlp'
BEGIN
INSERT INTO @result

SELECT
top 20 topic
,replace(open_nlp_pair,',',' ') word
,COUNT(*) Total
FROM ParseTopicData_V2
WHERE
open_nlp_pair<>'NA'
AND Topic=@Topic
AND open_nlp_pair_score >@gt
AND open_nlp_pair_score <@lt
GROUP BY
Topic,
replace(open_nlp_pair,',',' ')
ORDER BY
topic,
COUNT(*) DESC
END


SELECT @total=SUM(Total) from @result

SELECT topic,word,Total TotalCount,convert(int,(Total/1358.0*100)) Total
from @result
```