TEXT MINING PATIENT EXPERIENCES FROM ONLINE HEALTH COMMUNITIES

by

Mark Greenwood

This thesis is submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

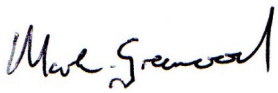
The School of Computer Science & Informatics
The Cochrane Institute of Primary Care and Public Health

Cardiff University
November 2015

DECLARATION

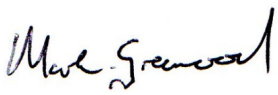
Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, not is being submitted concurrently in candidature for any degree or other award

Signed  Date: 17/11/2015

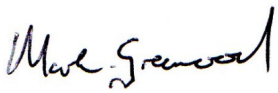
Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed  Date: 17/11/2015

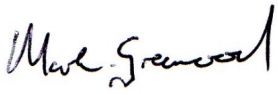
Statement 2

This thesis is the result of my own independent work, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed  Date: 17/11/2015

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed  Date: 17/11/2015

DEDICATION

To my Grandad
who lived for his family

ACKNOWLEDGMENTS

First and foremost, I have to say a huge thank you to my supervisors. I am privileged to have been given the opportunity to undertake this work and to have been guided through it by such enthusiastic, knowledgeable researchers. Irena, Alun, Nick and Glyn, thank you very much – you have each taught me a great deal and contributed in no small measure to the completion of this work.

Thanks also to my examiners - Professors Frans Coenen and Alex Gray. Their considered and valuable input has helped to make this piece of work something of which I am very proud.

I'd also like to thank the rest of the departments at the School of Computer Science and Informatics and the Institute of Primary Care and Public Health. I have never had an opportunity to work in such a creative environment and to collaborate on such interesting work as the time I have enjoyed in Cardiff.

Thanks also to Cardiff University and the President's Scholarship which funded this research.

My family have always stood by and supported me, and deserve a great deal of gratitude. It's not always easy, but I know their support will always be there and I couldn't have done this without them. Most of all, my Mum and Dad, Elizabeth and Eric, and my partner Victoria have each supported me through what has been a testing, but valuable experience. I love you all greatly.

Last but not least, thank you to my friends for the good times, for their input and their support. I am a very lucky person.

ABSTRACT

Social media has had an impact on how patients experience healthcare. Through online channels, patients are sharing information and their experiences with potentially large audiences all over the world. While sharing in this way may offer immediate benefits to themselves and their readership (e.g. other patients) these unprompted, self-authored accounts of illness are also an important resource for healthcare researchers. They offer unprecedented insight into understanding patients' experience of illness. Work has been undertaken through qualitative analysis in order to explore this source of data and utilising the information expressed through these media. However, the manual nature of the analysis means that scope is limited to a small proportion of the hundreds of thousands of authors who are creating content.

In our research, we aim to explore utilising text mining to support traditional qualitative analysis of this data. Text mining uses a number of processes in order to extract useful facts from text and analyse patterns within – the ultimate aim is to generate new knowledge by analysing textual data *en mass*.

We developed QuTiP – a Text Mining framework which can enable large scale qualitative analyses of patient narratives shared over social media. In this thesis, we describe QuTiP and our application of the framework to analyse the accounts of patients living with chronic lung disease. As well as a qualitative analysis, we describe our approaches to automated information extraction, term recognition and text classification in order to automatically extract relevant information from blog post data. Within the QuTiP framework, these individual automated approaches can be brought together to support further analyses of large social media datasets.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Term	Acronym	Brief Description
Automatic Term Recognition	ATR	A class of approaches which aim to automatically identify mentions of domain-specific concepts in text without the use of external lexicons or terminologies.
Blog		A website whereby user-authored articles (or posts) are collected and displayed to a potentially large audience.
Blog post		An individual article which is written on and available through a blog.
Chronic Obstructive Pulmonary Disease	COPD	Progressive condition which causes limitation of airflow within the lung which is not fully reversible.
COPD Exacerbation		An acute worsening of a COPD patient's respiratory symptoms. For instance, shortness of breath or a productive cough.
Cohen's Kappa	κ	Measure of agreement between two participants in an annotation task where each person labels each available item in the collection.
Crowdsource		An approach whereby a large task is split in to smaller parts and distributed among a number of distributed participants across the Web.
Data Mining	DM	A class of automated approaches which aim to help identify patterns and trends in data in order to help users extract new knowledge from large data collections.
Exacerbation		An acute worsening of the symptoms of a condition.
Forced Expiratory Volume in one Second Test	FEV_1	Respiratory test which measures the amount of air a subject can expel within one second.
Forced Vital Capacity Test	FVC	Respiratory test which measures the greatest volume of air that can be expelled by the subject.

Health 2.0		**Description of the impact that modern web-based tools have had on health care. Namely, tools that have enabled online authorship without extensive technical expertise – e.g. social media platforms.**
Information Extraction	IE	A set of Text Mining approaches which deal with identifying mentions of pre-defined classes of entities and relationships between entities from within unstructured, textual data.
Information Retrieval	IR	Approaches to identify documents relevant to some information need inside a collection of both relevant and irrelevant documents.
Inter-annotator agreement		Statistical measures of agreement between independent annotators of some data items. Measure coherence of the labels produced, often compared to agreement through random chance. (*see Krippendorff's Alpha, Cohen's Kappa*)
Linguistic/ Terminological Variation		Summarises a number of ways that language can vary – from inflections and tenses of words to synonymy and polysemy of phrases used to refer to concepts.
Krippendorff's Alpha	α	An inter-annotator agreement measure which can be tuned to a number of situations – where there are many annotators, and where each annotator has only labelled a portion of the available data.
Lexicon		A list of terms, which can be focused within one particular domain.
Machine Learning		Approaches which, when trained on previously seen data, can produce models which can inform users about new, unseen data.
Medicine 2.0		See *Health 2.0*
The National Institute for Health and Clinical Excellence	NICE	Public body which provides national guidance and advice to improve health and social care.
Objectivity		Describes a statement which relates to some general information. Opposite to *Subjectivity*.
Part-of-speech Tag	POS	Labels assigned to tokens (e.g. words) which describe the role it plays within a sentence. For instance *noun*, *adjective*, *verb*.
Part-of-speech Tagging	POS tagging	The process of assigning labels to tokens based on their expected role within the sentence. *See Part-of-speech tag*.

Patient Experience		The subjective experiences of an individual who is being treated for, or living with some medical condition.
Pulmonary Rehabilitation (Rehab)		Ongoing treatment for people living with COPD which involves a regime of supervised exercises designed to increase respiratory capacity.
Qualitative Analysis		Research methodologies which enable researchers to systematically interpret subjective data - for instance self-authored patient narratives.
Semantic information		Related to some real-world concept.
Sentiment analysis		Analysis of emotions information which is conveyed - in this document, by the author through text.
Social Media		Online platforms through which non-technical people can share content with a potentially large audience. May also support responses from that audience (i.e. through responses).
Subjectivity		Describes a statement which relates to the author – i.e. conveys a personal experience/opinion. Opposite to *Objectivity*.
Support Vector Machine	SVM	
Term		A word or phrase which represents some real-world concept.
Terminology		A coherent group of domain-specific terms. Potentially includes relationships between terms.
Text Mining		Knowledge intensive process whereby users can interact with a document collection using a suite of analysis tools.
Unified Medical Language System	UMLS	A collection of medical terminologies grouped. Includes a semantic network which relationships between terms.

CHAPTER 1

INTRODUCTION

User-generated content and the popularity of social media have enabled people to communicate online with potentially large audiences all over the world. The wide range of social media platforms which are available enable users to create and share content online in many ways, including social networking sites (e.g. Facebook, Google), content communities (e.g. YouTube, Instagram), collaborative projects (e.g. Wikipedia), blogging platforms (such as Blogger, Wordpress, etc.) and microblogging platforms (i.e. Twitter) (Kaplan & Haenlein, 2010). A recent survey of over 5000 adult internet users in the USA revealed that 73% use social network sites and 42% use more than one (Duggan & Smith, 2013). The authors asked roughly 1500 participants about their engagement with 5 particular sites which each showed different levels of engagement (see Figure 1).

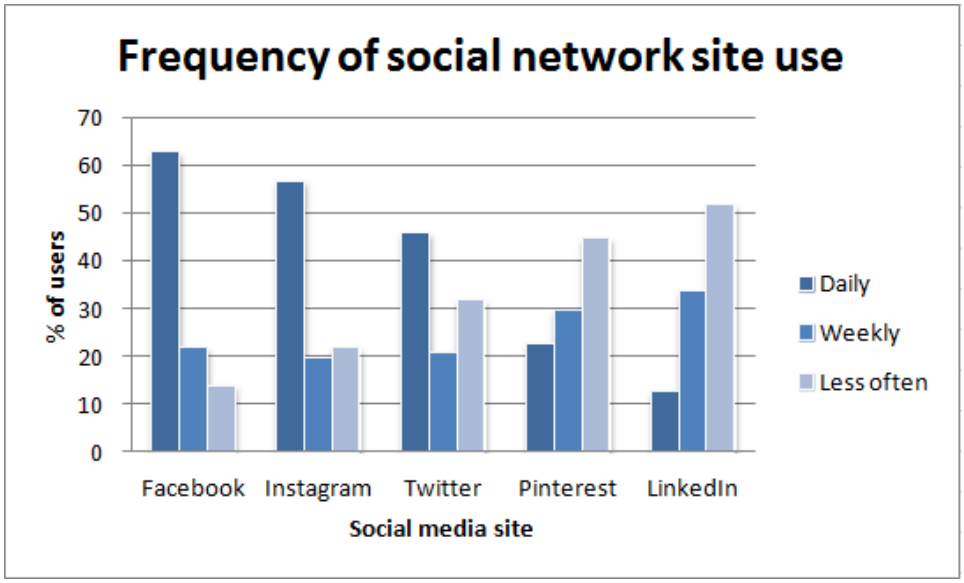


Figure 1: Social networking site engagement - adapted from (Duggan & Smith, 2013)

Three of the five social networking sites (Facebook, Instagram and Twitter) were engaged with by roughly 45-60% of users on a daily basis, showing just how important these channels of communication have become.

Social media have arguably had an impact on many facets of modern life, but it is the impact on healthcare that we focus on here. Patients are using social media as a means to find and share health information. Research by NM Incite (a market research company focusing on social media) have described an online community of hundreds of thousands of people taking part in an online discussion about a wide range of health conditions (see Table 1).

Condition	Prevalence	Authors
Cardiovascular disease	81,000,000	65,239
Arthritis	52,100,000	37,975
Asthma	24,400,000	45,633
COPD	24,000,000	12,445
Type 2 Diabetes	16,000,000	23,927
Depression	14,800,000	174,217
Alzheimer's Disease	5,300,000	36,175
Attention Deficit Disorder	5,300,000	115,296
Fibromyalgia	5,000,000	70,943
Breast Cancer	2,605,000	51,990
Prostate Cancer	2,276,000	19,940
Rheumatoid Arthritis	1,293,000	9,619
Colorectal Cancer	1,112,000	17,005
Ulcerative Colitis	619,000	13,718
Lung Cancer	371,000	23,746
Multiple Sclerosis	350,000	19,277
Type 1 Diabetes	340,000	14,241
Ovarian Cancer	177,000	11,585
Brain Cancer	126,000	13,891
Cystic Fibrosis	30,000	3,336

Table 1: Health and Social Media: Condition Prevalence And Online Authors – adapted from (NM Incite, 2011)

NM Incite also looked at who was taking part in online health discussions (see Figure 2). They show that as well as many participants, this online conversation is mainly patient led, with only a small percentage of the authorship made up of health professionals.

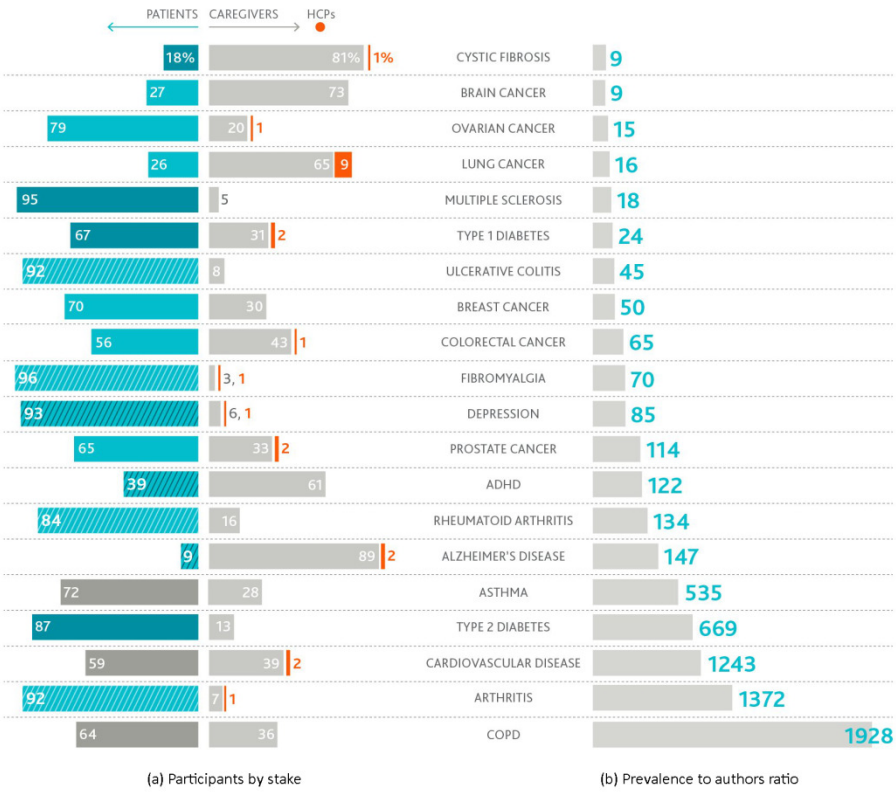


Figure 2: Health and social media: authors participation per condition by stake - adapted from (NM Incite, 2011)

The social media channels are used to share information, opinion and experiences. The information shared through these channels has the potential to inform other patients, healthcare professionals and researchers about living with these conditions. The aim of this work is to implement a text mining approach that will help unlock this information from large volumes of user-generated text data, providing faster and easier access to relevant text content shared by patients and carers. The following section will introduce and explore the existing impact of social media on healthcare and the potential for enhancing access to this rich source of information.

1.1 PATIENT EXPERIENCE AND ITS ROLE IN HEALTHCARE

Patient experience is an increasingly important factor within healthcare. Boote, Telford, and Cooper (2002) highlight the importance of patient experience by drawing on the definitions of two terms. They discussed the separation of 'disease' (i.e. a physiologic and clinical abnormality) from 'illness' defined as the subjective experience of the affected individual (Ong, 1996). While healthcare professionals can be considered the experts in disease, patients and healthcare consumers can be considered experts in illness – in the experience of disease (J. Wilson, 1999). Boote et al. argue that when both healthcare

professionals and patients' perspectives are considered together, they may offer a 'synergistic relationship' which could offer new insights towards 'improving the condition of the consumer'. Patient experience is used, and has proved useful, in a number of areas.

Firstly, at the individual level – patient-centered care has been championed for some years within the healthcare industry (Bauman, Fardy, & Harris, 2003) and there is evidence that it can improve outcomes for patients, including quality of life, patient satisfaction and anxiety reduction (Stewart, 2001). This paradigm of patient-physician interaction focuses on the inclusion of the patient in decisions about their care, and in relating the evidence about disease to their experience of illness in order to help make the most appropriate, personalised treatment decisions for that individual. As such, involvement of the patient and their family in decisions about their healthcare is one of the 7 key principles guiding the NHS, set out in their constitution (Department of Health, 2010). This has also been extended to decisions at the organizational level. The NHS' white paper 'Our Health, Our Care, Our say' (Department of Health, 2006) sets out how patients and communities will be involved in making decisions about local healthcare provision.

Another important area in which healthcare consumers' opinions are being sought is in medical research. Consumers have been both passively engaged with and actively participated in the instigation, design and execution of research (Boote et al., 2002). Studies have shown that health professionals have a poor understanding of their patient's views and that researchers do not necessarily represent the views of healthcare consumers (Coulter, Peto, & Doll, 1994; Dolan, Bordley, & Miller, 1993; Hares, Spencer, Gallagher, Bradshaw, & Webb, 1992) and so, by involving patients in setting research priorities and utilising their expertise of living with conditions, researchers aim to focus on areas which will have the biggest impact on patients' lives. The James Lind Alliance[1] (JLA) is an organisation that supports bringing patients, carers and clinicians together to involve all stakeholders in setting research priorities according to patient needs. The JLA have a structured method for engaging with healthcare consumers in order to set research priorities (Cowan, 2010). This method has been used to identify areas in accordance with patients' most pressing needs in areas including preterm births (Petit-Zeman & Uhm, 2012), strokes (Pollock, St George, Fenton, & Firkins, 2012, 2014), diabetes (Gadsby et al., 2012) and prostate cancer (Lophatananon et al., 2011; Tyndale-Biscoe, Malcolm, & Gnanapragasam, 2012).

[1] http://www.lindalliance.org/

In each of these scenarios, part of the challenge is how best to capture and represent patient experience. Gathering data about patient experience, whether at an individual or population level, can be approached in a number of ways. Face-to-face consultations are one approach utilised, for instance, by the NHS for stakeholder meetings to make decisions regarding local health services (Department of Health, 2004). Focus groups (Kitzinger, 1995) and structured interviews (Brédart, Marrel, Abetz-Webb, Lasch, & Acquadro, 2014; Hafeez et al., 2012; Rankin et al., 2014) have also been used to explore patient experience in recent qualitative research.

Questionnaires can also be used to gather more structured information relating to patient experience. Patient reported outcome measures (PROMs) are tools designed to capture patients' perspectives of health, illness and healthcare in a reliable way used to evaluate patients' perspectives on the quality of their care (Hibbard, 2003). One widely used example is the General Health Questionnaire (Goldberg, 1978). PROMs have been implemented to assess treatments from the patients' perspective in a number of areas (Meadows, 2011) including within the NHS (Department of Health, 2006). While their impact on evaluation of interventions has been positive, the potential of PROMs for improving general practice process is still being explored (Marshall, Haywood, & Fitzpatrick, 2006).

Tools to assist physicians in helping patients make personal decisions about their treatment have also been introduced by the Shared Decision Making research community. These decision aids aim to facilitate the patient-physician interaction. Firstly, to help physicians communicate medical information to the patient so that they can make informed choices, but also to help the patient communicate how their illness and various treatment decisions will impact their life. By directing this interaction, decision aids help to facilitate choices based on the best evidence and the opinions of at least the patient and physician, and can also involve the patients wider familial and social circle (Charles, Gafni, & Whelan, 1997). Shared decision making tools have been implemented for decisions about treatments relating to a number of conditions including angina, breast cancer, HIV, Lung cancer (see a comprehensive list at the Option Grid website[2]). They also exist to help patients make decisions about tests. AmnioDex is one recent example, which helps pregnant women make decisions about whether or not to undergo amniocentesis testing for Down's syndrome in their unborn children (Durand, Boivin, & Elwyn, 2012).

[2] http://www.optiongrid.org/

As experts in illness, patients' opinions and experiences help inform healthcare professionals and researchers at all levels of organization. From including patient perspectives in their individual treatment decisions to deciding how best to provide healthcare to communities; from evaluating healthcare to setting research goals that will have the greatest impact on them, patient experience is having a greater impact on how healthcare is provided. As discussed in the next section, the Web is one channel through which patients are expressing themselves, and means by which we can capture and use this information on a large scale are needed.

1.2 HEALTH 2.0

As discussed earlier in this chapter, the rise of social media and online social networks has had an impact on many aspects of life, and healthcare is no exception. The term 'health 2.0' is derived from 'Web 2.0' which was first used by Tim O'Reilly to describe the next generation of Web tools following the 'burst of the dot-com bubble' in the early 2000's (O'Reilly, 2005). In his article 'Web 2.0' he describes the properties of this *new* Web using cases studies about what companies operating on the platform would look like. The collaborative nature of Web 2.0, its dependence on collective intelligence and 'hard-to-recreate' data means that social media, as a platform on which people without technical or publishing knowledge could create content on the Web (Murthy, 2012) came to epitomise Web 2.0.

A consensus about the definition of Health 2.0 (also known as Medicine 2.0) has not yet been reached (Van De Belt, Engelen, Berben, & Schoonhoven, 2010), however working definitions usually focus on the use of Web 2.0 tools to facilitate participation and collaboration within healthcare across stakeholder groups (Hughes, Joshi, & Wareham, 2008). Health 2.0 describes the impact that social media and the content creation platforms have had on communication between patients, carers, physicians, other healthcare professionals, researchers and other interested parties (Eysenbach, 2008). This section aims to set out what the impact of Health 2.0 is and what opportunities it presents to connecting patients and physicians.

1.2.1 HEALTH COMMUNITIES ON THE WEB

As highlighted by NM Incite's survey of the health social media landscape (Figure 2), hundreds of thousands of patients are using social media platforms to discuss their conditions and their treatment.

While general purpose social networks can be utilised by people for discussing health, there are also specialised, health-focused social networks which allow patients to share in a more secure space. PatientsLikeMe[3] is one such example. This social network aims to help connect patients with similar conditions or undergoing similar treatments in order to inform and support each other. Conditions, symptoms and treatments are captured as structured information, meaning that patients in similar positions can be matched. At the time of writing (November, 2015) the PatientsLikeMe website reports more than 400,000 members (PatientsLikeMe, 2015).

Sharing information online is supporting patients who are living with long-term and chronic conditions. Analyses of online health conversations show that patients share experiences, support, information about their condition as well as reviews of physicians, institutions and treatments (Barker, 2008; Hoch & Ferguson, 2005; Lowney & O'Brien, 2011). A survey of over 1000 users of PatientsLikeMe showed roughly 70% of respondents found the site helpful for seeking information and various groups reported community-specific benefits (for example reduction of 'risky behaviour' in HIV patients) (Wicks et al., 2010). Patient highlighted benefits of online health communities include improved accessibility of information and practical advice for day-to-day management of their condition (Hoch & Ferguson, 2005; Wicks et al., 2010). Similarly, increased psychological well-being and empowerment has been found in online health community members (Barak, Boniel-Nissim, & Suler, 2008; Lieberman, 2007) as well as a reduction in fear (Rozmovits & Ziebland, 2004).

Experiences shared by patients through online media are an important resource for other people faced with similar issues. Recent surveys carried out by Pew Internet in the USA showed that 34% of internet users (25% of adults) had consulted other peoples' commentary or experience of healthcare shared online when faced with an information need (Fox, 2011) and that 26% had used this source of information in the previous 12 months (Fox & Duggan, 2013).

Online patient discussions have also been used by health researchers to help explore patient experiences. Sillence and Mo (2012) carried out a qualitative analysis of forum posts to understand how patients make decisions. Similarly, Hewitt-Taylor and Bond (2012) analysed online discussion boards to unravel patient expectations of their relationship with their physicians. Blogs (Keim-Malpass & Steeves, 2012) and video blogs (Chou, Hunt,

[3] http://www.patientslikeme.com

Folkers, & Augustson, 2011) of cancer patients have been analysed manually to better understand patients' experience of care. Understanding these information sources is essential in allowing them to be used more effectively by healthcare professionals and researchers as well as other patients.

Patient experiences shared online are an important source of data, for many stakeholders within healthcare. While many are sharing information in structured ways (through specialist networks such as PatientsLikeMe), other social media present this information in unstructured plain text which poses many challenges to interpretation and usage on a large scale.

1.2.2 Health experience Narrative on the Web – an Opportunity

As discussed throughout this chapter, sharing health information and experiences online may benefit the author, their readers and researchers who study their individual posts. However, aggregating the experiences of many people across this online community of patients has the potential to help unlock new information, improve healthcare and help set research directions that will have the biggest impact on patients' lives. This online discourse including self-reported experiences of a great many people could offer an unprecedented opportunity to learn about what patients experience on a grand scale. Studies (such as those described in section 1.2.1) are using patient's online communication channels to understand patient's experiences better, but manual analysis of these sources is limiting their scope to a few authors, when many thousands are sharing (see Figure 2).

Carrying out studies on large datasets or 'big data' in the modern age often relies on some form of automated processing. Areas such as data mining and machine learning focus on utilising computing power to find patterns in large datasets; the ultimate aim, to discover new knowledge. Automated processing of online health discussions could help us discover more about how patients live with their conditions on a grand scale, incorporating the experiences, thoughts and opinions of many thousands of people from all over the world.

However, whereas databases add structure to data, meaning that they can be interpreted automatically with less ambiguity, information shared through social media is unstructured (i.e. text, video, images, etc.) meaning that some human interpretation is needed.

1.3 Research Hypothesis and Objectives

It is our hypothesis that:

a) text mining tools can automatically extract and classify information from patient-generated narratives, and
b) text mining and qualitative analysis can be effectively combined to inform healthcare researchers about patients' opinions

Qualitative analysis involves systematic, manual interpretation of data and the combination of various points of view to draw objective conclusions from highly subjective opinions. This can be a time-consuming, expensive endeavor. Text mining reduces the manual effort required to analyse large text collections by automatically interpreting parts of text data and finding patterns within text collections to draw new conclusions or generate new hypotheses for further investigation. Our work aims to develop and evaluate a framework through which the qualitative analysis and text mining can be used in tandem, to help create tailored tools towards a specific dataset and enquiry, and to scale up that analysis in order to draw wider ranging conclusions about health communities. Combining the scalability of Text Mining approaches with the rigour and level of detail of qualitative analysis means that the corresponding tools have the potential to maximally utilise the patient experiences narrated on the web.

To explore this hypothesis, we defined the following research objectives:

1. To define a case study in a medical domain that aims to investigate patient perspective
2. To collect a relevant dataset of patient-authored narratives shared on the web
3. To conduct a qualitative analysis on such data
4. To develop methods to mine such data and evaluate their performance
5. To combine the findings from objectives (3) and (4) to provide a deeper insight into patient opinions expressed in the dataset (2).

Our framework (detailed in Chapter 3) uses an initial small-scale qualitative analysis as a template for designing the corresponding text mining methods, emulating some aspects of the manual interpretation. It is in this manner that we aim to scale up the analysis over large data collections.

1.4 RESEARCH CHALLENGES

In meeting our research objectives (see Section 1.3), the main challenges faced are related to the information overload problem we aim to address. To create a methodology whereby we can analyse potentially large datasets, we must create such a dataset upon which we can build and test our approach. The informal, subjective nature of the text can also make automatically extracting the required sentiment and semantic information very difficult . In this section, we explore these challenges.

Firstly, a representative sample of relevant documents will be required. Within a heterogeneous document collection, such as the Web, exists a relatively tiny set of documents which might be relevant to our needs. Finding this set of documents is a non-trivial task. In our approach, this Information Retrieval problem (see Section 2.2.1) will be approached in a semi-automated fashion, utilising online search engines and manual interpretation.

Text mining aims to emulate and scale up manual analysis, and as such they are built upon and compared to a manual analysis. The outcome of this manual analysis is the second set of data required for undertaking this work. This manual analysis data will take the form of labels taken from a discrete set which have been applied to text within the document by users qualified to interpret the text. These labels may be at various linguistic levels - i.e. word, phrase, sentence or indeed whole documents - and will encode the information we aim to extract automatically. Challenges include the avoidance of bias, imposed by the reader, in the labels obtained. A user may have a certain interpretation that may not generalise. The remedy for this problem is often to take in to account the opinion of more than one person, but this introduces the problem of scale - getting enough people to read and label the data - and how best to combine the information from a number of sources.

Text mining of patient-generated narratives is challenging and will require the development of novel approaches. Firstly, recognizing important references to medical terminology in text is one important facet of mining this data. While standardised terminologies and automated approaches have been developed for recognizing terms in text (see Section 2.2.2) they are often focused on text written by experts in a formal setting (e.g. scientific articles or medical notes). Patient-generated narratives are informal and neither written by or for medical professionals. The high degree of exophora, abbreviations, emoticons etc. make blogs substantially different from other text sources (Ulicny, Baclawski, & Magnus, 2007). This means existing text mining tools and approaches (designed to deal with more

formal language, i.e. scientific literature) may underperform when processing less formal types of documents (Doran, Griffith, & Henderson, 2006). In order to be applied in this kind of setting, novel methods would be required in order to make term recognition more robust to the challenges posed by the nature of the corpus.

Sentiment expressed by patients could provide valuable information for understanding the personal experiences of patients. It is this type of information that can inform healthcare researchers, but the interpretation of what is being conveyed by the author can be very subjective. Approaches (such as inter-annotator agreement) are used to consolidate various subjective interpretations of data (i.e. their agreements or disagreements) and settle upon 'objective' conclusions. That human experts find interpretation of these subjective accounts so difficult means that developing automated approaches will be especially challenging.

In summary, the challenges posed by our research include:

- Creating a set of relevant documents - a subset of the documents available on the Web
- Creating a set of annotations to support training and evaluation of machine learning approaches, which should:
    - annotate information we aim to extract automatically
    - reduce bias inherent in manually curated annotations, and;
    - be realistically annotated using available resources (i.e. annotators)
- Developing automatic term recognition approaches that are robust in finding references to domain-specific concepts in informal, lay-authored text
- Automatically extract opinions, or more specifically, classify sentiment and subjectivity status of individual sentences within patient-generated narratives

In the next section, we summarise the contributions of our work and how these challenges were met.

1.5 Research Contributions

As set out in this chapter, while physicians and healthcare professionals can be thought of as experts in disease it is patients who are experts in illness – in living with disease. Due to the rise in social media and the ease with which users can create and disseminate content relating to their experiences, many patients have taken to sharing information about their

illness with large audiences online. This source of information has the potential to inform healthcare professionals about illness and patient experiences, but scalable solutions to its interpretation are required to utilise it fully. As text mining has helped unlock information related to understanding disease (see Section 2.2.4), we aim to apply similar methods to help utilise data relating to patients' experiences. Our objective is to design and implement and evaluate a framework which utilises Text Mining methods to support qualitative analyses of patient-authored blogs relating to experience.

Our primary contribution is the QuTiP framework (see Section 3), designed to scale qualitative analyses of social media data through utilisation of Text Mining. In implementing this framework we addressed a number of challenges (see Section 1.4), and remaining contributions can be divided in to data collection and annotation methods, automated text mining methods and datasets.

In order to collect a dataset of patient blog posts and create the standoff annotation data, we had to overcome a number of challenges (as described in Section 1.4). We developed manual approaches to find relevant patient blogs and tools to gather the required content. In order to create some of the required annotation data, we used a crowd-sourced internet exercise, getting a large group of distributed annotators to help label sentences (see Chapter 7). Lastly, in Chapter 6 we describe how we used the coding framework developed as part of the qualitative analysis to augment standardised vocabulary in building a dictionary of terms relating to a particular medical area.

Automatic approaches to semantic information extraction and sentiment classification on patient blog posts were developed as part of our research. FlexiTerm (see Chapter 6) is an automatic term recognition approach tailored towards high performance on informal text collections, such as blog posts. Using term matching techniques which account for linguistic variation enables a more representative set of important terms to be automatically extracted from text. This approach was developed and evaluated using a number of text data sources including the blog post data created in Chapter 4 and the expert medical term annotations gathered in Chapter 6.

Lastly, the data produced in various ways during our research has the potential to be re-used in subsequent analyses. Firstly, the collection of 368 blog posts written by patients or carers living with chronic lung disease. These blog posts make up the unedited, self-disclosed accounts of 44 individuals relating to living with a long-term, debilitating disease.

The information contained in them has the potential to inform about patient experience and their needs or concerns.

We also developed resources which can be used to analyse other patient blog post data. Firstly, the coding hierarchy and themes highlighted (see Chapter 5) can be used in subsequent analyses of chronic lung disease patients' experiences of exacerbation. Also, we developed a manually curated vocabulary of over 3000 medical terms along with linguistic rules to extract information related to chronic lung disease exacerbations (see Chapter 6). These resources were developed utilising both standardised vocabularies and analysis of informal language used in patient blog posts. They could be applied to other patient social media content to provide faster access to relevant information.

The work presented in this thesis has contributed to the following publications:

Spasić I, Greenwood M, Preece A, Francis N, Elwyn G 2013. FlexiTerm: a flexible term recognition method. *Journal of Biomedical Semantics* 2013, 4:27

Greenwood M, Elwyn G, Francis N, Preece A, Spasić I 2013. Automatic extraction of personal experiences from patients' blogs: a case study in Chronic Obstructive Pulmonary Disease. *Proc. Of Third International Conference on Social Computing and its Applications (SCA)*, Karlsruhe, Germany (September, 2013).

In (Spasić, Greenwood, Preece, Francis, & Elwyn, 2013) we introduce and evaluate FlexiTerm - a novel method for automatically extracting domain specific terminology from documents. FlexiTerm forms part of the research contribution for extracting medical concepts from patient blog posts. In (Greenwood, Elwyn, Francis, Preece, & Spasic, 2013) we describe our method for classifying the subjectivity of sentences written by patients and carers in blog posts. This approach utilised a Naïve Bayes classifier trained on linguistic features of the text and a crowd-sourced set of training and test labels.

1.6 THESIS STRUCTURE

This thesis contains 9 chapters and the remainder is organised as follows:

Chapter 2 – Background

In this chapter, text mining is introduced and related work surveyed. The main focus of this chapter is how text mining has been utilised in the medical field and in social media as our work aims to bring these three areas together.

Chapter 3 – Method

Chapter 3 introduces our approach for supporting qualitative research of patient-authored accounts of health experience shared through social media. Our framework aims to allow medical researchers access to the most relevant information within this potentially valuable data source faster, by augmenting analysis with automatic text mining techniques.

Chapter 4 – Collection of COPD Patient Blog Posts from the Web

Here we describe our semi-automated approach to discovering relevant blogs from the Web and collecting posts in to a database. Through querying specialised blog search engines and manually reviewing the results we highlight candidate blogs from the Web. Utilising RSS feeds and linguistic pre-processing we store the results in a relational database.

Chapter 5 – Patient Experiences of COPD Exacerbation – A Qualitative Study

Our initial qualitative investigation of the collected blog posts is described in this chapter. We develop three questions relating to how patients identify and manage exacerbation events and the impact that they have on daily life. Our findings point towards the physical and psychological impact that, not only exacerbations themselves, but also the threat of those exacerbations can have during periods when their condition is stable.

Chapter 6 – Extracting Medical Concepts – Automatic Approaches

As described previously, automatically discovering mentions of medical concepts in text is an important step towards automatic interpretation of the textual data. In this chapter we describe two approaches to automatically extracting medical concept mentions from these blog posts. Firstly, using literature review, existing medical terminologies and the outcome of our qualitative analysis, we build a COPD exacerbation focused terminology – mixing the formal and informal terms that perform well in our blog post dataset. We also introduce FlexiTerm, a novel, flexible Automatic Term Recognition approach which is tuned to perform well in text which is subject to high variance (such as blog posts). FlexiTerm is evaluated against expert annotations of our dataset and performs well compared to a baseline method.

Chapter 7 – Sentiment Mining

Sentiment mining deals with building a picture of the author's *inner state* when writing – whether they were happy or sad about what they were writing, for instance. In this chapter, we explore mining this kind of information from text. Due to the subjectivity of this kind of interpretation, we use crowdsourcing to develop a set of annotations based on a wide range of opinions thereby creating a more objective view. We describe this exercise and its results in detail and how we use inter-annotator agreement to agree the final gold-standard annotations. In order to automate this interpretation of tone in text, we describe feature selection, training and testing of Naïve Bayes models to automatically applying labels to sentences in the blog posts, performing with over 90% precision.

Chapter 8 – Exploring the Data – Integration and Further Analysis

Having used a subset of our dataset to develop automated approaches to concept extraction and sentiment analysis, we apply these approaches to the remaining data. In this chapter, we describe the results along with the results of patterns discovered within the data and what they may mean for further research going forward. We also describe tools we have implemented to make use of open standards for qualitative data in order to be compatible with popular qualitative research software packages.

Chapter 9 – Conclusion

Our implementation of QuTiP shows how qualitative and text mining approaches can work in symbiosis to create scalable research plans for analysing patient experiences shared over social media. In this chapter, we describe the outcomes of our research and the contributions we have made as a result. We also describe the challenges that face future research in to health and social media and where our research fits in to helping make use of this valuable resource.

CHAPTER 2

BACKGROUND

The modern world creates and exchanges a vast quantity of information pertaining to all manner of things. Analysis of this information can help us to unlock new information and knowledge and there are many areas of research directed at how best to carry this out. Through systematic, manual interpretation and categorisation, qualitative analysis enables researchers to make sense and draw conclusions from a wide range of large sources of information – from academic literature, popular media and personal accounts. Automating the interpretation and categorisation of information allows the process to be carried out on extremely large collections of data. Data Mining enables conclusions to be drawn from very large databases of structured information, but the interpretation of unstructured text offers a separate set of challenges. Text Mining is an area of research which aims towards creating approaches that can extract individual data points from collections of textual data and methods for discovering patterns between data points and translating those in to new domain knowledge.

Both qualitative analysis and text mining have been utilised in the medical and social media spheres and in this chapter, we describe each area in more detail and relate to these particular applications. As described in Chapter 1, our work focuses on utilising qualitative analysis along with Text Mining in order to interpret the personal accounts of patients' illnesses shared through social media en-masse. This chapter sets out the state-of-the-art in both automated and manual analysis of this data source and applying each type of method to social media and medical data sources. Our approach unifies these approaches and applications in order to enable scaled analysis of a rich, valuable data source.

2.1 QUALITATIVE ANALYSIS

Qualitative content analysis is a research approach used to analyse text and to "provide knowledge and understanding of the phenomenon under study" (Downe-Wamboldt, 1992). This analysis aims to go beyond superficial analysis of keyword counts to unlock the

meaning behind text and place what the author/interviewee/participant says in to a wider context. Hsieh and Shannon (2005) define qualitative analysis as a "research method for the subjective interpretation of data through the systematic classification process of coding and identifying themes or patterns".

Text data for qualitative analysis can be gathered through transcripts of focus groups (Kitzinger, 1995), interviews (Patton, 2014) as well as through case studies (Yin, 2003) or collections of print media such as articles, books or manuals (Kondracki, Wellman, & Amundson, 2002).

Approaches to qualitative research are not uniform, and are designed based on the aim of the research and the data used (Bradley, Curry, & Devers, 2007). However as described in Bradley et al. (2007), there are some general steps that are used during interpretation and analysis of data. Classification (or *coding*) in qualitative research is undertaken by researchers interpreting and categorising units of text (units can range from individual characters to whole documents depending on the content and research method).

Codes are used to categorise the data, labelling sequences of text, audio or video in terms of a researcher's interpretation of the text. Coding schemes (i.e. the labels which are applied) are usually arranged hierarchically with very specific codes having parents describing more general categories. This allows generalisations based on the data to be made and interpreted. These codes are then reviewed and separated or combined in order to create key themes. Abstracting the data in this way and drawing key themes from the content is what enables qualitative research to draw new knowledge from data sources.

There are many qualitative methods which can be applied in order to analyse data. Qualitative methods may be *more* inductive or deductive (Elo & Kyngäs, 2008) in nature, depending on the aims of the research and the design of the analysis. Deductive methods approach data analysis with a pre-defined theory (i.e. one developed during a previous phase of research) in order to confirm or refute it. Inductive analysis, however, is more concerned with developing new theories which emerge from interpretation of the data.

2.1.1 QUALITATIVE RESEARCH AND MEDICINE

Due to the subjectivity and scale of qualitative research, its rigor and applicability in the medical domain has been the subject of much discussion in the literature (Green & Britten, 1998; Krumholz, Bradley, & Curry, 2013; Poses & Isen, 1998). However, as journals are publishing guidelines and methodological papers for qualitative research, the number of

articles appearing in the medical literature is increasing (Shuval et al., 2011) and applied to certain research questions, qualitative research has been accepted as a useful tool.

While quantitative research may inform clinical researchers of the efficacy of a particular treatment, it tells them little of the everyday experience of the patient. Qualitative research offers medical researchers the opportunity to systematically address questions which are harder to answer through quantitative means - the subjective beliefs, interpretations and practices of patient and physician alike. In their discussion about qualitative research being an important part of evidence-based medicine, Green & Britten (1998) give the example of qualitative studies which found that patients want to appear 'healthy' at work and so do not want to be seen taking medicine (S. Adams, Pill, & Jones, 1997). So while randomised controlled trials may prove the treatment works, this understanding of patient experience sheds light on barriers that might prevent adherence to a treatment plan.

2.1.2 Qualitative research and the online health community

Medical researchers have used online health communities as a source of qualitative data, for providing insight in to patients' experiences of healthcare. A wide range of social media are used by patients to communicate their experiences including forum posts (Hewitt-Taylor & Bond, 2012; Sillence & Mo, 2012), blogs (Keim-Malpass & Steeves, 2012) and video blogs (Chou et al., 2011) in order to inform about various aspects of patients' lives. Patients use these media to communicate their individual and subjective thoughts, feelings and experiences. However, in order to discover more general themes about patient experience, they require intense, systematic analysis and interpretation within the medical context. Through the steps detailed above, qualitative research has provided a framework upon which this analysis can take place.

Sillence and Mo (2012) carried out a qualitative analysis of forum posts to understand how patients make decisions. Analysing 137 messages from four discussion boards about prostate cancer, they investigated the different decision processes men reported. They grouped decisions in to non-systematic (e.g. deferring to medical professionals' expertise) from systematic decision processes (e.g. researching treatments and procedures themselves). In all, just under half of the messages reported non-systematic decision processes, and 36.5% reporting systematic inspection of treatment options and side-effects. This study implies that this information-seeking behaviour from patients should be reflected in the patient-physician relationship, in that people searching for information will

likely have questions for their physicians. The authors recommend that physicians must be prepared for this information-seeking and when faced with questions raised through personal research to relate them to the patient's specific circumstances.

In a related study, Hewitt-Taylor and Bond (2012) analysed diabetes discussion boards to unravel patient expectations of their relationship with their physicians. This study highlighted themes around ownership – that patients generally felt empowered to manage their own day-to-day management of their condition. They found that patients had a respect for their doctors and their opinion and that they accepted the doctor's role in managing their condition. However, many showed willingness to use other sources of advice and information when they felt this communication was not going smoothly.

While psychological processes, such as decision making, and the patient-physician relationship can have long-term implications for how treatment is managed, individual, self-reported patient narratives taken en-masse can directly affect how care is given. Blogs (Keim-Malpass & Steeves, 2012) and video blogs (Chou et al., 2011) of cancer patients have been analysed manually to better understand patients' experience of care. These analyses of patient narratives through various phases of diagnosis, treatment and management of cancer form the basis of improving care during these times for patients. Building a picture of patients' self-reported experiences of these most difficult times can help those who support them and help them through the processes involved.

Qualitative analyses have also been carried out to learn more about online health communities in general including what is shared, how it is shared and also why people feel the need to publish. For instance Fisher & Clayton (2012) conducted questionnaires on patients' interest in using social media in health care, finding increasing acceptance among over 100 respondents. Greene et al. (2011) analysed themes in Facebook posts to 15 of the largest diabetes groups on Facebook – including 690 comments from 480 users. They found that primarily, comments were conveying information to other users (65.7%), including condition management strategies. Content analysis of over 200 blogs looked at the features of blogs (Herring, Scheidt, Bonus, & Wright, 2004). This study concludes that the blogs analysed represented an important channel of self-expression, rather than a means by which they could join in a discussion of 'external events'.

As social media is used to both share and find information or experiences of health care, it is important to understand how it is used. It could also be a source of data through which

medical researchers can better understand patients' experiences of care, to understand how patients live with long-term illness outside of the direct care setting. Research on both of these fronts has been undertaken utilising qualitative approaches. The systematic, rigorous and manual nature of qualitative investigation of these sources adds credence to the conclusions drawn. Expert interpretations, verified through standardised process means that the information garnered can be regarded as authoritative. However, as highlighted by the number of documents involved in the studies listed above, the scale of the manual approach is limited within this potentially large sphere of health social media. Hundreds of thousands of patients participate in this online discourse, and analysing this growing data source would require a large amount of research effort.

2.2 TEXT MINING

As with qualitative analysis, the aim of Text Mining systems is to find new knowledge through examining sources of information. Through automation, text mining approaches offer high throughput systems which can analyse much larger document collections than would be feasible to inspect manually. The advantages of this capability are two-fold: it is more likely to (1) detect "below the radar" findings, and (2) lead to more statistically significant results. However, text interpretation requires both linguistic and domain knowledge, which make its automation a challenging task.

Text mining is analogous to data mining in its approach to knowledge extraction through identifying and analysing patterns (Feldman & Sanger, 2007). Where data mining deals with data held within structured databases, text mining approaches aim to find patterns in unstructured textual data. Unlike structured information, textual data and natural language require complex analysis in order to understand its content. The structure of a database and relationships inherent in their design implies some context on the data points represented within it. As mentioned previously, in textual data this context is communicated to the reader through the language used and the assumption of some background knowledge used for interpretation. This context is difficult to ascertain in an automated process. This is the problem that Text Mining methods address – automatically discovering and extracting information and facts from text and discovering the patterns that might constitute new knowledge. Figure 3 provides a general overview of a text mining system.

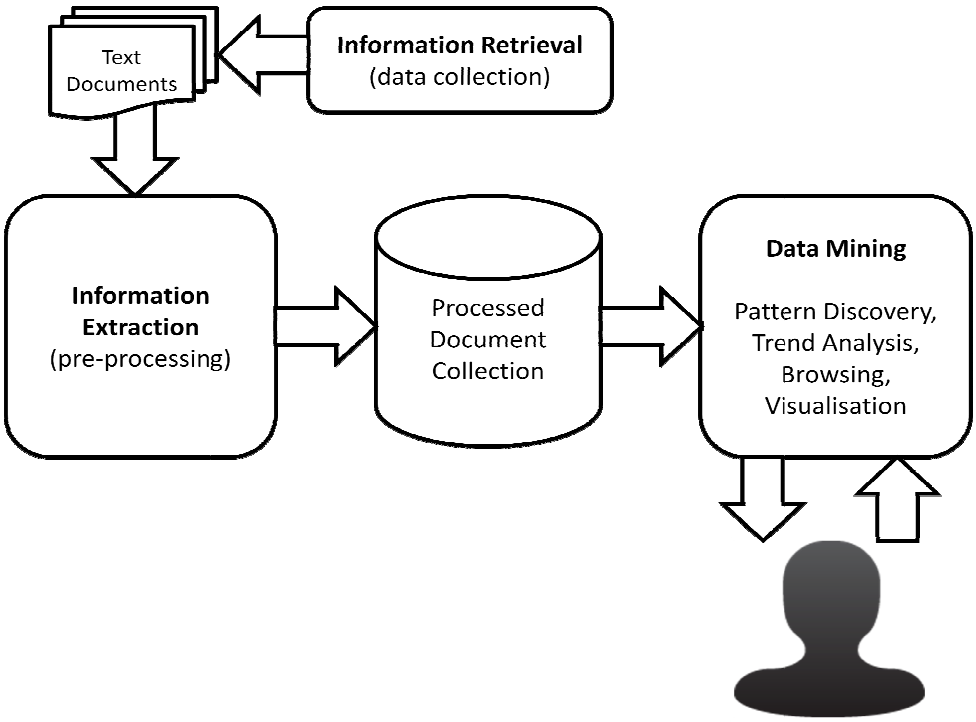


Figure 3: General overview of a text mining system
- adapted from Feldman & Sanger 2007 & Meystre et al. 2008

Three general processes often comprise a text mining system. Firstly, Information Retrieval describes the process of identifying documents relevant to a user's information need from a set of both relevant and irrelevant documents (Baeza-Yates & Ribeiro-Neto, 1999). Information Extraction relates to extracting pre-defined types of information from documents (Meystre et al., 2008). Whereas Information Retrieval aims to find relevant documents, Information Extraction deals with identifying specific information within them. Lastly, Data Mining focuses on discovering patterns in data (i.e. correlations between data points) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). It is these patterns that inform researchers of potential new hypotheses, generated by analysing large databases en mass. In a text mining system, such as the one depicted in Figure 3, the data that are analysed are the pieces of information extracted from the unstructured text.

In this section, we describe these processes and the technologies used to implement them in more detail. We also describe the impact that Text Mining has had on two particular areas of application. Firstly, how these technologies have been used to help categories and interpret medical information- e.g. medical records and literature. Secondly, we introduce text mining as a powerful mechanism in social media and interpreting peoples' experiences. We conclude by describing how these approaches, used together, have the

potential to help researchers unlock the semantic (i.e. medical domain knowledge) and sentimental (the experiences) of patients illness accounts shared through social media.

2.2.1 Information Retrieval

Information Retrieval (IR) is concerned with the representation, storage and access to information within a document collection (Baeza-Yates & Ribeiro-Neto, 1999); predominantly collections of information in the form of natural language (Salton & Mcgill, 1986). IR systems involve finding documents which contain information relevant to some information need from a collection of both relevant and irrelevant documents. Web search is one popular example of an IR application. Provided with a representation of the user's information requirement (usually search terms combined in a query), the search engine retrieves a list of potentially relevant documents from a large collection and attempts to rank them in order of relevance. In text mining systems, this process is usually similar – narrowing down the documents within a collection to those that are relevant and that *should* be processed further.

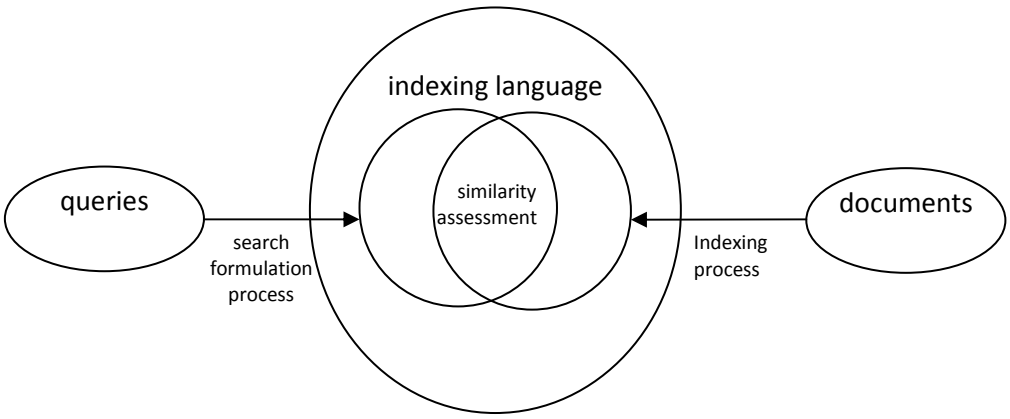


Figure 4: Functional overview of information retrieval systems (Salton & Mcgill, 1986)

Figure 4 shows an overview of a general IR system. The aim of the system is to translate both query and document in to a common indexing language allowing similarity assessment to take place. Earlier, the example of web search was given, where documents on the Web are represented in terms of their keywords, but Information Retrieval systems do not necessarily need to be completely automated. A classic card index in a library is one such manual example – books within a library are given a category or set of keywords manually and the user (i.e. a library customer) must use a system of categorised cards in order to find relevant material. This creation of a common index (the categories) and

translation of both documents (the content of the book) and the user's information need in to this language in order to satisfy their requirements is an example of an information retrieval system.

Automated indexing of documents stored digitally comprises finding suitable representations of the documents' content – as with the manual card system described above. One important method in information retrieval is the extraction of keywords or phrases to represent the document – an inverted index, where a document collection is represented as a matrix of words or phrases and in which documents they occur. An example from Berry & Browne (2005), Table 2 and Figure 5, show the representation of one collection of books as a group of words taken from their titles.

Documents	**Terms**
D1: Infant & Toddler First Aid D2: Babies & Children's Room (For your Home) D3: Child Safety at Home D4: Your Baby's Health and Safety: From Infant to Toddler D5: Baby Proofing Basics D6: Your Guide to Easy Rust Proofing D7: Beanie Babies Collector's Guide	T1: Bab(y,ies,y's) T2: Child(ren's) T3: Guide T4: Health T5: Home T6: Infant T7: Proofing T8: Safety T9: Toddler

Table 2: Representing documents as keywords

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Figure 5: Constructed document-keyword matrix

In Table 2, the titles are listed and each document given an identifier (D1 – D7). The highlighted words are extracted as the key 'terms' (T1-T9) listed on the right. It is these terms that are used to summarise each document. In Figure 5 the documents and terms are represented in an *m*x*n* matrix where the *m* rows represent the keywords and each document is represented by one of the *n* columns. In this matrix, a_{ij} is the frequency of keyword *i* in document *j* – assuming that the more important terms in a document occur more frequently. More sophisticated measures of the relative importance of a term to a document have also been developed, such as term-frequency inverse document frequency (*tf-idf*) (Salton & Mcgill, 1986) where the frequency of a keyword in a document is penalized if it is commonly found in other documents. The *tf-idf* measure strengthens the discriminative power of IR methods, assuming that these are more important and promoting them in relevance judgments. Binary representations of presence or absence of a keyword in a document have also been used to index documents (Robertson & Jones, 1976).

The translation of our information need in to a keyword vector format (as found in many popular Web search engines) means that we can match documents to our query based on their contents. Cosine similarity, where the angle between the two vectors represents their similarity is one popular method (Salton, 1971). As well as geometric measures of document-query similarity, probabilistic measures have also been developed, where the probability of relevance to the query is estimated based on the content of query and documents.

Whether manual or automated, the IR system depends on some summary representation of the content of the document and a user's information need. There are many approaches to extracting, representing and searching within these document collections. In this section, we described the general aim of IR systems and introduced some methods which have been developed to achieve them. Given the masses of information produced within the medical domain – for example medical records and literature – information retrieval has been an important area within medical informatics. Later in the chapter, we describe specific applications of these sorts of approaches within the medical domain.

2.2.2 Information Extraction

Information extraction (IE) systems operate at a finer level of granularity than IR systems. Where IR aims to identify potentially relevant documents and present them to the user, IE focuses on pinpointing the precise information and presenting it in a structured form

(Feldman & Sanger, 2007). IE approaches attempt to represent documents in terms of the *entities* present and their relationships as described in each document (Feldman & Sanger, 2007). These entities are instances of semantic classes that are usually defined a priori. For instance, an information extraction system might be designed to find companies and people who work there. How the system will identify the two semantic classes (*companies* and *people*) and the relationship between them (*employment*) within the text will be defined in terms of rules or a model trained on previous examples.

While text communicates information, not everything required for its interpretation is conveyed explicitly. The readers' linguistic skills together with their background knowledge and experience play a significant part in this process. The pre-processing steps in text mining systems augment the textual data with layers of information (i.e. lexical, syntactic and semantic information) in order to facilitate automatic processing of the text (Spasic, Ananiadou, McNaught, & Kumar, 2005). Figure 6 shows an example of utilising this layered information approach to extracting facts about a patient from a report. Using linguistic information and applying domain knowledge means that specific facts can be identified within text. For example, during the process depicted in Figure 6 medical terms (references to medical concepts, e.g. the symptoms mentioned) have been found and labeled.

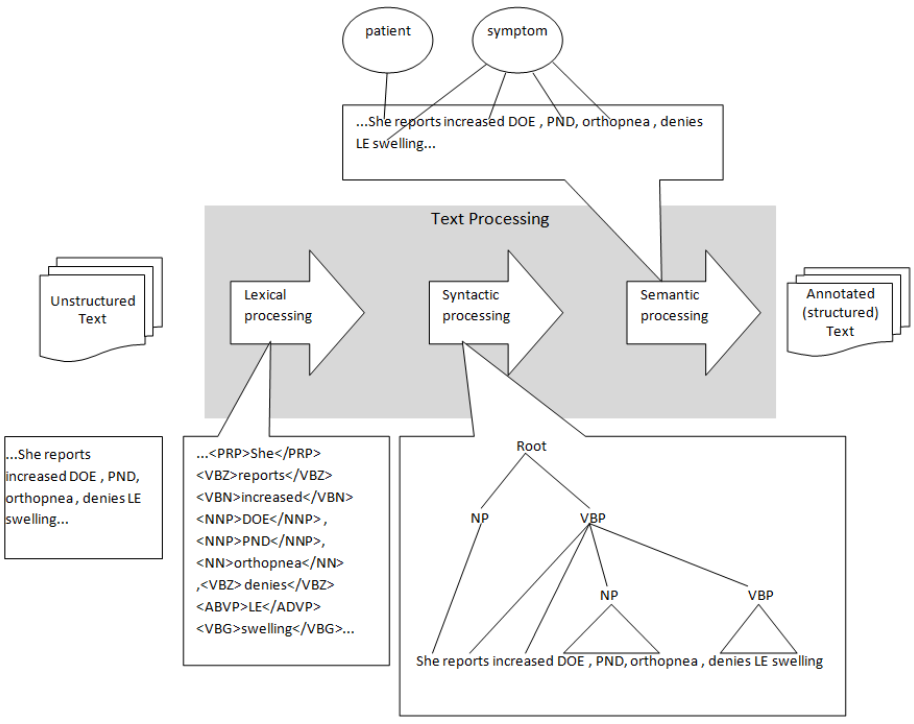


Figure 6: Text pre-processing overview- adapted from (Spasic et al., 2005)

Semantic analysis of text usually requires some background knowledge (Feldman & Sanger, 2007) and therefore text mining systems based on a concept-level representation are usually domain dependent (Tan, 1999). For instance, the process depicted in Figure 6 may be useful within the medical domain (i.e. recognising symptoms in text) but not necessarily in others.

The semantic information added does not necessarily need to be formal domain knowledge. For instance, opinion and sentiment analysis depends on the emotions conveyed by the author (Pang & Lee, 2008). These applications of text mining seek to discover concepts mentioned in text and also sentiment or opinion expressed about them by the author (e.g. 'positive' or 'negative' feelings towards a product) (Lakshmanan & Oberhofer, 2010). In these applications, knowledge sources concerned with emotive language can be used, e.g. SentiWordNet (Esuli & Sebastiani, 2006), WordNet-Affect (Valitutti & Stock, 2004) and EmoLex (Mohammad & Turney, 2010).

Each phase can be achieved using a number of approaches. Linguistic patterns, for instance, can be used to identify concept mentions in text and to extract pre-defined entity mentions (Carol Friedman, Shagina, Lussier, & Hripcsak, 2004). Machine learning can also be used to train models to identify entity mentions automatically from a set of labeled training data. Hidden Markov Models (HMM) (Rabiner & Juang, 1986) are one approach used for Information Extraction tasks (Feldman & Sanger, 2007). A finite state automaton, the HMM models a probabilistic generative process. Starting in some initial state, a symbol is chosen and then a transition to a new state is chosen whereby the symbol generation-transition cycle is performed again. The process continues until some final state is reached. Modeling these situations based on example data allows the outcome to be predicted based on the probabilities of state transition and of symbol generation meaning that the outcome can be predicted for new data. In text classification, this can be useful considering sequences of words or characters as individual states dependent on their context. Dependency between words or characters of a sentence is driven by the rules of language and grammar and as such, the HMM approach is useful for labeling this sequential data.

2.2.3 DATA MINING

Once textual data has been pre-processed and background information and structure has been explicitly encoded, it can be mined for trends and patterns in the last step – data mining. It is these patterns and trends which can potentially yield new information – for

instance, relationships between concepts which have not previously been considered. One practical example of this process from start to finish is work by Swanson & Smalheiser (1997) regarding migraines. Text mining was used on a corpus of biomedical research literature to automatically extract information about stress and migraines. Individual facts extracted from the text were combined to generate a new, testable hypothesis regarding a causal link between magnesium deficiency and some migraines.

Data mining approaches tend to satisfy two main goals – prediction and description (Fayyad et al., 1996). Prediction involves using some variable or model to predict future values based on new data whereas description involves finding some human-interpretable trend in data that could tell us about some real world phenomena. Each goal can be achieved using a number of classes of approach. Firstly, classification where the aim is to apply pre-defined classes to data (Hand, 1981) based on descriptions of previous examples (features) and outcomes which have already been observed (i.e. the label that was applied). Regression is a process through which a function is developed which maps data items to some predicted value which, unlike classification, may be continuous rather than discrete. During clustering, data items are grouped based on similarity in to a finite set of categories. Rather than being pre-defined as with classification, in clustering the groups are identified through relationships in the data and it is users' who describe those relationships. Summarization involves finding summary descriptions for data. A simple example given by Fayyad et al. (1996) is finding the mean and standard deviation in a population of numerical values – summarizing the dataset in to a reduced number of values. Dependency modeling is a process through which significant dependencies between variables in a dataset can be identified. Lastly, change and deviation detection approaches can be used to discover the most changeable aspects of data over time.

In Text Mining, the data mining processes are applied to document collections. Classification, for instance, can be automatically applying some category to a document based on its content. Through clustering, similar documents can be grouped together allowing relationships to be described. Dependency modeling can be used to find relationships between entities identified during Information Extraction (see Section 2.2.4.2). Data mining helps build upon previous processes in Text Mining in order to explore relationships through textual data and describe real world phenomena.

2.2.4 Text Mining and Medicine

Text mining has been applied in many disciplines within the biomedical domain. In his review of the impact of informatics (including text mining) on transitional medicine (i.e. the process of translating biological discovery to clinical adoption and community medicine), Indra Sarkar sets out the Translational Medicine Continuum (see Figure 7) and discusses the synergistic relationship between major areas in translational medicine (innovation, validation and adoption of new treatments), and the main areas of biomedical informatics (bioinformatics, imaging informatics, clinical informatics and public health informatics) and the concepts they deal with - from molecules to populations (Sarkar, 2010).

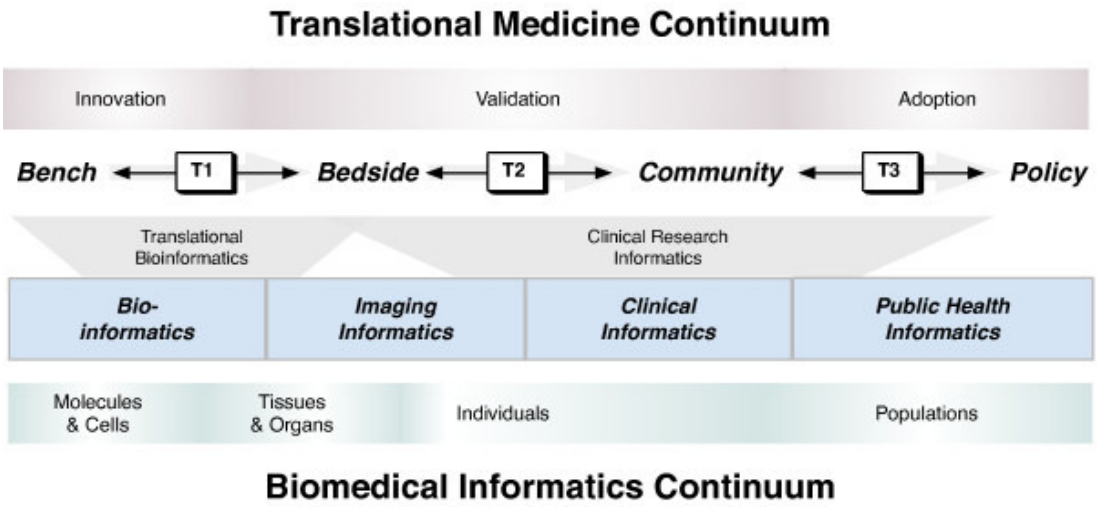


Figure 7: Translational medicine continuum (Sarkar, 2010)

In the remainder of this section, we will review specific examples of text mining applications within the biomedical domain, fitting them in to this continuum focusing on clinical and public health aspects of informatics.

2.2.4.1 Information Retrieval

Manual categorizations (albeit for documents often stored and disseminated through digital means) are used within both medical records and literature in order to satisfy information needs. Notes in a modern electronic health record are usually augmented by standardised codes attributed by physicians. For instance, the Read codes in the UK (O'Neil, Payne, & Read, 1995) and the International Classification of Disease (ICD) codes (World Health Organization, 2010) (currently version 10, with 11 due in 2017) provide taxonomies of health information including diagnoses, treatments and demographic information. This structure allows for the aggregation of medical information at the population level and easy retrieval of medical information from them. In these examples,

the indexing language is the set of codes and this is used to represent both *document* and *query* through manual translation of information by medical professionals.

Similarly, manually curated vocabularies have been used to organise medical literature. The Medical Subject Headings (MeSH) vocabulary[4], maintained by the United States' National Library of Medicine (NLM) is a controlled vocabulary of medical terms organised hierarchically (Rogers, 1963). This manually created indexing language is used by (among other things) medical literature search engine PubMed[5], where articles are labeled using the MeSH headings by professionals, and facilities exist to build queries using the same vocabulary along with binary operators (e.g. AND, OR, NOT, etc.).

While manual effort is still relied upon to both apply these vocabularies and search within document collections, work to automatically apply these indexing languages is well developed. A review by Stanfill et al. (2010) reviewed over 70 systems implemented which automated applying standardised diagnostic codes (e.g. ICD) to clinical notes. Approaches surveyed included 21 which were to support the retrieval of cases and 35 to support decision support and patient care.

2.2.4.2 Information Extraction

Automated IE has been applied to extract diagnoses, medical problems, treatments and demographic information from medical notes, discharge summaries and radiography reports. Good overviews of the field of medical information extraction are available in reviews carried out by Meystre et al. (2008) and Spyns (1996), reviewing approaches based on the kind of technology used, the information extracted and the documents it was extracted from. MedLEE (C. Friedman, Alderson, Austin, Cimino, & Johnson, 1994) is a good example of a medical information extraction system. Built to extract diagnosis and status information from radiological reports, its functionality has since been extended and has been applied to multiple problems on multiple document types (C Friedman, 2000; Carol Friedman et al., 2004). MedLEE is based on an external lexicon and grammar. The document is parsed and patterns based on cue-words and expressions are used to make inferences about surrounding phrases. Compounding these constituent semantic elements, facts and further information are identified and structured. Machine learning has also been used to train classifiers for semantic analysis of clinical documents including

[4] http://www.nlm.nih.gov/mesh/meshhome.html
[5] http://www.ncbi.nlm.nih.gov/pubmed/

maximum entropy (Bashyam & Taira, 2005) and Support Vector Machines (SVM) (Sibanda, He, Szolovits, & Uzuner, 2006).

2.2.4.3 Data Mining

More recently, health surveillance is one area in which text mining has been successfully employed to provide useful information to health professionals. The Real-time Outbreak and Disease Surveillance (RODS) system utilised text mining to identify outbreaks of diseases in a population (Tsui et al., 2003). It was developed to monitor outbreaks during the 2002 Winter Olympics in Utah, US in an attempt to identify possible biological attack on a public event, but in doing so highlighted the spread of more natural communicable diseases such as influenza (Gesteland et al., 2002). Natural language processing was used on the free-text hospital discharge summaries to identify the chief complaint of the patient from the notes and label it as belonging to one of seven syndromic categories (Chapman, Dowling, & Wagner, 2004). Utilising data mining techniques and a geographical knowledge-base, trends in the syndromic information were monitored in real-time, generating alerts if an outbreak was detected. Since the 2002 Winter Olympics, similar systems have been employed by the 2004 Olympic Games (Dafni et al., 2004) and other large public gatherings (Lombardo et al., 2008). This system is a useful example as it used each component discussed in this section in order to address a real-world information challenge – extracting knowledge from unstructured data with the added constraint of needing to work in real-time.

2.2.5 Text Mining and Social Media

The rise in user-generated content shared on the Web means that it is now an important information source for discovering more about opinions and sentiment expressed by people. Product reviews are just one example of opinion sharing on the Web. Many platforms enable users to share their opinion about products and services with other users (i.e. Amazon, Google Places). Individual posts are often made up of plain text comments and some description of structured overall rating (i.e. the '5-star' rating used by Amazon). Sentiment analysis (also known as Opinion Mining) deals with extracting, classifying and presenting information about opinion and sentiment expressed by people. While the scope of sentiment analysis is not limited to the Web, our discussion will be focused on this particular application.

As discussed in Chapter 1 in regard to experience in healthcare, information about patient opinion is relevant to a number of stakeholders in many contexts. Firstly, users trying to

make a decision may wish to consult people who have made similar ones. One useful example where opinion is important is making purchase decisions – users' may consult the opinions of others who have already made a purchase in order to make the decision themselves. A recent survey carried out by The Nielsen Company involving 28,000 people from 56 countries (The Nielsen Company, 2012) showed that consumer opinions shared online were trusted by 70% of respondents. Surveys of over 3,000 people in the USA showed that 58% of internet users confirmed have carried out research on products online (24% did so during a typical day) (Jansen, 2010). Consulting others' opinions online has also been reported to have an influence on not only what we buy, but what we are willing to spend. Respondents in a similar study reported that reviews shared by others online had a significant influence on their purchase decisions for a number of service categories (73% for 'home' – 87% for 'Hotels') (comScore, 2007). They also indicated that they were willing to pay between 20-99% more for a service with 5-star ('excellent') rather than 4-star rating. This kind of impact on behavior means that businesses which provide products and services may well be interested in garnering more online support. Reviews also offer direct feedback to service providers in order to evaluate and improve their products.

While structured information such as product ratings and recommendations can be quickly quantified and digested, much of the detail of users' opinions is in free-text and comment forms, shared through review platforms and social media. The unstructured nature of the medium and often casual, lay format of the text poses challenges to automated processing. Here we discuss three main tasks carried out for sentiment analysis: identifying opinion, categorising subjective statements and identifying the entities involved.

A core task in sentiment analysis is classification of the sentiment expressed. Classification can be applied at a number of document levels depending on the application. Firstly, text may be categorised in terms of the polarity of opinion – whether the opinion expressed is positive or negative. For example, one often tackled problem is deciding whether movie reviews express a positive or negative opinion of a film. A corpus of reviews gathered by Pang et al. (2002) from the Internet Movie Database (IMDb) website (including over two thousand reviews with a thumbs up/thumbs down rating - i.e. liked or disliked) has been used to test various methods for classifying positive and negative sentiment. In the paper which introduced this corpus, Pang et al. explored treating sentiment classification in a similar manner to topic classification, experimenting with unigram and bigram features alongside Naive Bayes, Maximum Entropy and Support Vector Machines classification

achieving 77-83% accuracy beating their 50-69% baselines set by human annotators. Feature sets classically used in Information Retrieval have been applied to this problem, including term presence, term frequency, parts of speech and syntax information as well as specially constructed sentiment-focus lexicons and knowledge-bases (Pang & Lee, 2008). Work with each has shown that the way each feature is employed for sentiment classification tasks differs from their use for topic classification. For instance, while tf-idf is a useful measure of term frequency in document topic classification Pang et al. (2002) found that term presence was a better indicator of polarity than term frequency. Besides binary classification tasks (*good*/*bad* opinion), regression has been used to discover *how* positive or negative an opinion is (T. Wilson, Wiebe, & Hwa, 2004) and the *strength* of the opinion held (Schler, 2005). Classifying opinions has also been carried out at the statement level, as well as summarising an entire document (Schölkopf, Platt, & Hofmann, 2006).

While much of the work related to sentiment analysis has focused on online product reviews, some studies have been undertaken in the medical domain. The SenticNet[6] sentiment analysis system has been applied in a number of areas within healthcare. Recently, SenticNet was used to analyse ratings of local NHS services and hospitals on a patient review website (Cambria, Benson, Eckl, & Hussain, 2012). The polarity of opinions shared online towards treatments has also been analysed automatically (Denecke, 2008)

Subjectivity analysis deals with *inner states* as defined by (J. M. Wiebe, 1994). Textual content can describe both the objective narrative of events (i.e. *what happened*) and also the thoughts, perceptions and inner states of the author, people or characters involved; their psychological point of view. Subjectivity analysis deals with determining this subjective content from objective descriptions and often to whose point of view the text relates. Filtering the subjective from less opinionated content contributes to opinion mining in increasing the precision with which opinion is identified. Likewise, identifying the point-of-view of a specific opinion means that the stakeholders involved can be highlighted, allowing more accurate sentiment information to be extracted.

[6] http://www.sentic.net/

2.3 CONCLUSION

Systematic interpretation and extraction of knowledge from textual collections is an important endeavour in order to make sense of the masses of content produced for all kinds of reasons. In this chapter, we've described the general aims and principles of qualitative analysis of data – the rigorous process of highlighting themes in textual data through manual inspection. Text Mining is a discipline of informatics which aims to automate the process of interpreting large collections of text and extracting new knowledge from them. We also introduced how these two separate approaches have been used within medical disciplines in order to analyse medical information and their application within social media. It is the intersection of these technologies that our work is focused upon.

Analysing content created by people on the Web has the potential to inform researchers on many aspects of modern life. In medicine, understanding patients' experiences of health care is an important aim and social networks offer an insight in to life outside of the direct care setting. Though qualitative research is increasing in importance within the medical research community, the scale of the online discussion being held through social media means that it is limited in this context. However, manual interpretation is the gold standard for textual content analysis and qualitative research adds the rigor of a systematic, formal process.

Text mining has been used to analyse professionally produced medical text – both academic research and medical notes. Likewise, text mining approaches have been developed to extract experiential information from social media content (i.e. sentiment analysis). Interpreting patients' social media posts in a medical context through qualitative approaches has shed light on patients' experiences of health care. However, the analyses have thus far been carried out on relatively small sets of documents, from within an extremely large community of content creators is available.

Our hypothesis is that Text Mining methods, while not replacing qualitative research of patient experience shared through social media, can at least augment interpretation of social media content on a larger scale. Applying information extraction and sentiment analysis on posts shared by patients on blogs or in online health communities in conjunction with manual qualitative analysis has the potential to help scale up analysis, and increase the representativeness and usefulness of the conclusions drawn. In the next chapter, we introduce our framework for unifying systematic, manual qualitative analysis

with the semantic and sentiment analysis offered by Text Mining approaches in order to scale the analysis of patient experiences shared through social media.

CHAPTER 3

THE QUTIP FRAMEWORK

The network of patient interaction online is large and complex, spanning many platforms and many media. Analysis of this data has the potential to inform about patient experiences at many levels of health care provision. However, due to the amount of data involved, large scale manual analysis of this online health discussion would be expensive to carry out (see Section 1.2.1). By enabling automatic interpretation of textual data, Text Mining approaches (as described in Chapter 2) provide potential ways to scale up qualitative analyses.

In this chapter, we will introduce our Qualitative Text Processing (QuTiP) framework which aims to utilise both the rigour and precision of qualitative approaches, combined with the high-throughput, scalable qualities, of Text Mining in order to support the use of the valuable information shared by patients every day.

We developed QuTiP in conjunction with a qualitative enquiry in to the experiences of Chronic Obstructive Pulmonary Disease (COPD) patients. COPD is a chronic, often debilitating, lung condition which is one active area of patient discourse online (see Figure 2). In Section 3.2, we introduce COPD and discuss the symptoms and complications that can affect people living with the condition. Lastly, in Section 3.3, we discuss our application of QuTiP to a qualitative investigation in to COPD patient experiences.

3.1 QUALITATIVE TEXT PROCESSING FRAMEWORK (QUTIP)

The QuTiP framework describes a general approach to qualitatively analysing large volumes of data by utilising automated methods and the outcomes of smaller-scale analyses. In Figure 8, we detail our proposal for a process through which text mining can support the scalable analysis of blog post data, which can involve multiple members of an interdisciplinary team (e.g. healthcare professionals, software engineers, etc.). After identifying relevant sources of information (e.g. patient blog posts) and collecting a set of

data, a development subset is created (100 documents in our case) which forms the basis of defining the approach. The initial qualitative analysis will mean that codes and themes in the subset can be discovered and will enable an initial exploration of the data for developing Text Mining tools. The aim of these tools is to support automated information extraction and annotation on the wider set of collected data, thereby enabling scalable analysis developed in synergy with qualitative approaches. Integration with qualitative analysis software packages (such as Nvivo[7] or Atlas.ti[8]) means that annotations and information extracted as part of the text mining process can be integrated within the coding structure used in qualitative analysis.

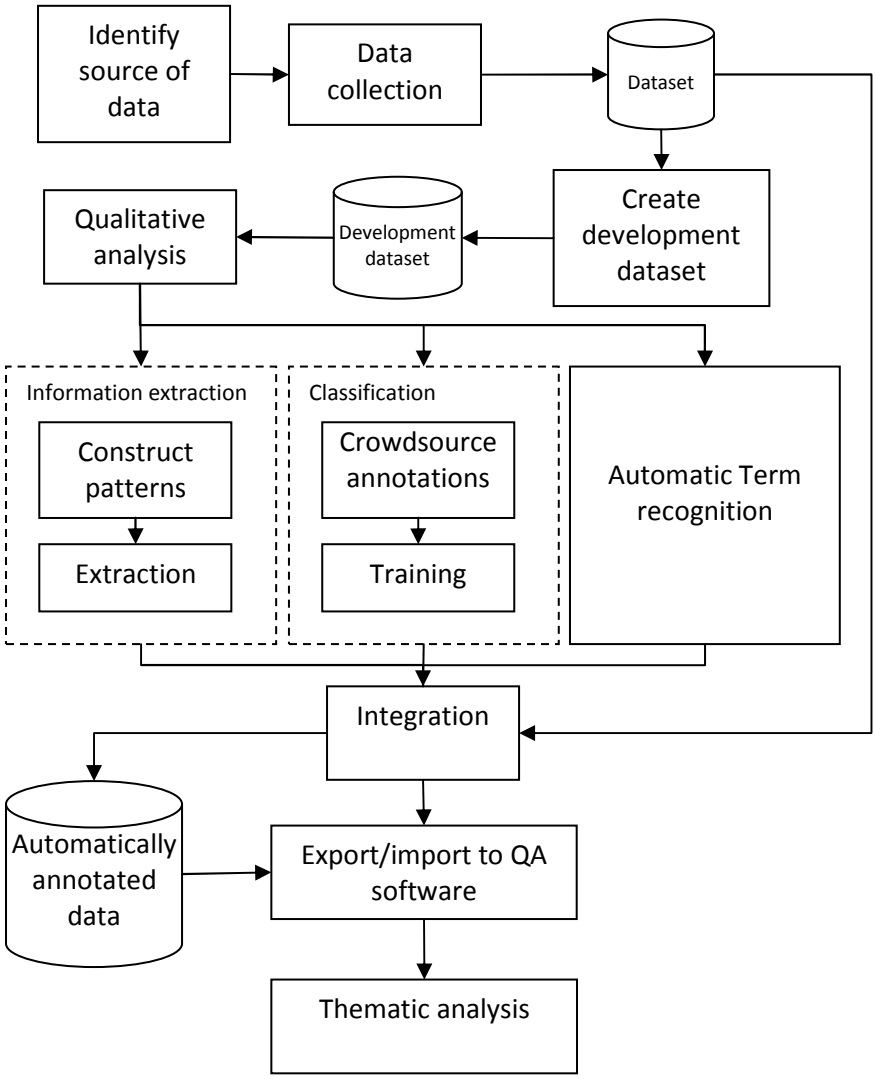


Figure 8: Text mining to support qualitative analysis of social media data: a process overview

[7] http://www.qsrinternational.com/products_nvivo.aspx
[8] http://www.atlasti.com/index.html

In this chapter, we discuss the Text Mining approaches introduced in Section 2.2 within the context of the initial qualitative work and the QuTiP framework and set out how they might assist in scaling up analysis of social media data.

3.1.1 Data Collection

The first stage is to identify and collect a relevant set of documents from targeted social networks. Existing analyses of social media sources have often identified relevant sources of data through a manual review process but an automated approach may also be appropriate. From the collection of documents, a smaller 'development set' is created which will be the target of an initial qualitative investigation alongside being used to train models or create rule sets for the subsequent automated processing. These tools will then be used to scale up the process and analyse the rest of the collected dataset.

3.1.2 Qualitative Analysis

This phase comprises a systematic expert annotation and rigorous analysis of the development dataset as introduced in Section 2.1. Again, the methods (e.g. more deductive or inductive) that are used to fulfil this phase of the process will be driven by research aims and the constraints of the data. Besides the primary outcomes of the qualitative analysis (i.e. some initial insight in to research questions), the annotated data and themes extracted form part of the input to development of the models and resources that will support mining the larger dataset.

3.1.3 Classification

Automated classification of textual data can help support the discovery of new knowledge. Utilising text classification techniques can help automatically organise documents within the collected corpus and allow for other patterns to be discovered. The classification phase is intended to provide labels to documents, paragraphs or sentences relating to general themes of interest to the researcher. Automatically labelling portions of text within the document collection can allow for rapid access to relevant information and for analysis of patterns or trends in the dataset.

While machine learning avoids the need to hand-craft classification rules, the main bottleneck here is the availability of the training data and the need for some type of manual labour still remains. To deal with such bottleneck, we proposed crowdsourcing as an efficient way of acquiring large amounts of training data, which also captures many interpretations –important aspects when dealing with automating the recognition of

essentially subjective information. By breaking down a large task (labelling thousands of sentences) in too much smaller tasks to be handed out to a large number of participants, we can collect a potentially large amount of data in a distributed fashion. The question that still remains open is who exactly would participate in large-scale data annotation. Some researchers have tapped into commercial platforms such as Mechanical Turk (Callison-Burch & Dredze, 2010; Kittur, Chi, & Suh, 2008; Rashtchian & Young, 2010; Sorokin & Forsyth, 2008), while others have involved individuals with a vested interest in the problem at hand (Pestian et al., 2012). The latter approach in particular seems like a good way to involve patients in healthcare research and improve communication between patients, researchers and healthcare professionals.

3.1.4 Information extraction

As discussed previously, information extraction aims to find mentions of specific classes of entity (e.g. symptoms, treatments or medical professionals) within textual data. As described in Section 2.2.2, Information Extraction aims to extract word, or phrase-level, mentions of those general classes of entity and relationships between those entities. Extracting mentions of specific entities and their relationships can enable researchers to discover patterns and trends between real-world concepts.

While related to classification (as described in the last section) Information Extraction is distinct in this process based on the granularity of data that is under consideration. While classification is aimed towards larger portions of text (i.e. whole documents, paragraphs or sentences). Information Extraction focuses on phrase and word-level labels. These individual mentions of certain classes of entity can often form part of a model used for classifying the surrounding portions of text, however the process involved is separated in this framework.

3.1.5 Term recognition

Term recognition is an important part of text mining, which unlike information extraction is not hypothesis-driven, but rather data-driven and, therefore, more likely to lead to "serendipitous" discoveries. Discovering the concepts that are discussed in text is a vital step in interpreting the content. The development of FlexiTerm (Spasić et al., 2013), which not only extracts terms but relates semantically similar variants together, means that the conceptual content of a document set can be better quickly characterised by the prevalent terminology used. As with information extraction, finding the concepts which are discussed in text is part of the initial coding phase of a qualitative analysis. With our work,

we related the terms discovered to the original coding hierarchy, created as part of the qualitative analysis carried out in Chapter 5.

As FlexiTerm utilises linguistic clues and statistical measure to extract domain-specific terms from text, and does not rely on external lexicons or knowledge bases, it can be applied on other document sets within other domains easily, with no retraining needed. FlexiTerm has already been evaluated and performed well on blog posts, medical notes and literature in various areas of the biomedical domain (Spasić et al., 2013).

3.1.6 INTEGRATION AND SCALABILITY

The ultimate aim of developing these approaches in parallel with a qualitative analysis is in order to replicate interpretation on a larger scale, in order to include more data in analyses. Integrating, not only the information extracted, but also the qualitative analysis is the next step in the process proposed in Figure 8.

Firstly, pattern discovery within the extracted information could yield interesting directions for further inquiry. In Chapter 8, we describe using point-wise mutual information to look at the relationship between terms and the subjectivity of their context. The aim being to highlight concepts discussed in a subjective, rather than objective way. Combining the outcomes of two previous steps, we were able to highlight useful patterns in the data and give examples from the data. While the outcomes themselves may not be significant, the patterns present may help form part of an initial inquiry in to the dataset – as part of the problem-defining process.

Integrating the outcomes of the text mining portion with the qualitative analysis is an important step. Qualitative research software (e.g. Nvivo and Atlas.ti) is well established within the field and rich with features to help support the qualitative analysis of text. In order to integrate automatically extracted information with these packages, exporting and importing information in the form of annotations requires some sort of standardised data format. Research towards standardising an open qualitative data exchange model has been undertaken by the Qualitative Data Exchange project (QuDEX)[9] and a schema for representing qualitative data (including coding hierarchies and annotations) is available. Support for this open format is built in to Atlas.ti, but the format is not currently supported by Nvivo.

[9] http://data-archive.ac.uk/create-manage/projects/qudex

3.2 MEDICAL DOMAIN – CHRONIC OBSTRUCTIVE PULMONARY DISEASE

In order to develop QuTiP we are applying the framework detailed above to address a particular medical information need. Chronic Obstructive Pulmonary Disease (COPD) is a long-term degenerative condition, and people who live with this condition are one of the more active communities online (see Table 1). In this section, we describe COPD and some of the ways it can affect a patient's life in order to motivate the information we aim to gather through our application of QuTiP. In section 1.8 we describe this approach in detail.

COPD is described by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) as a disease which causes a progressive limitation of airflow within the lungs, which is not fully reversible (GOLD, 2014). The damage characterised by COPD is usually caused by inhaled particles – predominantly tobacco smoke, but also other pollutants such as industrial chemicals. A genetically inherited deficiency of Alpha 1-antitrypsin can also lead to the condition.

COPD kills around 25,000 people in the UK each year, and accounted for 4.8% of all deaths between 2007 and 2009 (Department of Health, 2011). A report produced by The British Thoracic Society (2006) shows there are around 835,000 people living with diagnosed COPD and estimates of a further 2.2 million undiagnosed cases have been made from representative samples of the population (Shahab, Jarvis, Britton, & West, 2006). This report also estimates over 1 million hospital bed days were related to COPD. Due to the cumulative nature of the damage inflicted on the lungs, COPD predominantly affects elderly people. It has been predicted that by 2020 COPD will be the fifth leading cause of disability and the third leading cause of death worldwide (MacNee & Rennard, 2004).

Much of COPD patients' day-to-day care is self-managed (NICE, 2004) and therefore takes place outside of the direct care setting. This means that patients need to become more informed about their condition and online health communities are particularly important to this patient population. COPD is therefore a useful case study for our approach. In this section, we will introduce details of the symptoms of COPD, the complications that can occur and the current advice for treating and managing this chronic condition.

3.2.1 SYMPTOMS AND DIAGNOSIS

COPD presents with respiratory symptoms such as chronic cough, wheezing and dyspnea. A diagnosis is confirmed by physicians using spirometric assessment which measures the amount of obstruction present in the lungs. Spirometric tests include the forced expiratory

volume in one second (FEV_1) and forced vital capacity (FVC) tests which measure the volume of air that can be exhaled in the first second of breath and the greatest volume of air that can be expelled in one breath, respectively. A fixed ratio of $FEV_1/FVC<0.70$ is used to define airflow limitation and the FEV_1 can be used as an indication of the severity of the patient's condition (Vestbo et al., 2013). The GOLD COPD severity classification is set out in Table 3.

Level	**Severity**	**FEV_1**
GOLD 1	Mild	$FEV_1 \geq$ 80% predicted
GOLD 2	Moderate	50% ≤ FEV_1 < 80% predicted
GOLD 3	Severe	30% ≤ FEV_1 < 50% predicted
GOLD 4	Very severe	FEV_1 < 30% predicted

Table 3: GOLD COPD severity scale (Vestbo et al., 2013)

3.2.2 MANAGEMENT

As mentioned previously, COPD is a chronic condition which is largely managed outside of a direct care setting. The damage related to COPD is non-reversible and progressive and as such, therapeutic interventions focus on managing the condition and slowing deterioration of the lungs. Symptoms of the disease can be controlled using bronchodilators (e.g. inhalers) which increase airflow in the lungs and can be used to relieve breathlessness. For people with severe COPD, long term oxygen therapy may also be supplied.

Non-medication interventions may also be used to improve patients' condition. Smoking cessation is a recommendation for all COPD patients. Patients may also receive a referral for pulmonary rehabilitation, which is a multidisciplinary approach to improving physical and social autonomy including physical training, disease education and behavioural intervention.

3.2.3 EXACERBATION

Exacerbation of COPD is defined as an acute worsening of a patient's condition beyond the range of day-to-day variations. This may include worsening breathlessness or cough or increased sputum production. As these symptoms characterise COPD, it is the severity relative to the patient's stable state which signifies an exacerbation of the condition.

Exacerbations can be caused by many lung irritants including respiratory infection, pollutants, allergens and air quality. They could also be brought about by excessive physical activity. Exacerbations vary in severity, some can be managed by the patient themselves while others may require a visit to their GP or possibly a hospital stay. As so much of a patient's day-to-day care is self-managed, social media has the potential to inform about patients experiences of exacerbation – their understanding of exacerbation, how they monitor their condition and how they make treatment decisions when their condition is worsening.

3.3 USING QUTIP TO ANALYSE PATIENT EXPERIENCES OF COPD EXACERBATION SHARED THROUGH BLOGS

As described in Chapter 1, social media is an important source of information relating to patient experience. While qualitative approaches (see Section 2.1) are important tools for making sense of these subjective accounts and drawing more general conclusions or recommendations based upon them, the scale at which they can be applied is limited as they are based upon manual interpretation. It is our hypothesis that Text Mining methods (see Section 2.2) can help augment this manual process and ultimately scale up this potentially rich source of information.

In this chapter, we have introduced QuTiP - a framework for pairing the manual, qualitative analysis with the automated, scalable Text Mining in order to support analysis of social media data (see Section 3.1). QuTiP was developed in conjunction with a particular medical domain, but the methods used are designed to be generalisable to other media or areas of inquiry. We used QuTiP to analyse blog posts written by those living with or caring for people suffering from COPD (see Section 3.2). As COPD is largely managed by the patient, social media has the potential to inform researchers about patients' experiences living with the condition. We applied QuTiP in order to analyse social media for information relating to how patients manage their condition and potential acute worsening of symptoms through exacerbations. Figure 9 shows how the framework was applied and is colour-coded according to which chapter details the work associated with each phase.

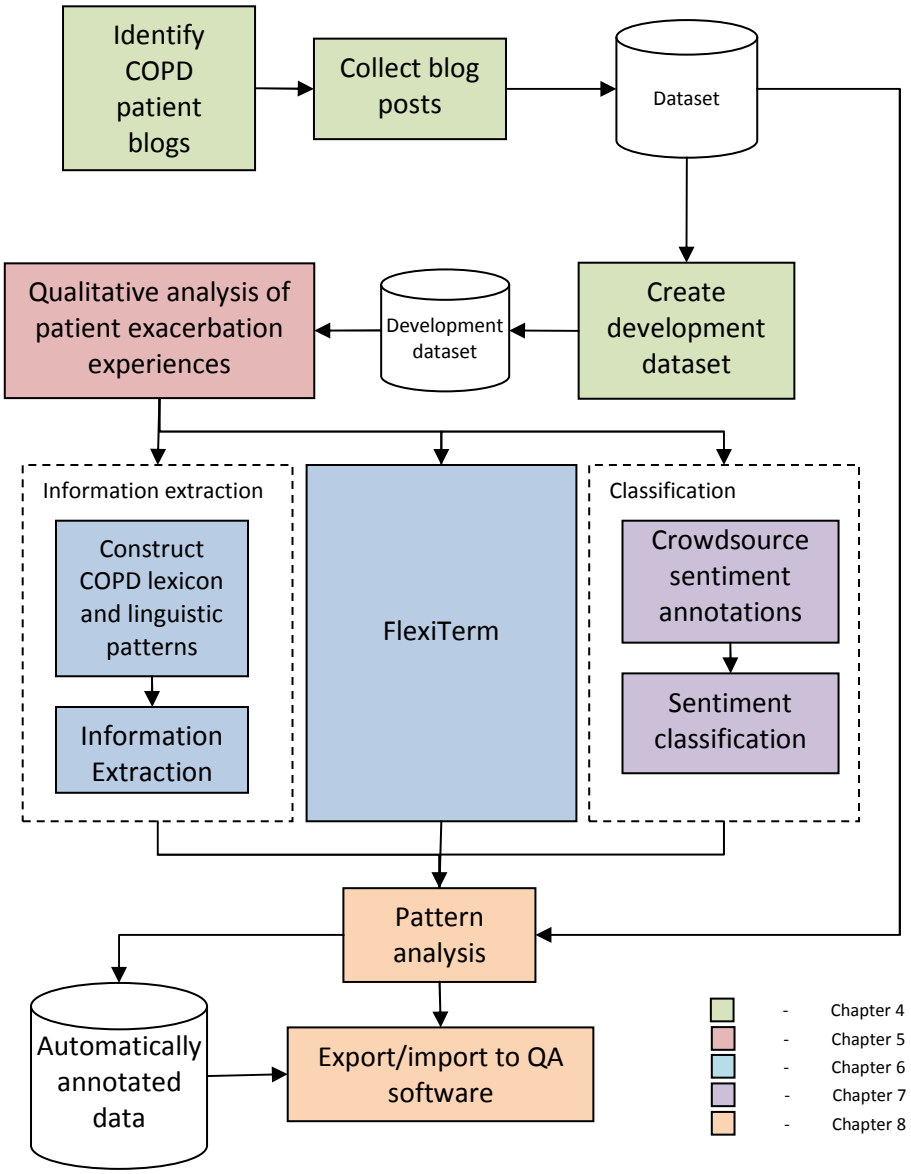


Figure 9: Applying QuTiP to investigate COPD patient experiences

Figure 9 summarises our overall approach, but in addressing each stage (as introduced in Section 3.1) we have developed and evaluated approaches to collect and process patient blog post data.

Data Collection

Our semi-automated approach to data collection combines specialised internet search tools with manual inspection of blog content for inclusion or exclusion from our investigation. This allows a potentially large amount of data to be collected with the rigour of manual inspection.

Qualitative Analysis

Our qualitative analysis of the COPD blog posts retrieved focuses on patients' conceptualisation and understanding of exacerbation as well as how they manage these events themselves.

Information Extraction

In order to extract mentions of entities related to COPD exacerbations we used literature review and standardised vocabularies alongside the outcomes of our qualitative analysis to construct lexicons and a set of linguistic rules to identify mentions of specific concepts in blog post text (e.g. symptoms, treatments, etc.). We also use these mentions to classify sentences as relating to COPD exacerbation or not.

Automatic Term Recognition – FlexiTerm

As well as labelling known concepts, we harness FlexiTerm to find concept mentions more generally and enable the serendipitous discoveries described earlier. FlexiTerm is tuned to perform well in informal data collections (such as blog posts) and its performance is evaluated here against another Automatic Term Recognition approach.

Text Classification

In this phase of our implementation of the QuTiP framework, we targeted sentiment information – relating to emotions and requirements expressed in text as well as the subjective nature of the authors' narrative (i.e. whether they are describing something personal or more general). We used a crowdsourcing exercise in order to generate labels for our dataset which we used to train and evaluate an automated approach to classifying the sentiment of sentences in patient blog posts.

The primary contribution of our work is the QuTiP framework, our approach to performing text mining in conjunction with qualitative research. However, our application of this framework also contributes methods to achieve the following:

- Gather patient blog posts from the Web (Chapter 4).
- Crowdsource annotated text (Chapter 7).
- Automatically extract domain-specific terminology from informal textual content (FlexiTerm – see Chapter 6).
- Automatically extract subjectivity of statements in patient blog posts, based on linguistic features (Chapter 7).

We also produce the following reusable resources:

- A database of manually curated COPD patient blog posts (Chapter 4), including annotations for:
    - Medical concepts (Chapter 6).
    - Subjectivity of sentences (Chapter 7).
- A node hierarchy for qualitative analysis of COPD patient experiences of exacerbation (Chapter 5).
- A manually constructed lexicon of medical terms related to COPD (Chapter 6), gathered from both:
    - Formal, medical terminologies.
    - Colloquial expressions derived from qualitative analysis.
- Tools for integration of labels from Information Extraction and classification software in to qualitative analysis packages (Chapter 8).

Lastly, contributions to our chosen medical domain include potential areas of further inquiry, drawn from both qualitative analysis (Chapter 5) and pattern analysis of further automated analysis (Chapter 8). In the remainder of this thesis, we discuss the phases described in Figure 9 including covering these contributions in more detail.

CHAPTER 4

COLLECTION OF COPD PATIENT BLOG POSTS FROM THE WEB

As with any analysis of documents or social data, the first step in the QuTiP framework is to collect a set of documents which are relevant to the questions posed. Whether this analysis is manual or automated, qualitative or quantitative, it is real-world documents that are the subject and finding this relevant data can be a challenge in itself. In this chapter, we discuss those challenges and describe our approach to finding relevant patient-authored blog posts on the World Wide Web.

This step can be considered as addressing an information retrieval problem – of all the potential documents, the aim is to find the ones that satisfy the constraints of our study. Approaches to this problem may be very strict and stringent, applying static rules for judging inclusion and exclusion for instance. In medicine, for instance, the Cochrane methodology for systematic literature reviews (Higgins & Green, 2011) sets out in great detail how researchers should design the search strategy to find relevant studies from the vast collection of medical literature. Methods for designing initial queries, finding relevant data sources and finally manually reviewing the results and applying inclusion/exclusion criteria to qualify them to be considered in the analysis are all defined. On the one hand, it is important to ensure coverage of all the potentially relevant areas, but ultimately paramount to be strict and focused in the documents analysed.

Analogous to the strict nature of data collection followed in systematic literature reviews, analyses of health social media tend to work with highly focused, relevant document collections. Authors often concentrate on specific networks they know are active and relevant to the topic of their study. Sillence & Mo (2012), Hewitt-Taylor & Bond (2012) and Keim-Malpass & Steeves (2012) focused on specific, topically relevant message boards and forums on the Web which they had vetted previously, considering each post to those platforms for inclusion. Likewise, much more general social networks can be searched for potentially relevant posts. Adams et al. (2011) used keywords relevant to current events

relating to stem cells in order to gather streamed social media reaction in real-time. Similarly, Chou et al. (2011) constructed queries to find YouTube users sharing their experiences of cancer survival.

While platforms such as Blogger and Tumblr offer blogging services to users in a centralised location on the web, blogging platforms (such as WordPress) can be installed as part of a users own website. This means that, as with the rest of the Web, there is no way to collect blog posts from any centralised place – individual blogs will be found on many different websites. Where previous studies have found particular platforms through which documents can be collected, our approach must allow for collection of documents from a more disparate collection.

As with the studies detailed above, our approach relies on a manual inspection of content to determine inclusion or exclusion for subsequent analysis. However, initially we use general purpose search engines and a query constructed with the help of standardised vocabularies to identify potentially relevant blogs from the Web. In the next section, the process we used to identify and collect relevant documents is set out, followed by a description of the resultant dataset.

4.1 METHOD

To create our dataset we combined the scalability of online search tools with the authority of a manual, expert review guided by systematically laid out inclusion and exclusion criteria. The general process is set out below in Figure 10 and each step will be described in detail in this section.

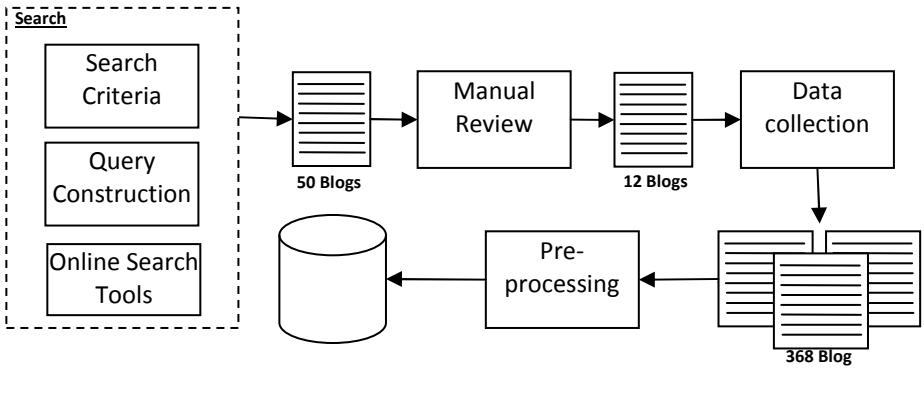


Figure 10: Data collection method

In the initial search step, we defined the information need (i.e. the inclusion and exclusion criteria), utilised expert knowledge and standardised knowledge sources to construct a query that would yield relevant results and utilised online search tools to identify an initial set of potentially useful blogs. Each of these blogs was analysed manually and judged against the criteria set out. All the available blog posts were then collected from the 12 blogs identified as good sources of COPD patient experiences and pre-processed before being stored in a database.

4.1.1 SEARCH

Using public search tools to find relevant blog posts poses two main challenges. Firstly, we must sufficiently represent the target domain as a search query. This means identifying useful terms and search operators to both specify the target topic to a sufficient degree to avoid noise (i.e. documents falsely identified as relevant), but also widely enough so that we get a useful coverage of available material – that there are few relevant sources that are not returned by the search engine. Secondly, we must also find ways of representing the *type* of document that we are looking at. In technical terms, a blog is another website and distinctions are not often made in search engines as to the type of document that is returned. Blogs are only singled out in terms of how they are used by authors and readers alike. For our information need, we have to find tools that allow us to identify blogs on the web based on a keyword query.

In order to address the first challenge, we utilised medical lexicons (through UMLS) to expand the term "COPD", using Boolean operators to include synonymous expressions for the condition. While we could have included terms relating to other concepts related to COPD (e.g. symptoms or treatments) to increase the coverage of our query, there is a lot of overlap between the condition we are focussing on and other respiratory diseases. While more documents might have been returned were we to expand our query in this way, a lot of noise would have been introduced. As such, we focused on terms relating to the name of the condition and related conditions. The query had the following structure:

("*COPD"* OR "*chronic obstructive {pulmonary|lung|airways|respiratory} disease"* OR "*bronchitis"* OR "*emphysema")*

In order to focus our search specifically on blogs we utilised online blog search tools (Google Blog Search[10], Technorati[11]). Each tool allows the user to specify whether they are

[2] http://www.google.com/blogsearch

looking for individual posts or for blogs which are related to the query string. In each of the search engines used we chose the 'blog'-level search.

Blogs were then selected from their titles and the descriptions returned by the search engines that were related to COPD. A total of 50 blogs were selected for an in-depth manual review.

4.1.2 Manual Review

Before a manual view of the blogs discovered through online search, we first set out criteria for what qualifies a blog as a good source to satisfy the requirements of this study. These criteria were then used to judge each blog for inclusion or exclusion in our research.

- Included blogs
    - written by patients or people caring for people living with Chronic Obstructive Pulmonary Disease (COPD); and
    - which primarily focuses on COPD
    - Published under sufficient license for inclusion in research
- Excluded blogs
    - do not principally focus on COPD
        - i.e. focus is on related issues, such as smoking
    - not written by patients or carers
        - e.g. written by physicians or companies relating to COPD

Each blog was visited and assessed based on the content of recent blog posts, the user descriptions and any description that was attached to it. A total of 38 of the 50 blogs identified were excluded. 14 blogs were not written by patients or carers (e.g. authored by physicians or organisations related to COPD). 19 blogs were marketing tools used by companies selling drugs or treatments for respiratory conditions. And finally we were not satisfied that 5 blogs were published under sufficient license for inclusion.

4.1.3 Collecting Content

In order to collect the content of the selected blogs, we used RSS feeds supplied through blogging platforms. RSS feeds provide a stream of recent posts in an XML format. While the web page presents content in an easy to read manner, the RSS feed provides access for automatic processing – for instance, for newsfeed aggregation services such as Feedly[12]. An

[11] http://technorati.com/
[12] http://feedly.com

example of RSS content is given in Figure 11 and Figure 12. Figure 11 shows an example blog, with the latest post listed first and the resultant XML stream from the RSS feed in Figure 12.

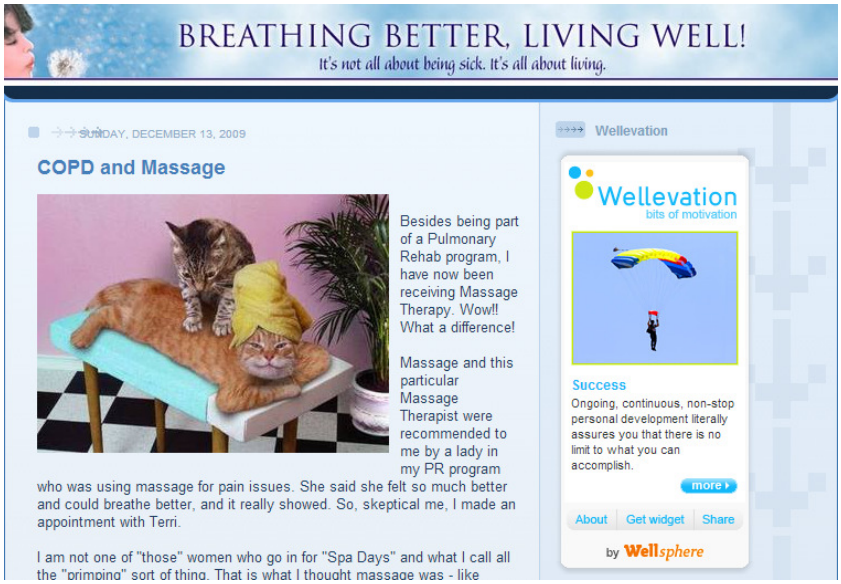


Figure 11: Example blog post – taken from http://breathingbetterlivingwell.blogspot.co.uk

```
<?xml version='1.0' encoding='UTF-8'?>
<?xml-stylesheet href="http://www.blogger.com/styles/atom.css" type="text/css"?>
<feed xmlns='http://www.w3.org/2005/Atom' xmlns:openSearch='http://a9.com/-/spec/opensearchrss/1.0/' xmlns:blo
xmlns:georss='http://www.georss.org/georss' xmlns:gd="http://schemas.google.com/g/2005" xmlns:thr='http://purl
    <id>tag:blogger.com,1999:blog-32722308</id>
    <updated>2014-03-18T23:43:28.252-04:00</updated>
        <title type='text'>Breathing Better Living Well</title>
    <subtitle type='html'/>
    <link rel='http://schemas.google.com/g/2005#feed' type='application/atom+xml' href='http://breathingbetter
    <link rel='self' type='application/atom+xml' href='http://www.blogger.com/feeds/32722308/posts/default'/>
    <link rel='alternate' type='text/html' href='http://breathingbetterlivingwell.blogspot.com/'/>
    <link rel='hub' href='http://pubsubhubbub.appspot.com/'/>
    <link rel='next' type='application/atom+xml' href='http://www.blogger.com/feeds/32722308/posts/default?sta
    <generator version='7.00' uri='http://www.blogger.com'>Blogger</generator>
    <openSearch:totalResults>60</openSearch:totalResults>
    <openSearch:startIndex>1</openSearch:startIndex>
    <openSearch:itemsPerPage>25</openSearch:itemsPerPage>
    <entry>
        <id>tag:blogger.com,1999:blog-32722308.post-8850483369796743483</id>
        <published>2009-12-13T16:31:00.004-05:00</published>
        <updated>2009-12-29T17:28:44.255-05:00</updated>
        <title type='text'>COPD and Massage</title>
        <content type='html'>&lt;a href=&quot;http://3.bp.blogspot.com/_do5Qi-P1HRc/SyVqK-a8kfI/AAAAAAAAAK4/vr
        style=&quot;MARGIN: 0px 10px 10px 0px; WIDTH: 319px; FLOAT: left; HEIGHT: 234px; CURSOR: hand&quot; id
        &quot; border=&quot;0&quot; alt=&quot;&quot; src=&quot;http://3.bp.blogspot.com/_do5Qi-P1HRc/SyVqK-a8k
        &quot; /&gt;&lt;/a&gt;&lt;br /&gt;&lt;div&gt;Besides being part of a Pulmonary Rehab program, I have n
        difference!&lt;/div&gt;&lt;br /&gt;&lt;div&gt;Massage and this particular Massage Therapist were recom
        using massage for pain issues. She said she felt so much better and could breathe better, and it reall
        appointment with Terri.&lt;/div&gt;&lt;br /&gt;&lt;div&gt;&lt;/div&gt;&lt;div&gt;I am not one of &quot
        &quot; and what I call all the &quot;primping&quot; sort of thing. That is what I thought massage was
        were okay to do, but I thought of as a waste of money. Luxuries, you know?&lt;/div&gt;&lt;br /&gt;&lt;
        shoes...or a something for my grandson ...than &lt;em&gt;waste&lt;/em&gt; the money on myself! Go out
        &gt;&lt;br /&gt;&lt;div&gt;I have now gone five weeks in a row, and even my husband is very impressed
```

Figure 12: Example RSS feed – taken from http://breathingbetterlivingwell.blogspot.com/feeds/posts/default

4.1.4 PRE-PROCESSING AND STORAGE

Pre-processing the blog posts involves augmenting the plain text with linguistic information to support further processing. The text was stored in the data structure described in Figure 13.

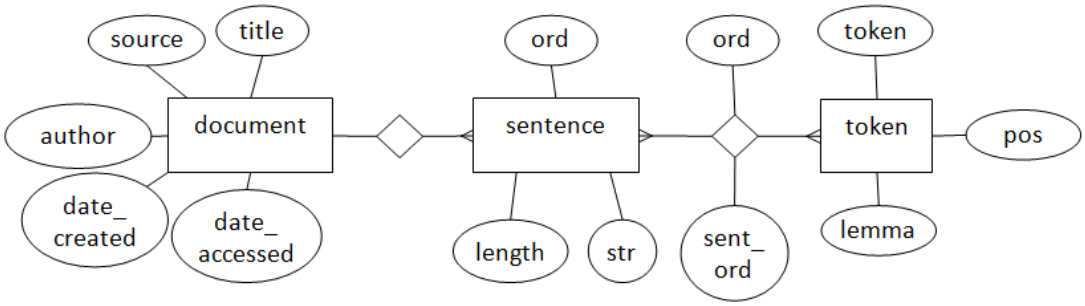


Figure 13: Document E-R diagram

Linguistic information was extracted automatically, using the Stanford part-of-speech (POS) tagger (Toutanova, Klein, Manning, & Singer, 2003). The Stanford POS tagger uses a maximum entropy model, trained on a corpus of articles from the Wall Street Journal and splits documents into sentences and tokens. Although other POS taggers are available, Stanford offers pre-packaged general purpose language models, which would prove useful for out particular corpus. Along with splitting tokens, their part-of-speech tag is automatically assigned. A token's part-of-speech tag (for instance, *noun*, *adjective*, etc.) is a category assigned to describe its syntactic and morphological properties. As described in Chapter 1 this low-level linguistic information is used to extract semantic and sentiment information later on (see Chapters 6 and 7). Figure 14 shows an example blog post in its plain text and pre-processed (XML) forms.

Plain text

I hope you are all staying clear of nasties this summer and able to enjoy what sunshine we have had so far. I have been unable to access this site for several weeks and finally sought Derek's help again via e-mail. I took his advice and have now managed to post this blog using Google Chrome instead of Internet Explorer which I would normally use. So thanks very much Derek for your helpful input. Your recent postings have been very interesting - I hope you can look forward now to a period of wellness with no more colds and infections for the foreseeable!

Pre-processed text

```
<document>
<sentence><PRP>I</PRP> <VBP>hope</VBP> <PRP>you</PRP> <VBP>are</VBP> <DT>all</DT> <VBG>staying</VBG> <JJ>
clear</JJ> <IN>of</IN> <NNS>nasties</NNS> <DT>this</DT> <NN>summer</NN> <CC>and</CC> <JJ>able</JJ> <TO>to
</TO> <VB>enjoy</VB> <WP>what</WP> <NN>sunshine</NN> <PRP>we</PRP> <VBP>have</VBP> <VBN>had</VBN> <RB>so
</RB> <RB>far</RB> <PUN>.</PUN> </sentence>
<sentence><PRP>I</PRP> <VBP>have</VBP> <VBN>been</VBN> <JJ>unable</JJ> <TO>to</TO> <NN>access</NN> <DT>
this</DT> <NN>site</NN> <IN>for</IN> <JJ>several</JJ> <NNS>weeks</NNS> <CC>and</CC> <RB>finally</RB>
<VBD>sought</VBD> <NNP>Derek</NNP> <POS>'s</POS> <NN>help</NN> <RB>again</RB> <IN>via</IN> <NN>e-mail
</NN> <PUN>.</PUN> </sentence>
<sentence><PRP>I</PRP> <VBD>took</VBD> <PRPS>his</PRPS> <NN>advice</NN> <CC>and</CC> <VBP>have</VBP> <RB>
now</RB> <VBN>managed</VBN> <TO>to</TO> <VB>post</VB> <DT>this</DT> <NN>blog</NN> <VBG>using</VBG> <NNP>
Google</NNP> <NNP>Chrome</NNP> <RB>instead</RB> <IN>of</IN> <NNP>Internet</NNP> <NNP>Explorer</NNP> <WDT>
which</WDT> <PRP>I</PRP> <MD>would</MD> <RB>normally</RB> <VB>use</VB> <PUN>.</PUN> </sentence>
<sentence><RB>So</RB> <NNS>thanks</NNS> <RB>very</RB> <RB>much</RB> <NNP>Derek</NNP> <IN>for</IN> <PRPS>
your</PRPS> <JJ>helpful</JJ> <NN>input</NN> <PUN>.</PUN> </sentence>
<sentence><PRPS>Your</PRPS> <JJ>recent</JJ> <NNS>postings</NNS> <VBP>have</VBP> <VBN>been</VBN> <RB>very
</RB> <JJ>interesting</JJ> <PUN>-</PUN> <PRP>I</PRP> <VBP>hope</VBP> <PRP>you</PRP> <MD>can</MD> <VB>look
</VB> <RB>forward</RB> <RB>now</RB> <TO>to</TO> <DT>a</DT> <NN>period</NN> <IN>of</IN> <NN>wellness</NN>
<IN>with</IN> <DT>no</DT> <JJR>more</JJR> <NNS>colds</NNS> <CC>and</CC> <NNS>infections</NNS> <IN>for
</IN> <DT>the</DT> <JJ>foreseeable</JJ> <PUN>!</PUN> </sentence>
</document>
```

Figure 14: Plain and pre-processed text

4.2 SUMMARY OF THE DATASET

The corpus of collected blog posts is described in Table 4.

Blogs	12
Authors	44
Blog posts collected	368 (819KB)
Mean length (tokens)	461 (std dev: 402)
Post dates	2006-2012
Sentences	7955
Tokens	165042
Distinct tokens	13861
Mean sentence length (tokens)	20.7 (std dev: 14.5)

Table 4: Blog corpus properties

Sentence, tokens and distinct tokens are based on the result of the linguistic pre-processing described in Section 4.1.4. Tokens were grouped by their string representations alone and

not by meaning. The dataset includes posts from 44 authors, published over a period of 7 years.

4.3 CONCLUSION

Following the process laid out in this chapter, we have created a corpus of 368 blog posts written by 44 people living with, or those caring for people with, COPD. As with other studies of health-related social media, the aim is to utilise the experiences shared to learn about how patients manage their conditions and the issues and difficulties they face. The blog posts we have collected have been linguistically processed to enable further analysis.

While supported by search tools, the process we followed to identify relevant sources of blog posts was primarily a manual one. Through interpretation of the document sources, we can ensure to a greater degree that the documents satisfy our information requirement and inclusion criteria. However, the time consuming nature means that automating this process would be required to scale blog post analysis. Manually curated collections, such as ours, offer training and testing data to create models for automating document relevance decisions based on high-confidence examples.

For Web documents, topic-focused web crawling (De Bra & Post, 1994) is an active area of interest where networks of documents on the Web are exploited to find those related to some pre-defined topic. As each document is retrieved from the Web it is classified as either relevant or irrelevant and included or excluded from the collection, and hyperlinks in that document are then followed to find more potentially relevant documents (M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, 2000). Approaches in topic-focused Web crawling aim to increase the efficiency of this process by prioritising visits based on the target documents' expected relevance. Topic-focused Web crawling offers the potential to help automate this stage in the QuTiP frame work. In future work, we aim to explore how the resultant document set could help train models to automatically garner relevant social media from the Web.

Through manual, qualitative analysis in conjunction with automated interpretation and mining of this textual data resource, the remainder of this thesis is dedicated to making use of the data collection curated during this chapter.

CHAPTER 5

PATIENT EXPERIENCES OF COPD EXACERBATION – A QUALITATIVE STUDY

As discussed in Chapter 3, the initial qualitative analysis of the collected data has a number of roles in applying the QuTiP framework. Firstly, the outcomes of this qualitative investigation can help define the aims of the research as a whole. The initial exploration of the development subset can help develop the questions asked of the larger dataset. The secondary outcomes of a qualitative study – the annotations and themes extracted – can aid in developing the automated text mining approaches. The aim of QuTiP is to utilise the outcome of this small-scale investigation to develop tools and approaches to allow the analysis to be scaled up and to support researchers in carrying out qualitative research on much larger, more representative sets of patient-authored blog posts.

In applying QuTiP in our research, we defined an initial set of questions relating to patient experiences of COPD exacerbations. Through systematic qualitative analysis of the dataset, we extracted annotations and themes related to the way patients understand and identify exacerbations of their condition and the impact that this acute worsening of their condition has on their lives. In this chapter we discuss existing qualitative approaches to investigating patient experiences of COPD exacerbation, motivating this area of inquiry. We also discuss why social media might offer further insights in to how patients identify and manage their exacerbations. We introduce our approach to analysing the development set of 100 patient blog posts (see Chapter 4), the questions we set out to answer and the outcome of our analysis. In concluding the chapter, we highlight the primary outcomes of this research (the themes extracted from the data) and identify how the secondary outcomes (the annotations and thematic hierarchy) will be used in developing text mining approaches.

5.1 Qualitative analysis of exacerbation accounts

A number of qualitative studies have described COPD patients' perceptions of their illness experience, including how they manage their condition. Interviews with patients have been used to investigate how they monitor their condition and discover and manage exacerbations, focussing on their motivations for seeking consultations (R. Adams, Chavannes, Jones, Ostergaard, & Price, 2006) and how they conceptualise, communicate and respond emotionally to exacerbations (Kessler et al., 2006; Michaels & Meek, 2004; O'Neill, 2002; Thomas, 2009).

As well as understanding patients' experience of living with a condition, their experiences of healthcare have also been investigated using qualitative analysis. Access to healthcare services is one common theme in this area of research. While the debilitating nature of COPD is one barrier patients face while accessing services (Gysels & Higginson, 2008), patients' reluctance to seek treatment through social and environmental pressures (Gysels & Higginson, 2008), normalisation of their condition (i.e. believing their condition to be 'normal') (Habraken, Pols, Bindels, & Willems, 2008) and doctors attitudes towards diagnosis (Walters, Hansen, Walters, & Wood-Baker, 2008) have all been highlighted by patient groups as causes of delaying treatment. One possible cause is lack of information or support (Gruffydd-Jones, Langley-Johnson, Dyer, Badlan, & Ward, 2007; Rodgers, Dyas, Molyneux, Ward, & Revill, 2007).

That the information comes from the patient's point-of-view makes these insights in to living with COPD valuable. While traditionally these studies rely on data gained through, for example, interviews (Clarke, Sohanpal, Wilson, & Taylor, 2010; Pinnock et al., 2011; Reinke et al., 2008; Schofield, Knussen, & Tolson, 2006), focus groups (Gruffydd-Jones et al., 2007; Kessler et al., 2006), questionnaires (P W Jones et al., 2009; Ojoo et al., 2002; Pilling, Bassett, & Wolstenholme, 2003) and patient diaries (Paul W Jones, Chen, Wilcox, Sethi, & Leidy, 2011), this chapter focuses on carrying out similar analysis on patients' self-authored blogs.

5.2 Methods

We analysed COPD patients' accounts shared online in order to explore their experiences of acute exacerbation. To make the most of the self-reported nature of the blog posts, we opted for an inductive approach, allowing key themes to emerge from the data without a pre-formed hypothesis. Analysis of the data was carried out by the author in three phases

adopting an inductive approach. Firstly, familiarity with the dataset was established through an initial review of the blog posts. After this phase the questions that the analysis would aim to answer were designed. The questions we aimed to answer were based on the nature of the discussion in the blog posts, previous literature on qualitative analysis of COPD exacerbation accounts (see Section 5.1) and input from two medical professionals. The questions we aim to answer are:

1) How do patients conceptualise a COPD exacerbation?
2) What is the impact of exacerbations on patients?
3) How do patients assess their condition and make self-management decisions?

The second phase involved the initial coding of the data - creating a hierarchical code set based on the content. As described in the QuTiP framework (Chapter 3), this analysis was carried out on a development set of 100 blog posts, of the 368 collected in Chapter 4. This analysis focused solely on discussions about COPD exacerbations in the blog posts, leaving out discussions of wider COPD-related topics, or non-related issues. Lastly, we identified core themes related to our research questions from the coded data. This process involves merging and organizing nodes in the coding hierarchy in order to illustrate key themes within the data.

				docs	ref
Exacerbation				42	74
	General comment			2	3
	Physical exertion			6	6
		Impact		2	2
	Prevention			0	0
		Medication		3	4
		Avoiding physical activity		3	3
		Slowing down		1	1
		Exercise		5	5
		Improve air quality		1	2
		Avoid risks		2	2
	Infection			4	4
		Discovering exacerbation		11	15
			Patient notices change in condition	7	11
			Diagnosis at doctor's appointment	3	3
			Emergency hospital (ambulance)	1	1
		Treating Exacerbation		8	15
			Vitamin D for exacerbation prevention	2	2
			Sleeping in a chair for comfortable breathing	1	2
			Prescribed	2	2

			antibiotics		
			Home supply antibiotics	3	3
			Prescribed steroids	3	3
			Prescribed nebulizer capsules	1	1
			Decline steroids or antibiotics	1	1
			Over the counter cold remedies	1	1
		Monitoring Symptoms		11	39
			Sleep disturbance sign of exacerbation	2	2
			Phlegm	2	2
			Vigilance of Infections Close by	2	3
			Tiredness	1	2
			Pain	2	3
			Temperature	1	1
			Waiting for chest symptoms before treating	5	8
			Breathing	5	8
			Weather as cause of SOB	1	1
			Heart rate	1	3
			Oxygen levels	1	3
			Waiting too long for treatment	1	1
			Upper respiratory symptoms	1	1

			Cough	1	1
		Impact of exacerbation		5	8
			Disturbance to work	2	3
			Hospitalizatio n	3	4
			Miserable	1	1
		Issues		3	3
			Availability of medication	2	2
			Medication side-effects	1	1
		Self-diagnosis		1	1
	Personal limitations			4	4
	Learning techniques			3	4
	Weather_air quality			10	14
	Sharing information			6	7
	Unpredictable			1	1

Table 5: Initial exacerbation node hierarchy - phase 2

			Documents	References
How do patients conceptualise COPD exacerbation			**19**	**31**
	Exacerbation as limiting		11	13
		Limiting	10	12
		Torturous	1	1
	Exacerbation as relative		2	3
	Powerlessness		11	15
		Controllable	4	7
		Uncontrollable or inevitable	7	8
What is the impact of exacerbations on patients			**5**	**6**

	Physical		4	5
		Aesthetic impact	1	1
		Incapacity	2	2
		Slowing down	2	2
	Psychological		1	1
		Risk avoidance	2	2
		Avoiding physical activity	3	3
		Confidence	2	2
		Distress	1	1
		Loneliness	1	1
		Nervousness	1	1
How do patients assess their condition and make self-management decisions			**4**	**5**
	General energy levels		1	1
	Incidental diagnosis		1	1
	Speed of change		2	3

Table 6: Thematic analysis outcome

In Table 5 and Table 6 the outcomes of both the initial coding phase and the thematic analysis carried out on the data are set out. The analysis focused on COPD Exacerbation, so Table 5 contains only those nodes. Other nodes were used to annotate the data as part of this initial coding phase, the remaining nodes can be found in Appendix B. As shown in Table 5, 42 of the 100 documents analysed included discussion about some aspect of exacerbation. In the remainder of this section, the themes which were extracted will be explored in more detail along with their contribution to answering the questions posed.

5.3 RESULTS

This section sets out the themes identified in the data using the qualitative approach described in the previous sections. As shown in Tables Table 5 and Table 6 there is diverse discussion about exacerbations throughout the blog post which were analysed. While the initial codes ascribed to the data (Table 5) offer a low-level view of what is discussed in the blog posts, the thematic analysis (Table 6) draws them together through interpreting the relevant references in order to extract higher-level themes relevant to the inquiry. In this

section, we describe those themes, present relevant examples from the blog posts and describe the interpretation and how it supports our observation.

5.3.1 How do patients conceptualise a COPD exacerbation?

How patients conceptualise or understand exacerbations is likely to have an impact on how they decide to manage them. For instance, if increased shortness of breath is equated with having an infection, then this could affect how a patient is likely to manage their symptoms. Conversely, if respiratory symptoms are thought to be due to a general worsening of the condition rather than an acute exacerbation caused by external factors, this could delay consultation with a physician and prolong or worsen impact.

We sought to utilise COPD patients' discussion about their condition in order to discover *how* patients discussed exacerbations.

5.3.1.1 How patients describe exacerbations

As previously mentioned, discussion of exacerbation appeared in 42 of the 100 blog posts analysed. However, the word 'exacerbation' (or its variants) appears in just 9 posts. Exacerbations are mostly discussed implicitly, relating to causes, symptoms, treatments and the personal circumstances faced.

5.3.1.2 Exacerbations as limiting

Throughout the data, exacerbations were frequently discussed in terms of the limitations they imposed on people living with COPD and the difficulty they had in adjusting to these limitations.

"Over the past year, I have been forced to move even slower, now at a snail's pace. The surprising benefit that I found is better breathing. It took broken bones to make me pace even more, and move even more slowly while doing things. I feel pretty comfortable doing the snail's pace, and I don't feel embarrassed as I had previously."

– Quote 1: author A, post 139

While describing an overall decline in their condition, the author discusses the limitations that the threat of exacerbations has placed on them. As their condition is worsening, their tolerance to physical activity is weakening. Exercise (e.g. as part of a pulmonary rehabilitation course) is one way of increasing tolerance to physical activity, so accepting the decline and 'slowing down' may not always be the best course of action. The author also implies that they were once embarrassed about the fact that they were no longer as

physically fit as they once were. This extra psychological stress may have further implications for a patient's psychological and social well-being.

5.3.1.3 EXACERBATIONS AS RELATIVE

Exacerbation is defined as an acute worsening of a patient's condition – i.e. beyond what might be expected through deterioration as part of a long-term condition. The online discussions analysed show both understanding and confusion relating to the distinction between deterioration and exacerbation.

In Quote 2 the blogger discusses getting back to normality, describing it as 'what would be normal' for them – describing their current condition as a deviation from that personal level of wellness. Due to the length of the exacerbated period (the post is dated January and mentions it began in October) and the attempts to treat as an infection (antibiotics and steroids) they are now investigating other, longer term causes of a deterioration in breathing function. It appears that while this patient is discussing the worsening of their condition as an acute period of exacerbation it may instead be related to a longer-term decline in their condition (i.e. comorbidities).

"Having read the posts so far it would appear that we are all having a torrid time since the beginning of Autumn with long term infections etc. I seem to be in the same boat as everyone, having been on various antibiotics and courses of steroids since early October and am still struggling to get back to normal or what would be normal for me! I am currently undergoing heart tests to establish if my deterioration in breathing is possibly caused by a heart problem."

– Quote 2: author E, post 333

However, in Quote 3 the author describes improvement after treatment for two respiratory infections ('*nasties'*) in terms of relative improvement in their condition.

"I have noticed this last month my breathing since the two nasties have been dealt with has been much better. And that when I do become breathless recovery is much faster than before as I am able to breathe deeper."

– Quote 3: author B, post 350

5.3.1.4 POWERLESSNESS

Patients often discussed their exacerbation in terms of how powerless they feel over this aspect of their illness. Patients often discuss their condition in terms of factors that are outside of their control – for instance, the weather or air quality.

"During last week I had two whole days when I felt on top of the world. My breathing was so much better, and I was much less breathless than usual – and I felt better than I had done for well over a year. This did come to an end on the day I was due to see my respiratory nurse and I mentioned how I had felt the previous two days. It was explained that what I experienced happens when a combination of good things come together. Less polution, the right weather be it temperature, humidity, and I expect a lot of other things. I now think the best thing to do when I get one of those rare good days is to just enjoy even though I know it is not going to last for too long. "

– Quote 4: author C, post 4

"It seems this changeable weather is taking its toll once again very SOB."

– Quote 5: author C, post 4

One carer discusses the fine balance of these external factors:

"Of course the colder and rainier weather has had an impact and he is certainly more breathless today and feeling a little deflated. As soon as the air is more damp the cough becomes worse but on the other hand, too dry and the same! Such a fine balance."

– Quote 6: author D, post 5

Avoiding infections at certain times of year or from close family members is also highlighted by bloggers as a cause of difficulty.

"I guess it was bad luck to be hit by one and ending up in hospital, only for shortly after to get a flue like cold that pushed me into another bout of illness...On this occasion on my part it would have been hard to avoid as it ran through the family. I can't exactly leave the wife, or ask her to move out if she catches a bug."

– Quote 7: author B, post 13

"It's always difficult this time of the year still fighting off infections hoping for a respite. So Far this year I have had six lots of antis fingers crossed again."

– Quote 8: author C, post 296

The authors conceptualise exacerbation in terms of how it affects their everyday lives and how this impact can be reduced. Information and advice is gathered from a variety of sources and propagated through the community in order to support others in similar circumstances. The distinction between acute exacerbation and overall decline in condition seems present in some cases, but not in others and reluctance to seek help when feeling worse is mentioned in the analysed posts.

5.3.2 What is the impact of exacerbations on patients?

The impact of exacerbations was mostly discussed within the data in terms of *avoiding* exacerbation. The authors shared the many ways that suffering from COPD and the stresses this places on them; both physically and emotionally. In this section, we discuss the impact exacerbations have on COPD sufferers from these two perspectives.

5.3.2.1 Physical impact

Patients use many signs and symptoms through which to monitor their condition and keep exacerbations in check. Breathlessness and cough symptoms are discussed including phlegm and discomfort as a result (see Quotes 9 and 10). Aching and pain are also mentioned (Quotes 11 and 12). One author in particular gave a day by day account of their heart rate and oxygen levels (Quote 13).

"But by the end of last week, and more so this week, I have felt better and more able to tolerate movement without disabling shortness of breath."

– Quote 9: author A, post 149

"Lots of nasty looking phlegm and can't stop coughing this evening. It's back to the settee or chair tonight. I had managed to get a couple of nights in bed, well, until the early hours when I woke up feeling uncomfortable and needing to go get a warm drink to clear the gunk in my throat. I hate that feeling, and I only seem to get it when laid down in bed."

– Quote 10: author J, post 1169

"I knew I had it (respiratory infection) when I felt very tired early one evening, very unusual for me, and laid on the settee to sleep feeling like I had been run over by a hundred steam rollers. I kid you not I had parts of me aching that I did not know I had."

– Quote 11: author B, post 350

"I had an appointment with my Pulmo. Doc the following week and guess what...yupper Pneumonia again! My back hurt so bad I couldn't even take a deep breath without excruciating pain!"

– Quote 12: author K, post 159

"Day four my pulse was racing and my oxygen levels were dropping"..."So yesterday (day five) my pulse were racing near 100 when resting, my oxygen levels were very low at rest"

– Quote 13: author B, post 293

As discussed in Quotes 14 and 15, as the frequency of exacerbations increases, patients may need to make decisions about whether or not they can continue to work. Quote 15

also suggests that allowances afforded to people who are forced to leave work by disability can be difficult to obtain.

"Over the last couple of years I have had more frequent exacerbations and now no longer work."

– Quote 14: author E, post 333

"There comes a point in time where we start to realize that working is becoming quite an impossible task. Many with COPD have to go on social security disability and just as many dread having to go through the process."

– Quote 15: author F, post 191

Authors also shared with others what aspects they found difficult due to their condition, and what they did to try and alleviate these difficulties. In Quotes 16 and 17, the author shares aspects of their life which have changed due to their condition. Firstly, in Quote 16 the author shares that they've learned how to adjust their behavior in order to reduce strain and the chance of exacerbation of their condition. Likewise, in Quote 17, they share their experience of a massage therapist and the difficulty they had with their breathing while lying flat.

"Over the past year, I have been forced to move even slower, now at a snail's pace. The surprising benefit that I found is better breathing. It took broken bones to make me pace even more, and move even more slowly while doing things."

– Quote 16: author F, post 139

"I do not have a problem lying flat, if I do it in stages. I cannot go from moving around a lot to just lying down. Some people with breathing issues simply need to have their head raised a bit and <massage therapist> can work with that."

– Quote 17: author F, post 132

The physical impact of exacerbation and the prevention of exacerbation is discussed in blogs in terms of the things that people find they cannot do anymore due to the difficulties in breathing they suffer from as a result. These may not necessarily be particularly taxing physical activities, but specific positions or movements that are no longer possible or as easy as they once were. The impact this has on the sufferer can make things like maintaining employment difficult.

5.3.2.2 PSYCHOLOGICAL IMPACT

The psychological impact of exacerbations and the threat of exacerbation stem from the physical factors discussed in the previous section. The distress of not being able to breathe

is the source of much of the psychological impact described in the blog posts (Quote 18). Likewise, Quote 19 graphically illustrates the burden of suffering from an exacerbation.

"Yes two [respiratory infections] in two months. I have to admit it has made me feel rather miserable. As anyone with copd will know it is distressing to struggle to breathe."

– Quote 18: author B, post 13

"Finally, I'm able to finish getting ready. I was praying it would be one of those days that would get better as it goes on. Not today, each step I took in my work shoes felt like they weighed 25 lbs each. I'd just take a couple of steps and I'd be short of breath. Each time I would get back from getting a grade I would feel as if I was going to collapse, darkness enveloping my head and a very powerful weight on my chest. My day was just pure torture."

– Quote 19: author G, post 122

Two main themes emerged which illuminate the psychological impact of exacerbation on COPD sufferers. Firstly, the fear described by authors when faced with activities beyond what they thought themselves capable of is evident throughout. In Quote 20, a carer discusses the self-defeating nature of the fear the person they care for feels towards physical activity.

"that his fear of doing something and becoming breathless actually is a downward spiral"

– Quote 20: author D, post 348

Exercise is utilised in pulmonary rehab to raise tolerance to physical activity, but Quote 21 and 22 describe two people's fear about being able to participate in the exercises.

"I walked in and looked at all those people on the various machines, even the ones wearing oxygen, and nearly charged back out the door! Believe me, I was petrified! No way was I ever going to be able to do this exercise machine stuff!! I just couldn't see how I could ever do any of it. (And I wasn't even on oxygen yet!!)"

– Quote 21: author A, post 134

"I was aghast, how could I withstand the rigors of exercise when I couldn't walk from room to room in my home without being breathless?"

– Quote 22: author G, post 118

Vigilance to risk is also a common theme within the data. Avoidance of exacerbation risks (e.g. physical activity and respiratory irritants or infections) was discussed in a number of ways including sharing strategies (see Quote 23) and frustration with trying to stay well (Quotes 24 –26) including impact on social interaction (Quotes 25-26).

"Also avoid the company of people who smoke, never take smoking seats in the restaurant even if you have to wait."

– Quote 23: author I, post 23

"Normally you get on with life and avoid things you cannot do...but sometimes you forgot. I did try carrying some panels with him but fell, took my breath away."

– Quote 24: author C, post 279

"I have came across several people with flue. The twins, my two grandchildren, have gone down with several colds, I can t of course push the kids away can I."

– Quote 25: author B, post 239

"I feel pretty comfortable doing the snail's pace, and I don't feel embarrassed as I had previously. I guess there's been that pesky pride thing going on that didn't want anyone to think I was less of a person because I had to go slow. I had to learn the hard way, as most lessons are learned, that pride never helped me breathe better."

– Quote 26: author F, post 139

Exacerbations and efforts to avoid exacerbations seem to lead to a loss of confidence in the ability to participate in physical activity, and fear at the prospect of engaging in what can seem like an impossible challenge. As engaging in physical activity is part of the treatment to increase tolerance and alleviate patients' symptoms this trepidation could serve as a barrier to treatment uptake. Similarly, while being vigilant of risk factors could be beneficial, patients' anxiety could be increased and their relationships with others could suffer as a result. Future work could explore how patients perceive risk factors, what strategies they employ to avoid them and what impact this has on their well-being.

5.3.3 How do patients assess their condition and make self-management decisions?

As shown through the self-reported experiences discussed so far, the authors of these blogs are vigilant of their condition, what exacerbates it and how these things can be prevented. In this section, we discuss some signs and symptoms specified by bloggers that indicate to them that their condition is worsening – that their condition is becoming exacerbated.

5.3.3.1 General energy levels

Fatigue (tiredness) was commonly described as an early indicator of an exacerbation. In Quote 27 a COPD sufferer describes the onset of a respiratory infection and the first sign being tiredness that 'unusual' for them. Exacerbation is relative (as discussed in Section 3.2.3) and so patients must monitor their condition in terms of this norm and look for

peculiarities in good time. This vigilance of condition and of risk factors was covered in Section 5.3.1.2.

"I knew I had it [respiratory infection] when I felt very tired early one evening, very unusual for me, and laid on the settee to sleep feeling like I had been run over by a hundred steam rollers. I kid you not I had parts of me aching that I did not know I had."

– Quote 27: author B, post 13

In Quote 28 the author has been required to make a number of journeys by foot and describes the physical limitations (see Section 5.3.1.2 and 5.3.2.1) felt when making them. Fatigue is described as the primary presentation of the difficulties brought about as a result of their emphysema.

"I found myself more short of breath and noticed that the distance in my walking ability seems to have shortened. Fatigue was at the forefront, as were some emotions that seemed to take on a life of their own. Anger and confusion seemed to grab hold and they felt so foreign to me and I couldn't control them. It's so weird and disappointing to realize how the effects of the emphysema take just a little bit more from me each day."

– Quote 28: author F, post 198

5.3.3.2 SPEED OF CHANGE

As mentioned previously, COPD is a degenerative, progressive condition and the condition of those suffering from the disease is likely to worsen over time. Exacerbation describes an acute period of worsening at a rate faster than would be expected as part of the natural course of the disease. It is this speed of change that patients may use to monitor their symptoms. In Quotes 29 and 30 the author describes their rapid deterioration, which they had accepted at first, before a consultation at a hospital revealed it was respiratory infections (or '*bugs*') affecting their condition rather than natural progression. This author shares advice with their peers to seek help when taking a 'major turn for the worse' rather than accepting it as a natural part of living with chronic lung disease.

"In the last year, even though my mobility was severely limited due to getting breathless, I noticed I had suddenly started to worsen rapidly. This was unexpected, as although we do become disabled due to COPD, as I had done after many years with the condition, and there is no cure, the downhill spiral of the illness is a slow one rather than rapid. A visit to the hospital showed my lung function had declined at an alarming rate. Much faster than was expected or the norm. What was the most alarming was if it continued to fall at the rate it had in such a short time, I would be on oxygen 24/7 within a year, and could even succumb

within two or three. What I had not realised was that I was in fact seriously ill with two major bugs that had set up home deep in my lungs, and were slowly killing me. It was only when the bugs were rampant and making me so breathless I could not stand and was rushed to hospital that the problem was found. I have to warn you guys and gals that us with COPD are more prone than usual than fit people to get these infections, so be aware."

– Quote 29: author B, post 350

"The moral of the story here for you my friends with COPD is – if you suddenly take a major turn for the worse. Don't accept it as I had. Question it and get down the doctor. Ask for a simple test, a sputum test to ensure you do not have something not nice in your lungs. It could save you a lot of distress, or even your life."

– Quote 30: author B, post 350

In post 13 (mentioned in Quote 27) author B uses their blog post as an illness diary. Continuing from the first onset of the disease (described in Quote 29), Quote 31 shows the patient charting the course of their condition over a number of days to the point at which they decided to use the antibiotics kept in the home.

"From then on it was a bit like a roller coaster, feeling fine, then ill again, for three days but it did not at that time hit my chest or effect my breathing. Day four of it did though. I slept most of day three, my breathing was getting worse. Then at 2am a trip to the bathroom and alarm bells were ringing. I was so breathless it felt like I had run a marathon. That was the time to get to the medicine cabinet, find the antibiotics I keep in the house, and get them down me fast. "

– Quote 31: author B, post 13

The patient waited for sudden dyspnea (or the infection to 'hit their chest') before treating as a respiratory infection and as an exacerbation. Besides administering themselves antibiotics they also sought to address potential irritants to lung function and an emergency appointment with their doctor (see Quote 32).

"Coming back up my wife woke and was quite alarmed, asking if I wanted her to get me to the hospital. I said no and with the air purifier on in the bedroom to cool me as I was burning up, lay on the bed to sleep. An emergency trip to the doctor of course followed. From where I got some steroids, looks like I will be climbing the walls by the weekend..."

– Quote 32: author B, post 13

5.4 CONCLUSION

The experiences shared online by COPD patients are valuable in understanding the physical and psychological implications of the condition from the patients' own, unprompted perspective. Blogs used as diaries to chart patients' lives can help us understand the factors and processes used to make treatment decisions and the information upon which these decisions are made. The themes highlighted in Section 5.3 relate strongly to one another. The limitations that this condition places upon patients can be a cause of great struggle, both physically and psychologically - leading to fear and anxiety about risk factors as well as sadness and embarrassment. Educating patients about recognising and distinguishing exacerbation from natural progression is important. Within the blog posts analysed, there were many instances of patients being vigilant of their condition and of risk factors, but deciding between acute exacerbation and natural progression can be difficult. The themes highlighted in this analysis require further investigation and could be built upon through further analysis of blogs posts or as a basis for further study through interviews with stakeholders.

Besides the primary outcomes of the analysis (i.e. the experiences described, the themes highlighted and the conclusions drawn) the secondary information created through the process described in Section 5.2) have the potential to be incorporated in to subsequent analyses of social media content relating to patient experiences of COPD. Coding and thematic frameworks can be reused in further analyses of similar datasets. Similarly, and more important in our work, this information can be used to create new tools and approaches to discovering information relating to patient experience. The data produced here will form the basis of lexicons, rules and information extraction methods to help automatically garner some of the information which has been manually extracted here. With example annotations, covering a number of levels of interpretation, we aim to develop approaches to apply similar labels to sentences in unseen data so as to reduce the effort required to interpret other blog posts related to COPD.

CHAPTER 6

EXTRACTING MEDICAL CONCEPTS – AUTOMATIC APPROACHES

The semantic information conveyed in patient blog posts is an important component of interpreting this data in the medical context. Semantic information includes the medical concepts discussed in patients' accounts of their experiences – e.g. symptoms, treatments, diagnoses, medical professionals, etc. As described in Section 2.2.2, Information Extraction has a number of applications, including the organisation and automated processing of medical information (Section 2.2.4.2). Firstly, identifying the concepts referred to in patients' accounts of illness can help filter, or search through a large collection to find those relevant to a particular research question. For example the Read codes, as discussed previously, allow medical professionals to organise large collections of medical notes. However, annotating concepts can also form the basis of further analysis – for instance the analysis of relationships between concepts, or more general topic classifications. Continuing with the Read code example, these standardised conceptual mappings enable population-scale enquiries of general medical record databases.

Automating the extraction of these concept mentions directly from text data can help to reduce the manual effort required to organise these document collections and to enable further processing of the data. Information Extraction approaches, such as the ones detailed in Section 2.2.4.2, can help identify concepts in text. Many Information Extraction systems, such as MedLEE (C. Friedman et al., 1994), identify concepts using structured, standardised vocabularies – sets of terms known to relate to domain-specific concepts. However, linguistic variation can mean that these vocabularies are difficult to curate and maintain. Concepts may be referred to by any number of phrases which can change over time. Similarly, the same term can refer to a number of concepts in different context. This can create noise in the output from systems based on standardised, static vocabularies. Automatic Term Recognition (ATR) tools use linguistic and/or statistical analysis to automatically identify potential terms from linguistic clues reducing the static nature of the identifiable terms providing more dynamic output.

In this chapter, we described two approaches to extracting medical terms from COPD patient blog posts. First, we introduce an approach to manually curate medical lexicons, based on existing structured, standardised knowledge sources and literature as well as the qualitative coding hierarchies created as part of our initial qualitative analysis (see Section 6.1.1). The resulting terminology represents both formal medical terms and informal colloquial expressions used by patients. Alongside this terminology, we developed a rule set to perform topic classifications at the sentence level (see Section 6.1.2). This will help annotate sentences in blog posts which are related to COPD exacerbation. Secondly, we applied FlexiTerm - a novel ATR approach for informal text sources - on our blog post dataset. FlexiTerm uses linguistic and statistical analysis alongside term normalisation and matching which accounts for linguistic variation. Both the lexicon and ATR approaches were evaluated against a manually annotated set of terms.

6.1 Constructing a COPD Lexicon and Rule Set for Topic Identification

Topic identification in patient blog posts involves identifying medical concepts involved in the authors' accounts. In order to achieve this, we must cover both explicit and implicit mentions. Our focus is on exacerbation of COPD – an acute worsening of a patient's symptoms. As mentioned previously, COPD exacerbation can be caused by many respiratory risk factors and treated in many different ways. The variation in how exacerbation may be referred to poses a challenge for this information retrieval task. In this section, we detail our work to manually curate a lexicon of terms relevant to COPD exacerbation and a rule set with which to identify exacerbation mentions. Terms are phrases used to represent concepts from within a particular domain (e.g. biology or medicine). By identifying the terms that can signify this more general concept *exacerbation*, we can build rules and tools to extract these statements automatically.

The nature of the topic classification task and, moreover, the informal nature of the corpus – the procedure used to build the lexicon involved sources from medical literature and existing medical terminologies were used as well as a reference sample of blog posts from our corpus. The National Institute for Health and Clinical Excellence (NICE) guidelines for COPD (NICE, 2004) were reviewed and its structure used to specify the core concepts and initial terminology. These guidelines cover definitions related to COPD as well as diagnosis, management and treatment guidelines. These core concepts and terms (e.g. *diagnosis*, *management*) were then expanded using existing medical terminologies and a reference set of posts from our collection of data in order to create a lexicon with a wide coverage of

terms and variants. Two General Practitioners who were involved with the project also advised to help guide development of the lexicon. To label sentence relating to exacerbation of COPD a rule set was produced using the reference set of 50 blog posts. The rules developed combine terms from our manually curated lexicon and linguistic features in order to label sentences as either related to exacerbations of COPD or not.

6.1.1 IDENTIFICATION OF CORE CONCEPTS

NICE are an independent body in the UK dedicated to producing evidence-based advice on best practise in health and social care in order to improve patient outcomes. As discussed in Chapter 3, COPD is a chronic condition and can affect many aspects of a patient's life and as such the guidelines published by NICE involved a multidisciplinary team to cover all aspects of diagnosis, management and care.

In order to begin to organise related terms in to some structured lexicon, we used the main themes in the guidelines. The resultant categories in to which we would place terms relating to COPD is as follows:

1. Diagnosis
    a. Symptoms
    b. Tests and Measurements
2. Management
3. Treatment
4. Exacerbation
5. Medical Professionals

Concepts were highlighted in the NICE guidelines and placed in to this general structure while grouping and categorising them further.

A collection of medical terminologies was then used to expand terms found. The Unified Medical Language System (UMLS) (NLM, 2012) combines and expands on existing medical terminologies and a wide-ranging medical semantic network is curated manually by the National Library of Medicine (NLM, USA). This resource helped expand terms through finding synonyms and related concepts. For instance, the NICE guidelines for treatment of COPD exacerbations caused by respiratory infection with antibiotics names several general classes of antibiotics (aminopenicillin, macrolides and tetracycline). Each general class contains many individual antibiotics and includes many brand names as well. The UMLS

allowed us to expand this to over 900 anti-biotic terms including around 250 brand names, providing much wider coverage of concepts relating to treatment.

The last stage was to take in to account informal references to medical concepts – highlighted through manual analysis of a reference set of COPD patient blog posts. Mentions of concepts were manually extracted and the terms used were expanded using online thesauri. How patients referred to their symptoms gives a good illustration of the how we were able to expand our lexicon. Productive coughs are often a sign of respiratory infection and informal references were common in the blog posts we analysed. Phlegm produced in the lungs was often referred to as 'gunk' or 'goo', expanded via the thesauri to cover terms like 'ooze' and 'sludge' to add further coverage. Including these colloquialisms from the online patient community, rather than relying solely on formal language sources, means that we can better identify the concepts involved in online discourse.

The resultant lexicon includes 3073 terms relating to COPD. The structure identified above was expanded and built upon, resulting in the following:

1. General COPD terms
2. Diagnosis
    a. Symptoms
        i. Respiratory
        ii. Other
    b. Tests and Measurements
        i. Devices
        ii. Measurements
        iii. Tests
3. Management
    a. Pulmonary Rehabilitation
    b. Smoking Cessation
    c. Surgery
    d. Vaccination
    e. Risk Management
4. Treatment
    a. Alpha-1 Antitrypsin Replacement Therapy
    b. Antibiotics
    c. Antioxidents

    d. Bronchodilators
    e. Mucolytics
    f. Oxygen
    g. Steroids
    h. Delivery Devices
5. Exacerbation
    a. Causes
6. Medical Professionals

This vocabulary represents terms relating COPD exacerbation, including standard terms and their expanded forms – i.e. synonymous expressions gathered from a number of formal and informal sources. Utilising this vocabulary, we can help automatically extract concept mentions in text.

6.1.2 Classifying sentences relating to patient experience of COPD exacerbation

Our aim here was to build a ruleset based on a custom lexicon in order to automatically label sentences relating to the author's personal experience of COPD exacerbation. A lexicon of COPD-specific terminology will help identify the concepts patients discuss. By utilising a ruleset to extract relationships between these concepts, we aim to identify sentences which specifically relate to the exacerbation of a patient's condition. As already described, understanding how patients recognise and manage exacerbation is an important step to supporting them in managing their condition. Automatically extracting this information from large-scale online discussions has the potential to help build a better picture of patient experience.

Topic identification was approached using manually curated rules. The set of 100 blog posts annotated online was split into two sets of 50 documents, a reference and test set (1,199 and 1,010 sentences respectively). Both the test and reference sets were manually annotated for exacerbation experience mentions. Each sentence in these posts was labelled either positive or negative for being related to exacerbation experiences or not. The reference set was then used to develop a set of 52 rules by which these sentences could be extracted, utilising the information from the lexicon and subjectivity classifications already obtained. The rules were developed using a rule-based information extraction tool called Mixup (part of the MinorThird library (Carnegie Mellon University, 2009)). Example rules are given below:

```
defDict oxygenWords = oxygen, o2;
defDict nose = nose,noses,nasal,nostril,nostrils;
defDict cannula = cannula, cannulae, tube, tubes;
defDict therapy = therapy, treatment,supplement,supplemental;
defDict preMaskWords = face,venturi;
defDict o2exclude = saturation,level,levels,sats,sat,exchange,intake;

defSpanType nasal =sentence:
       ... [ ai(nose) ] ...;

defSpanType o2delivery =sentence:
       ... [ @nasal any{0,2} ai(cannula) ] ...|| //e.g. 'nasal cannula'
       ... [ ai(cannula) any{0,2} @nasal ] ...|| // e.g. 'tubes  in my nose'
       ... [ ai(oxygenWords) ai(preMaskWords){0,1} eqi('mask') ] ...||
       ... [ ai(preMaskWords) eqi('mask') ] ...; // 'face mask'
```

This set of rules aims to identify mentions of oxygen therapy, utilising 6 dictionaries of relevant words and two rules. The rule *o2delivery* combines words in the dictionaries provided (`ai(<dict>)` denotes a case insensitive match with a word in the specified dictionary) along with other conditions to denote a match. This rule contains 4 patterns (|| signifies an *or* relationship between rules, meaning that the label will be applied if any of the following patterns are a match) and should any of the conditions satisfy for any span of text, then the label will only be applied to the tokens matching the specific pattern between the brackets. The complete rule and dictionary set can be found in Appendix D.

		Manual	
Automatic		Exacerbation	Non-exacerbation
	Exacerbation	86	72
	Non-exacerbation	64	788

Table 7: Exacerbation sentence classification confusion matrix

The annotations were assessed in terms of precision, recall and f-measure (*P*, *R* and *F* respectively) based on the number of true positives (*TP*), false positives (*FP*) and false negatives (*FN*) see Equations 8-10.

$$P = \frac{TP}{TP + FP} \quad \textbf{1}$$

$$R = \frac{TP}{TP + FN} \quad \textbf{2}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad \textbf{3}$$

The precision, recall and F-measure (0.54, 0.57 and 0.55 respectively) show a relatively precise approach to discovering exacerbation sentences, but more work is needed to improve the rule set.

We also applied an inter-annotator agreement measure – a chance corrected measure of performance. By comparing agreements between multiple sets of labels to what would be expected by chance (i.e. picking a label at random) we can understand how our approach performed within this particular dataset. There are a range of measures, (e.g. Cohen's kappa (Cohen, 1960), Scott's pi (Scott, 1955), Fleiss' kappa (Fleiss, 1971) and Krippendorff's alpha (K. Krippendorff, 1970)) and each makes different assumptions about the data making them applicable in different situations. Cohen's Kappa rates the agreement between two annotators who have labeled each available example. Cohen's Kappa measures the observed relative agreement among annotators (P_o) against the hypothetical probability of agreement by chance (P_e).

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad \textbf{4}$$

In our case, we can compare the automatic labels to the manual ones to understand the strength of our classification method. The agreement (κ=0.477) achieved was better than would be expected by random assertions, but less than perfect. This initial evaluation, while promising, shows that this rule set would need to be expanded further to improve coverage and precision of classifications. In order to facilitate the expansion of this rule set, automatic rule generation could be used to develop a rule set based on training examples. Similarly other, non rule-based classification approaches could be used in order to facilitate the automatic annotation of this data.

As introduced previously, one limitation of the lexicon based Information Extraction approach is the brittle nature of the knowledge source. While some variations are accounted for in the lexicon, language can be very expressive, and some variations on themes may be missed. In the next section, we discuss the application of Automatic Term Recognition to provide wider coverage of semantic information extraction.

6.2 TERM RECOGNITION

While standardised lexicons are useful in identifying known concepts from text, variations in the way that concepts are represented in text can mean information is missed. Term recognition is the process of marking which words or phrases in text relate to some concept (Krauthammer & Nenadic, 2004) and can be carried out manually (see Section 6.1.1) or automatically. Term recognition carried out manually relies on human indexers using a controlled vocabulary or dictionary (Salton & Mcgill, 1986) or on domain experts. However, manual term recognition is time-consuming and expensive and so Automatic Term Recognition (ATR) is used to process electronic documents more efficiently. Relying on dictionaries is also problematic as domains and their vocabularies are dynamic (Nenadić, Spasić, & Ananiadou, 2002). Likewise, new channels of communication, such as social media, mean that lay users may discuss specialised domains without using standardised language, so standardised vocabularies may not be suitable for successfully extracting non-standard concept mentions.

While manual approaches to term recognition may rely upon the domain knowledge of the human performing the task, automated approaches rely on other forms of information to annotate terms – for instance linguistic (i.e. morphological, syntactic or semantic) or statistical information. Linguistic ATR approaches rely on pattern-matching to locate candidate terms based on their syntactic properties (Justeson & Katz, 2008). Statistical approaches rely on information about frequency in order to compute a phrase's *termhood* or *unithood.* Unithood relates to the strength of collocations of constituent words, characters or phrases, where termhood refers to the degree to which a phrase relates to a domain-specific concept (Kageura & Umino, 1996). Collocation measures such as mutual information can be used to compute a unithood score – i.e. the relatedness of particular words or the likelihood of them appearing together (Church & Hanks, 1990). One example measure of termhood is the C-value score (Frantzi, Ananiadou, & Mima, 2000). The C-value method extracts term candidate using a pattern based approach. The procedure is as follows:

1. Part-of-speech (POS) tagging
2. Linguistic filtering
3. Stop list filtering
4. Termhood computing

POS tagging provides the syntactic information used in the linguistic filtering where pattern-matching is applied. The role of these patterns is to extract only those words sequences that conform to syntactic rules that describe a typical inner structure of terms. In the statistical part of the C-value method, each term candidate is quantified by its termhood following the idea of a cost-criteria based measure originally introduced for automatic collocation extraction (Kita, Kato, Omoto, & Yano, 1994). The termhood calculation is based on a number of statistical criteria:

$\|a\|$	The length of the candidate string (number of words)
$f(a)$	Frequency of the candidate term *a* in the corpus
T_a	Set of candidates that *a* is nested in
$\sum_{b \in T_a} f(b)$	Frequency of the candidate as part of another, longer candidate term

The principles underlying C-value are that longer, more frequent candidates are more related to the corpus' domain, but penalties are applied for candidates which appear in longer terms – i.e. there are more specialised uses of the term. C-value is computed as follows:

$$\text{C-value}(a) = \begin{cases} \log_2|a| \cdot f(a) & a \text{ is not nested} \\ \log_2|a| \cdot \left(f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b) \right) & \text{otherwise} \end{cases} \quad \textbf{5}$$

While unithood and termhood are useful if term occurrences are consistent, terms are subject to several forms of variation which pose a challenge:

- morphological variation, constituent words are inflected (e.g. '*COPD exacerbation*' and '*exacerbated COPD')*
- syntactic variation, where constituent words appear in their original form (e.g. *chest pain* vs. *pain in chest*)

- semantic variation, where there is a semantic relation between constituent words (e.g. *pulmonary function* vs. *lung function*)

Typographical and spelling errors along with regional differences (e.g. *haemoglobin* vs. *hemoglobin*) also cause variation in term representation. If minor variations on terms are considered new term candidates termhood and unithood scores may be inaccurate and the concepts related to a corpora bloated.

FlexiTerm (Spasić et al., 2013) is a novel approach to ATR which addresses the challenge that terminological variation poses. Like C-Value, FlexiTerm uses pattern-matching for term candidate extraction and then computation of a termhood score based on the frequency of the candidate in the corpus. However, FlexiTerm also relies on two normalisation steps in order to match term variants which relate to the same concept. Firstly, term candidate normalisation employs stemming and a bag-of-words term representation to overcome the problems caused by word order and inflection as seen in the morphological and syntactic variations described above. Secondly, flexible token-level matching based on edit distance is employed to match further token variations which may be missed by stemming – for instance, spelling or typography errors. FlexiTerm is described in detail in the remainder of this section.

Term candidate normalisation

Applying a normalised format to term candidates enables matching between them, despite syntactic and morphological variation. This normalisation process is similar to that presented by McCray et al. (1994) which consists of the following steps:

1. Remove punctuation, numbers, and stop words
2. Remove lowercase tokens containing 2 or fewer characters
3. Stem remaining tokens
4. Sort remaining tokens in alphabetical order

Stemming tokens in step 3 accounts for morphological variation where inflection is used, while steps 1, 2 and 4 address the problems of syntactic variation, where the same tokens appear but not necessarily in the same order or with added prepositions. While the McCray et al. method required token sorting, we relied on a *bag-of-words* approach, where word order is disregarded altogether. The result of this process is that similar candidates such as '*airway inflammation*' and '*inflammated airways*' will both be matched to the normalised form {*airwai*, *inflamm*}. This enables information about each candidate to be

aggregated, and the concept they represent to be more adequately represented in the output of the ATR process.

Token-level matching

While stemming deals with token variations caused by inflection, other variations we have discussed can cause issues with exact string comparisons. Variations caused by spelling and typographical errors as well as some regional differences are addressed using a flexible string matching approach. To support flexible token matching, FlexiTerm uses an open-source spellchecking API called Jazzy (Jazzy, 2013). Jazzy is based on the Aspell (Atkinson, 2011) algorithm, combining both string comparison and phonetic token matching.

The similarity of two strings (i.e. character sequences) can be represented by their edit distance (Damerau, 1964). Edit distance (ED) assigns operations to transform one token to another (e.g. substitution of one character for another, deletion of a character or insertion of a new character) a cost. The lower the cost to transform one string to another, the better matched the strings are. For example, comparing 'exacerbate' and 'exasperate', the edit distance would be three, as there are three operations to convert one string to another:

exacerbate	exasperate	
exacerbate →	exaperate	delete 's'
exa**p**erate →	exa**c**erate	replace 'p' with 'c'
exacerate →	exacer**b**ate	insert 'b'

Metaphone (Philips, 1990) is an algorithm which indexes tokens using an approximation of their English. Characters or combinations of characters are mapped to a standardised set of 16 characters based on a generalisation of how they would be pronounced. Maintaining the examples of 'exacerbate' and 'exasperate' – both these words sounds similar when pronounced. Using Metaphone, their pronunciation would be represented as the string 'EKSS'. The fact that both share the same string representation is due to the fact that they are formed using similar sounds. For comparison, the word 'pronounce' is represented by the string 'PRNN'. Normalising strings in this fashion means that strings that are phonetically similar, but lexically different can be found. Jazzy uses both the ED and phonetic similarity of strings when generating lists of spelling suggestions and it was this

that we used to map similar strings together. The process used to find suggestions is as follows (White, 2004).

Calculating termhood

FlexiTerm's termhood calculation is based on *unithood* and *termhood* as described earlier and applied in the C-Value ATR. Nestedness (the set $T(a)$ in Equation 5 which represents the set of candidates in which a appears nested) within FlexiTerm is not limited to the candidate exactly appearing as part of a longer candidate. The outcome of the normalisation and token-level similarity portions of the FlexiTerm approach gives a bag of expanded words which represents a fuzzy profile of the candidate, allowing the set of potentially similar candidates to be found. It is this set that is used in place of $T(a)$. In summary, the FlexiTerm approach is as follows (Spasić et al., 2013):

1. Pre-process text to annotate it with lexico-syntactic information.
2. Select term candidates using pattern matching based on part-of-speech
3. Normalise term candidates by performing the following steps.
    a. Remove punctuation, numbers and stop words.
    b. Remove any lowercase tokens with ≤2 characters.
    c. Stem each remaining token.
4. Extract distinct token stems from normalised term candidates.
5. Compare token stems using lexical and phonetic similarity calculated with Jazzy API.
6. Expand normalised term candidates by adding similar token stems determined in step 5.
7. For each normalised term candidate t:
    a. Determine set S(t) of all normalised term
8. Candidates that contain t as a subset.
    b. Calculate C-value(t) according to formula 5.
9. Rank normalised term candidates using their C-value.

FlexiTerm was evaluated on 5 corpora – a mix of formal, edited text (i.e. biological and medical literature) and unedited text (e.g. 2 groups of medical notes) including blog posts from the collection detailed in Chapter 3. The manual annotations of term occurrences in the posts covered in section 3.3.1.4 were used as a gold standard (the same procedure was used for the other corpora) and the results were compared to C-Value as a baseline. The term lists produced by FlexiTerm and C-Value on this corpus were assessed in terms of

precision, recall and F-measure (see Equations 1-3). The results from FlexiTerm were comparable to those achieved by C-Value, achieving the best performance gain in the unedited text (medical notes and patient blog posts) where terminological variation is more likely. The results achieved on the blog posts corpus are summarised in Figure 15 - Figure 17 and the top 10 terms identified by the two methods can be found in Table 8.

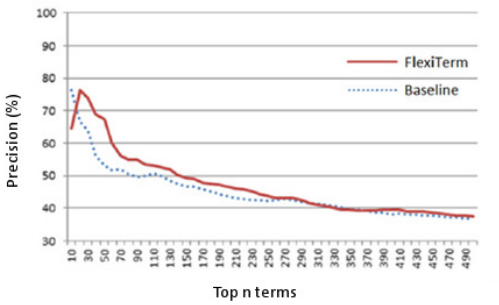


Figure 15: Flexiterm precision – COPD patient blog posts

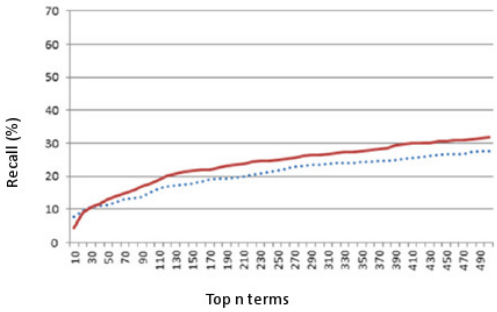


Figure 16: Flexiterm recall – COPD patient blog posts

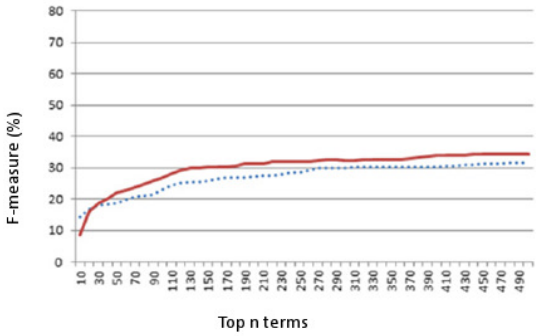


Figure 17: Flexiterm F-measure - COPD patient blog posts

Figure 15-Figure 17 show the precision, recall and f-measure of the FlexiTerm approach against the baseline method. As both approaches rank relevant terms according to their importance, we used various numbers of proposed terms (i.e. the top 10, 20, 30, etc.) and analysed the performance of each method in slices. We would expect that more of the manually identified domain-specific terms would appear higher in the ranked term list. As shown in these graphs, FlexiTerm improves performance of term extraction on this medical social media dataset over the baseline method.

Rank	FlexiTerm	TerMine
1	pulmonary rehab	pulmonary rehab
	pulmanary rehab	
2	breathe easy	breathe easy
3	vitamin d	vitamin d
4	lung transplantation	lung function
	lung transplant	
	lung transplants	
	lung transplantations	
5	breathe easy groups	severe copd
	breath easy groups	
	breathe easy group	
6	chest infection	blood pressure
	chest infections	
7	quality of life	lung disease
8	blood pressure	lung transplant
9	lung function	chest infection
10	rehab room	rehab room

Table 8: Comparison of terms identified by Flexiterm and baseline method

The terms identified by both approaches (Table 8) are similar, but FlexiTerm has identified variants and successfully grouped them together (e.g. 'lung transplant', 'lung transplantation', etc.) which has improved their termhood scores. The terms extracted help inform us about the main topics discussed in the blog posts collected.

Each term relates to some medical concept, despite not being gathered from some standardized vocabulary – the term candidates, identified from their linguistic features, and their variants have been identified as important to this document collection based on

statistical measures of termhood. That mentions of medical concepts can be automatically extracted and found, providing a number of benefits to processing large collections of patient blog posts. Firstly, researchers can navigate through the document collection more easily, by looking at particular areas related to medical concepts that are under investigation (i.e. indexing for Information Retrieval). Secondly, these annotations can relate to coding hierarchies (such as the one developed in Chapter 5) and help to identify potential themes. Lastly, patterns relating to these concepts, and where they appear in text, can be mined in order to generate new facts or information about patient experiences. Relationships between these medical concepts and correlations with other data that can help identify new hypotheses and drive further investigation. These applications of the data are discussed further in Chapter 8.

6.3 Conclusion

In this chapter, we described two approaches to extracting semantic information from COPD patient blog posts. The first is through utilising standardised medical terminologies alongside the qualitative analysis (Chapter 5) to build a lexicon of topic-specific medical terms. We also demonstrated using pattern-based information extraction to add further information, relating to symptoms, treatments and medical professionals related to our area of inquiry (COPD exacerbation). Secondly, we introduced our approach to automated term recognition, based on linguistic features requiring no manual curation. FlexiTerm is an ATR approach which aims to perform well in more informal text sources – where terms may be subject to more linguistic variation and synonymy.

This chapter details some important contributions of our work. Firstly, the gold standard annotations added by independent annotators add value to the blog post corpus for reuse in subsequent experimentation with semantic information extraction methods. Secondly, the two approaches to semantic analysis and utilising the outcome of qualitative analysis could be replicated for other, non-COPD related, inquiries.

Unlocking semantic information contained in patient blog posts is an important step to automating their interpretation – recognising the medical concepts involved in patients' descriptions of their experiences. In chapter 8 we investigate one way that we could use semantic analysis to summaries trends in blog post content. Besides being used for further processing, the semantic information added to the blog post text in this phase has the potential to help medical researchers explore data sets – exploring topics or trends within

the data. In Chapter 8 we also look at ways that this information could be exported to existing qualitative research software packages in order to support further research.

CHAPTER 7

SENTIMENT MINING

While the concepts and topics extracted from textual content can help us discover *what* is described, sentiment analysis can help unlock the psychological context – the authors' feelings towards what they are describing and how that impacted them. Linguistic clues, such as emotionally charged words can help classify the kinds of emotion expressed through statements. This can be a binary classification (i.e. positive/negative sentiment) or the specific *kinds* of emotion shown, usually expressed as a node in an emotional taxonomy.

The sentiment information expressed through patient blog posts are an important facet of the patients' narrative. As well as the symptoms, conditions and treatments they describe, the emotional context in which they appear could be an important component of highlighting patients' priorities, needs or concerns. As well as the emotional information, we introduced subjectivity classification as a component of sentiment analysis in Section 2.2.5. Subjectivity analysis aims to classify a sentence as related to the author's own personal opinion or experience as opposed to some objective fact about the entity under discussion. This step also has a great potential in automatically interpreting patients' experience – helping extract the sentences that most explicitly summarise their experience.

As mentioned in Chapter 2, sentiment analysis deals with classifying very subjective interpretations of textual data and poses a great challenge in terms of creating a gold standard set of annotations with which to train and evaluate models. In this chapter, we describe our approach to annotating our blog dataset with sentiment information. We utilise a crowdsourcing approach – a distributed online task in which many people can take part. We gathered multiple interpretations of each sentence, from across stakeholder groups in order to create a set of annotations and reduced the inherent bias. We also discuss our approach to evaluating this dataset with measures of inter-annotator agreement and utilising the dataset alongside probabilistic machine learning approaches to

train and evaluate models to automatically label sentiment information in patient blog posts.

7.1 CROWDSOURCED DATA ANNOTATIONS

In order to add semantic annotations to the dataset, we used crowdsourcing to collect multiple annotations from a wide range of people over the Internet. Crowdsourcing is a method of problem solving or content creation through distributed participation of many individuals following an open call (Brabham, 2008). The use of the Internet as a crowdsourcing medium bridges the physical gap and allows a large number of participants to be reached.

We implemented a web-based annotation tool[13] to display 20 random sentences from the dataset to each annotator during a single session, along with the context it appeared in originally in order to aid interpretation of the indented meaning (see Figure 18).

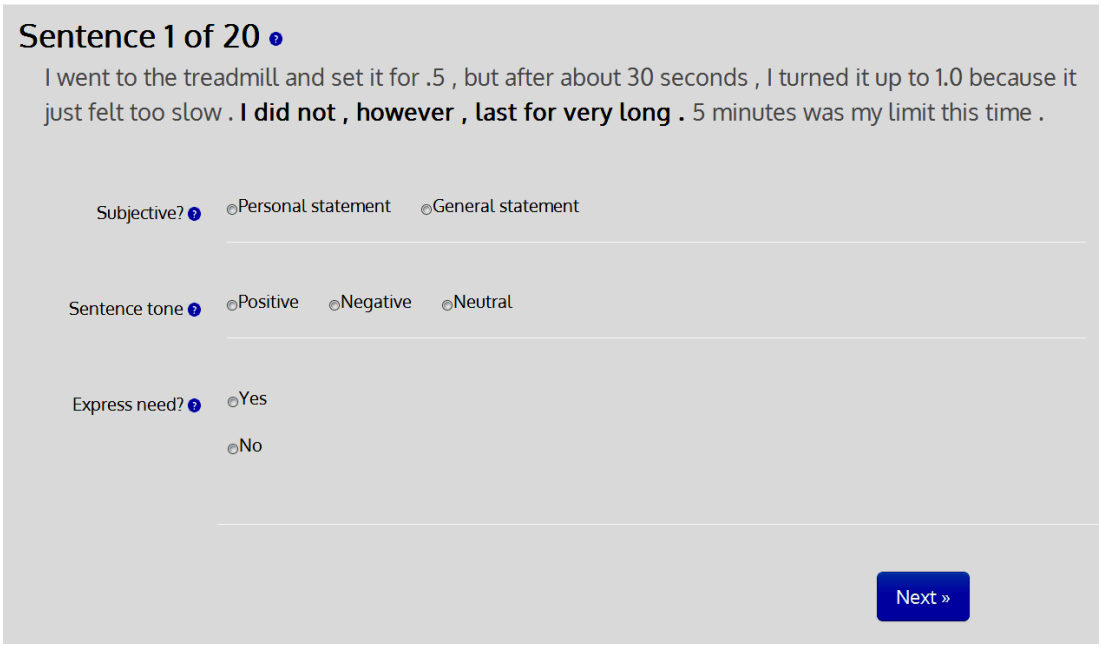


Figure 18: Annotation site screenshot

A random subset of 100 blog posts was selected (from a total of 368), and the sentences from those documents which contained more than 10 tokens were selected for annotation (1770 sentences from a total of 2209). Each annotator was asked to annotate the sentence in terms of three criteria, described on an initial instructions page. The descriptions are as follows:

[13] http://users.cs.cf.ac.uk/M.A.Greenwood/annotation

Question 1:	Subjective?	Is the sentence about something **personal** (e.g. a personal experience or opinion) or something more **general** (e.g. sharing information)
Question 2:	Sentence tone	What is the general tone of the sentence? • **Positive** – happiness, joy, excitement, etc. • **Negative** – sadness, anxiety, anger, etc. • **Neutral** – neither positive nor negative
Question 3:	Express need?	Does the patient express an **unmet need** within this sentence? Do they mention something that, in their opinion, could... • ...**provide extra assistance** to patients? • ...be **deemed necessary** for better care? • …be **improved**?

Table 9: Annotation exercise questions and definitions

A subsequent portion of Question 3 asked annotators to highlight portions of the sentence which indicated that a need was present. An instruction video[14] was provided on the instructions page to indicate how this feature could be used. Further introduction of each question and the results obtained are set out in Sections 7.1.1-7.1.4.

[14] http://www.youtube.com/watch?v=uTQcdiFABas

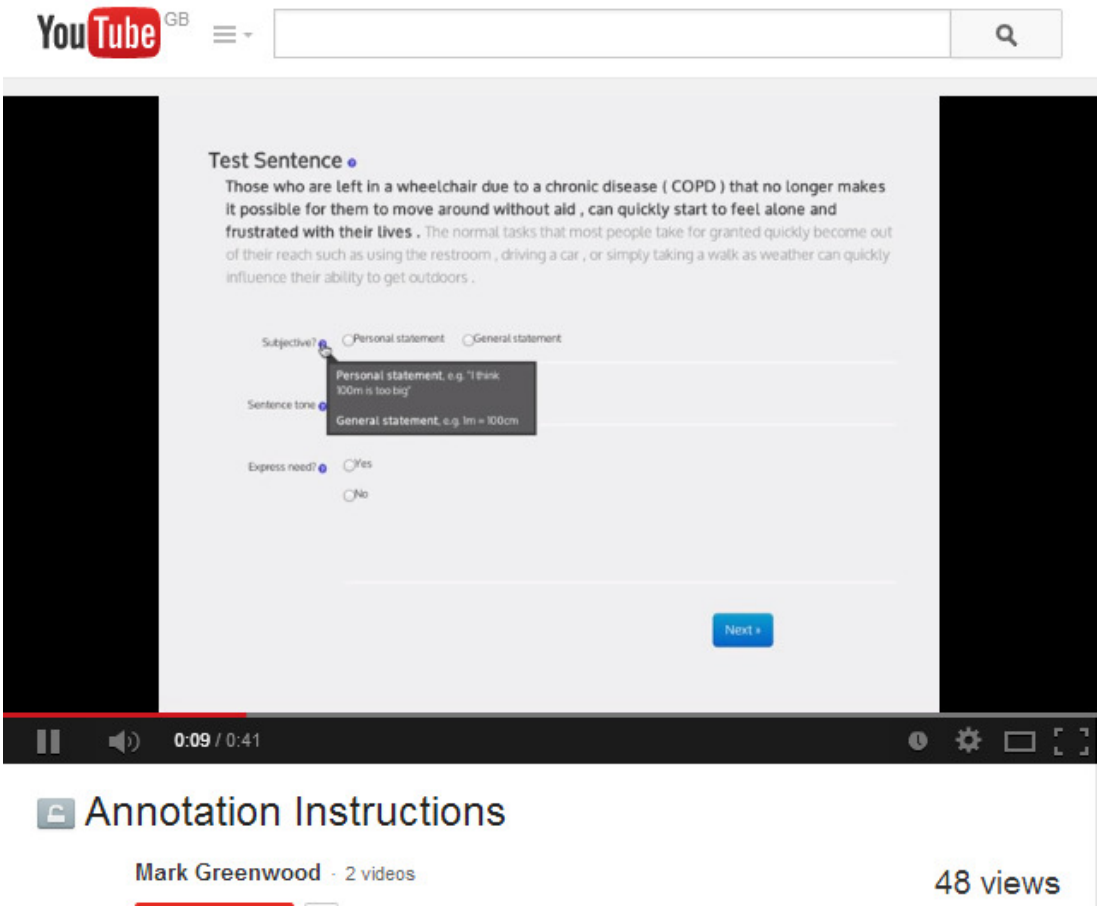


Figure 19: Annotation instruction video screenshot - http://www.youtube.com/watch?v=uTQcdiFABas

An open call for participation enabled us to collect opinions from various stakeholder groups. While the interpretation of the sentences should not require any specialist knowledge, users who are close to the subject matter (e.g. patients or healthcare professionals) may have a different interpretation. Aiming for a spread of interpretations means that interpretation bias could be reduced. In order to represent the spread this information was captured in an initial question posed to each user before the first sentence was displayed. The annotator was asked which category best described their relation to the subject-matter of the task (i.e. COPD) – the following categories were offered:

- Patient
- Carer (i.e. carer for a friend or relative who is ill)
- Healthcare professional (e.g. doctor, nurse, etc.)
- Health researcher
- Researcher
- Student
- Other

The annotator could also select an 'I would rather not say' option if they wanted to complete the task with greater anonymity. An open call may introduce bias if certain groups of participants are reached while others are not. We collected annotations in three phases:

1. Small scale pilot (12 participants) validating online exercise.
2. Call to participate through mailing lists, department newsletters and social networking sites (Facebook, Twitter).
3. Repeat invitations, plus recruitment through Amazon Mechanical Turk crowdsourcing platform[15].

While representation of various interpretations is important, anonymous data collection online is prone to gaming and malicious behaviour, which increases the probability of poor quality annotations (Raykar et al., 2010; Vuurens & de Vries, 2012; Zhao & Zhu, 2012). In order to address this risk, each sentence was annotated by more than one user. Firstly, this allowed us to assess the quality of the annotations using inter-annotator agreement. Secondly, applying a majority vote method, we could select the annotations with the highest agreement to later train and test our method.

Krippendorff's alpha was used to rate inter-annotator agreement. As mentioned in Section 6.1, different inter-annotator agreement measures are applicable in different situations. Where Cohen's Kappa was useful for comparing the resultant labels generated by two annotators, Krippendorff's alpha can be used to measure agreement among any number of annotators, and allows for missing data, i.e. the case when not all annotators annotated each data item. Crowdsourcing encourages a large task (in our case, annotating a large number of sentences) to be divided in to smaller parts and to encourage a large number of participants to join in. By distributed completion of these smaller tasks, the larger goal can be achieved. Our task is designed in this way, and as such, Krippendorff's alpha is the most appropriate measure for inter-annotator agreement.

Krippendorff's alpha is calculated using the ratio between the observed disagreement between annotators and that which would be expected by chance (see Equation 6). The disagreement is measured using incidence matrices (of the form described in Figure 20) where the agreement between observers 1 to *m* on units 1 to *N* are mapped. Equations 7, 8 and 9 are then used to compute the disagreement between each pair of annotators and

[15] http://www.mturk.com

the agreement that would be expected by chance. Equation 6 is then used to combine these measures in to one overall agreement measure between -1 (complete disagreement) and 1 (perfect agreement) with 0 meaning an agreement which could occur by chance (Klaus Krippendorff, 2013).

Units ***u:***	**1**	**2**	.	.	.	***u***	.	.	.	.	.	***N***
Observers: **1**	c_{11}	c_{12}	.	.	.	c_{1u}	.	.	.	.	.	c_{1N}
i	c_{i1}	c_{i2}	.	.	.	c_{iu}	.	.	.	.	.	c_{iN}
j	c_{j1}	c_{j2}	.	.	.	c_{ju}	.	.	.	.	.	c_{jN}
.	.	.				.						.
m	c_{m1}	c_{m2}	.	.	.	c_{mu}	.	.	.	.	.	c_{mN}
Number of observers valuing ***u:***	m_1	m_2	.	.	.	m_u	.	.	.	.	.	m_N

Figure 20: Incidence matrix form for computing Krippendorff's Alpha – adapted from (Klaus Krippendorff, 2013)

$$\alpha_{\text{nominal}} = 1 - \frac{D_o}{D_e} \quad \textbf{6}$$

$$D_o = (n-1)\sum_c o_{cc} - \sum_c n_c(n_c - 1) \quad \textbf{7}$$

$$D_e = n(n-1) - \sum_c n_c(n_c - 1) \quad \textbf{8}$$

$$o_{ck} = \sum_u \frac{Number\ of\ c-k\ pairs\ in\ unit\ u}{\mathrm{m_u} - 1} \quad \textbf{9}$$

We computed Krippendorff's alpha using Thomas Lippincott's (2008) publicly available 'agreement.py' script which has since become part of the Natural Language Toolkit (NLTK)[16] Python[17] library.

Crowdsourcing task results

The crowdsourcing approach was successful in collecting annotations for a total of 1770 sentences. Table 10 shows details of the annotators along with the properties of the annotations collected. Figure 21 shows the number of annotators involved in each phase of collection. Lastly, Figure 22 shows the proportion of annotators reporting each option in

[16] http://www.nltk.org/
[17] https://www.python.org/

the initial stake question, showing that a range of stakeholders were involved in labelling the data.

Blog posts	100 (233KB)
Sentences with >10 tokens	1770
Annotators	286
Completers (annotated all 20 sentences)	226
Total annotations	4745
Average annotations per sentence	2.68 (std dev: 0.5)

Table 10: Annotations properties

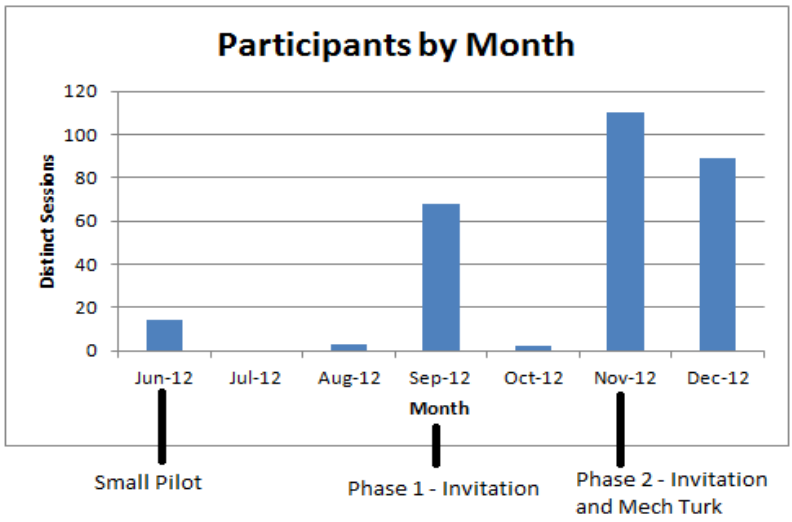


Figure 21: Crowdsourced annotation participants by month

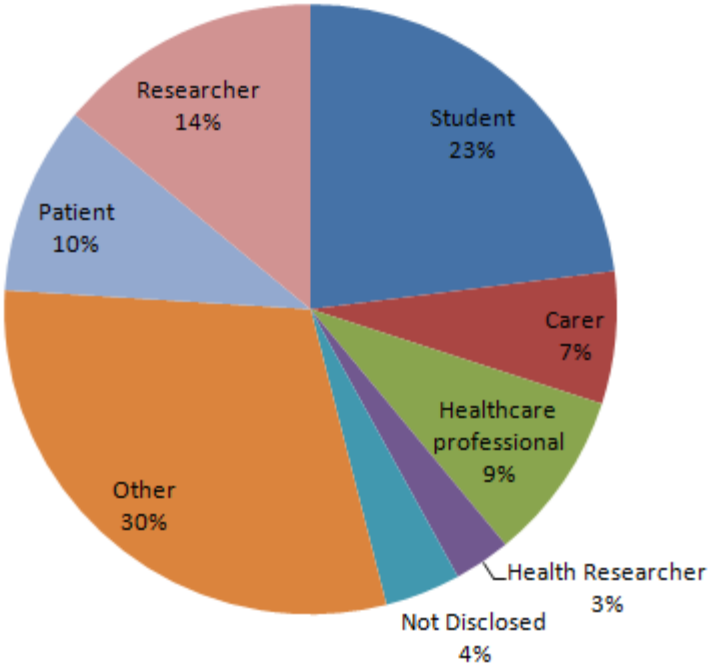


Figure 22: Annotators self-reported stake

In total, 286 annotators were involved in our annotation gathering exercise. We were successful in collecting multiple annotations per sentence, with each sentence gaining at least 2 annotations relating to the three questions. Figure 21 shows that phase 2 (utilising open invites on crowdsourcing platforms) succeeded in attracting most people to take part. Figure 22 shows a good spread of stakeholders taking part in the annotation task, thereby including many sources of opinion in our dataset. In sections 7.1.1 - 7.1.4 the annotations pertaining to each of the three questions will be analysed in more details and the results of the collection task set out.

7.1.1 Question 1: Subjectivity of statements

Question 1 of the annotation task aimed to determine whether the sentence pertained to some personal experience of the author, or to more general information. Ultimately, we aimed to annotate the data according to whether the author was talking about themselves or those close to them or sharing more general information about their condition – e.g. facts or advice.

The agreement amongst annotators for this question is set out in Table 11.

Agreement	Total	%	Label	Sentences
No agreement	135	7.63%	N/A	
Maj. agreement (2/3, 3/4)	388	21.92%	Subjective	198 (50.1%)
			Objective	190 (48.9%)
Total agreement	1247	70.45%	Subjective	697 (55.9%)
			Objective	550 (44.1%)

Table 11: Sentences by annotation agreement - subjectivity

There was total agreement between annotators for the majority of sentences in this task (1247, 70%). However, as mentioned previously using percentages for agreement can skew our perceptions as they disregard chance. Krippendorff's alpha was computed to be α=0.55 for this dataset, which implies less than perfect agreement, but better than that expected purely by chance (α=0). The observed disagreement between annotators is not necessarily a consequence of poor quality annotations, but indicative of the difficulty in dealing with subjective interpretation. While the agreement score gives us an indication of confidence in the annotations, it can also provide a benchmark to evaluate automated approaches to data annotation – i.e. performing *as well as* human annotators in a highly subjective interpretation task.

7.1.2 Question 2: Emotional tone of statements

Question 2 asked annotators to determine whether the sentence portrayed a positive (e.g. happy, excited, etc.) or negative (e.g. angry, sad, upset, etc.) sentiment, including a neutral option when no sentiment was expressed. The agreement between annotators for this question is set out in Table 12.

Agreement	Total	%	Label	Sentences
No agreement (1/2, 1/3, 0)	358	20.23%	N/A	
Maj. agreement (2/3, 3/4)	633	35.76%	Positive	208 (32.9%)
			Neutral	315 (49.8%)
			Negative	110 (17.4%)
Total agreement	779	44.01%	Positive	325 (41.7%)
			Neutral	294 (37.7%)
			Negative	160 (20.5%)

Table 12: Sentences by annotation agreement - emotional tone

As Table 12 shows, there was little agreement amongst annotators about the emotional tone of sentences in the dataset – only 44% had total agreement between annotators, meaning there were disagreements on roughly half of the data items. Krippendorff's alpha was α=0.32, agreement better than that expected by chance (α=0), but not strong agreement.

The failure to draw agreement between annotators for this portion of the task could be due to a number of factors. Firstly, as there were three options for this question and between 2 and 4 opinions, there may not have been enough annotations per sentence in order to draw stronger agreement. Also, the definition and the way the question was posed may have been confusing for the annotators. Analysing the disagreements, there were often explainable disagreements. While some annotated sentences that were, for instance, introducing a new treatment that the author had just read about as neutral (I.e., according to the task, sharing information) others marked this as 'positive' – implying the author was sharing good news.

7.1.3 Question 3: Expression of patient needs

Question 3 asked annotators whether the sentence expressed something that the patient needed. This was defined in the exercise as something that the author thought could be provided to better support them (see Table 9). This question aimed to directly answer one of the main facets of our research by asking annotators to point out the patient needs within blog posts. However, two main challenges made this information particularly difficult to extract. Firstly, the question may be ambiguous without a more in-depth definition than could be provided through an open online exercise. 'Need' is still a largely

ambiguous concept – medical literature has still not firmly defined it – and can vary depending on the context (see Chapter 1). Secondly, there is potential for blogs being sparsely populated with explicit mentions of needs which could be labelled by an (in many cases, as was the aim) lay annotator. Despite the data sparsity problem the percentage agreement between annotators looks promising (see Table 13).

Agreement	Total	%	Label	Sentences
No agreement	170	9.60%	N/A	
Majority agreement (66%, 75%)	448	25.31%	Need	120 (26.8%)
			No need	328 (73.2%)
Total agreement	1152	65.08%	Need	54 (4.7%)
			No need	1098 (95.3%)

Table 13: Sentences by annotation agreement - patient need

However, Krippendorff's alpha (α=0.14) shows that much of the agreement can be explained by chance. This is because the majority of sentences were annotated unanimously (1098 out of 1770) or with at least a majority vote (1426 out of 1770) as not expressing a patient need.

7.1.4 Summary of Sentiment Data

An interpretation of the sentiment an author was expressing through a particular sentence is highly subjective. In utilizing crowdsourcing to annotate our blog post dataset with sentiment information, we aimed to take in to account a number of different opinions, spanning the stakeholder groups involved. Our 3 questions asked about different aspects of sentiment – the subjectivity of a statement, the nature of the opinions or emotions expressed (whether they were positive, negative or neutral) and lastly, the highly specific interpretation of whether or not a sentence expressed a 'need'. The first two questions are quite general, where the last one is specific to our particular dataset and research aim – to discover accounts of patient experiences.

Involving more than one annotator in labeling each sentence means that bias towards a particular participants view point can be reduced. The data were collected in a number of phases, where participants were recruited from within Cardiff University, through social media and through general purpose crowdsourcing platforms. In order to assess the data produced (i.e. the trustworthiness of the labels obtained) we used Krippendorff's alpha measure of inter-annotator agreement. We chose this method because it was a suitable measure given the circumstances under which the data was produced – where many annotators labeled many, but not all sentences in the collection. This measure assesses the labels assigned to each items in terms of whether agreement as to the label assigned could be attributed to chance. It can be used to assess whether, when annotators agree on a label they did so because they interpreted the sentence in the same way as opposed to, for instance, picking the same answer at random.

All but one of the questions we posed of our data resulted in inconclusive data. For both the question on emotional tone and relating to patient needs the computed alpha score was too similar to random to be able to say with any confidence that the labels on which annotators agreed were usable. In the case of emotional tone, this may be due to the number of opinions captured about each answer. This question differed from the others in that it has more potential responses. In order to say whether answers given were any better than chance, more responses to this particular question would be needed. The question of whether a sentence expressed some patient need may have suffered for a number of reasons. Firstly, the concept of a patient need may be too abstract or require too much knowledge of the particular medical domain to reasonably ask participants to pick out of the textual data. Also, occurrence of such a specific type of sentiment may be quite rare in the dataset, which means that the results in a general collection will always resemble the chance-case.

Improving the number of participants would be difficult, but the response from the crowdsourcing platforms was promising. One way to increase the number of opinions on each sentence would be to tune the offering on this platform. To offer greater reward for taking part and potentially a more straight-forward, easily explained task. In order to collect 'need' labels, a more focused task might succeed where this one failed. Focusing on one, well described concept and targeting expert communities may be one approach. Pruning the dataset to a more even distribution of need and non-need sentences may also

help improve the quality of the results and make interpretation of the inter-annotator agreement more useful.

The results of the subjectivity question, however, were promising with an alpha value of 0.55. As described earlier, the subjectivity of statements is of particular importance to our chosen research aim – highlighting patient experience. Identifying personal statements in the text means that we can more easily find those related to some experience or opinion as opposed to sharing facts about their conditions. Annotators totally agreed on 70% of sentences in the collection, with a further 21% receiving a majority vote in favour of one label. In the remainder of this chapter, we describe how we used this data to train and evaluate a model to automatically apply labels to sentences.

7.2 Identifying Personal Experience

Blogs are used to share personal experiences and opinions as well as information and advice. For example,

Personal experience:

"After my discharge from hospital a couple of weeks ago I continue to monitor my condition and so far, so good with no sign of another infection."

Information:

"Some COPD patients live alone, and are in many aspects isolated."

Advice:

"A better plan might be to let your doctor know what is going on so that he or she can find a way to relieve the problem."

Our focus is within the medical domain and so our definition of personal experience is based there. We define personal experience as a description of events from the author's point of view, discussing events that they were part of and thoughts and feelings they had. These statements relate specifically to the author's personal circumstances and point of view, as opposed to more general information and advice relating to the condition or medical area.

As with the discussion of sentiment analysis above, when analysing people's subjective opinions and experiences, these kinds of content need to be filtered from the others. While sentiment analysis relies on a definition of subjectivity relating to whether a sentence is

opinionated or not (i.e. it conveys the author's opinion or sentiment towards some entity), we use 'subjectivity' to refer to whether a sentence relates to a personal experience or some other kind of information. In this section, we discuss this definition, set out the classification problem and our approach to automatically extracting this subjectivity information.

7.2.1 Approaches to subjectivity classification

Subjectivity analysis is a binary classification task (i.e. classifying a sentence as *subjective* or *not subjective*), meaning that approaches often involve finding the optimal feature set or method upon which to base a decision. Wiebe, Bruce, & O'Hara (1999) created a corpus of 1,001 sentences which three annotators had annotated manually in terms of subjectivity. Automatic classifiers were trained on linguistic (five part-of-speech tags), lexical and paragraph level features and a supervised learning approach achieving 72% accuracy. Using the same corpus, they applied statistical analysis which highlighted the presence of adjectives as significantly positively correlated with subjective sentences (Bruce & Wiebe, 1999). Subsequently, work has focused on building subjectivity lexicons (J. Wiebe, Wilson, Bruce, Bell, & Martin, 2006; J. Wiebe, 2000) and automatically generating rules for subjectivity classification (Riloff & Wiebe, 2003).

One challenge with applying supervised machine learning is the creation of an annotated corpus for training and evaluation. Recent work has explored overcoming this issue with semi-supervised approaches. Wang et al. (2008) used a semi-supervised approach to subjectivity classification employing decision tree algorithms which re-trained on high-confidence automatically classified sentences. Su and Markert (2009) also used a semi-supervised approach in order to label word senses according to their subjectivity achieving similar results to the supervised approach with less that 20% of the training data.

Subjectivity classifiers have been employed to improve and enhance a number of other classification and information extract systems. Pang and Lee (2004) used subjectivity classification in order to improve sentiment analysis of movie reviews. They involved subjectivity labelling as a pre-processing step, only providing subjective segments of text, improving performance for Naive Bayes sentiment classifiers. A similar approach was used by Jiang et al. (2011) for classifying the sentiment expressed in posts on Twitter. Likewise, Riloff et al. (2005) used the objective classifications to improve the precision of an Information Extraction system focused on terrorism. They found that selectively filtering

extracted information from subjective sentences improved precision with minimal impact on recall.

7.2.2 METHOD

Our approach to automatically label sentences as relating to some subjective experience of the author, or more general information was based on the crowd-sourced labels described in Section 7.1.1. Given the subjective and highly variable nature of the problem, we have opted for a supervised machine learning approach as opposed to a rule-based approach. Such an approach relies on a training set of annotated examples. In this case that implies the collection of relevant documents (i.e. blog posts) and manual annotation of individual sentences that refer to a subjective patient experience. Finally, it remains to select an appropriate set of features that adequately characterise subjective sentences. This section details how we utilised crowd-sourced data, linguistic information and supervised machine learning in order to train a model to enable automatic labelling of new examples.

7.2.2.1 DATA COLLECTION

The data used to train and evaluate the models were collected as described earlier in the chapter. Participants in the online annotation exercise (see Section 7.1.1) were asked whether each sentence related to a personal experience or some more general information. 1770 sentences were included in the exercise. As described in Section 7.1 we used Krippendorff's alpha coefficient (K. Krippendorff, 1970) to rate the inter-annotator agreement. The coefficient α=0.55 implied less than perfect agreement between annotators, but greater than that expected by chance (α=0).

Table 14 describes the annotations in terms of inter-annotator agreement. The majority (1247 out of 1770) of sentence annotations were agreed on unanimously between all annotators to whom they were shown. All but 135 sentences had a majority vote in favour of a single annotation.

Agreement	**Sentences**
No agreement (i.e. 1/2)	135
Majority agreement (2/3, 3/4)	388 (379, 9)
Total agreement	1247

Table 14: Sentences by subjectivity annotation agreement

In order to train models for extracting subjectivity information, the annotated data was randomly split in to two datasets. The first (dataset A) contained only sentences where the annotation was agreed on unanimously (i.e. agreement = 100%, *n*=1247) and the second (dataset B), where there was a majority in favour of one annotation (i.e. agreement >50%, *n*= 1635). Datasets A and B overlap, but will be used separately in order to compare their usefulness. Using agreement as a measure of annotation quality, we will train and test classifiers on each dataset independently in order to assess their impact on classification performance.

7.2.2.2 FEATURE SPACE

Our approach focused on a token-level representation of each sentence within the dataset. Following a similar method to Weibe et al. (1999, 2000) we focused on lexical features and statistical analysis to analyse their descriptiveness. Tokens were coupled with their part-of-speech tag (e.g. noun, verb, adjective or adverb) extracted during the pre-processing step of blog post collection (see Section 3.2.2) and generalised according to their meaning using WordNet (Princeton University, 2013) synsets using the WordNet API. WordNet is a lexical database providing information about general English words. Synsets allow grouping of words with the same meaning (see Figure 23) and were used to group similar tokens (i.e. 'disbelieving' and 'skeptical') into more general features thereby creating a more useful model for generalisation.

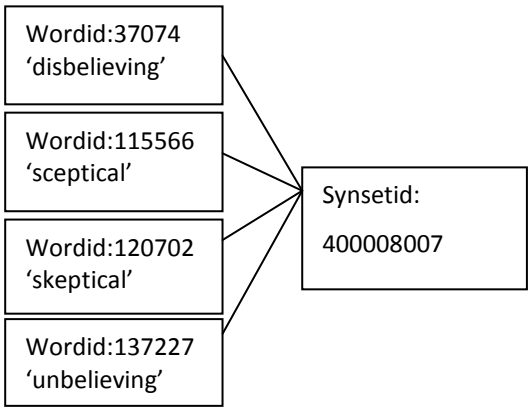


Figure 23: Wordnet synsets

Personal pronouns used in sentences were also included in the feature set. Pronouns were grouped in to three classes – first person ('me', 'I', etc.), possessive ('my', 'mine', etc.) and third person ('you', 'them', 'they', etc.). When discussing a personal experience, statements in the first person would be expected intuitively. Pronouns would therefore be a potentially

valuable feature for classifying experiential sentences. A summary of the feature groups and their 5371 features can be found in Table 15.

Feature	Description
Length	Length of the sentence in tokens
Pronouns	Relative frequency of personal pronouns present separated into three classes 1. First person ('I', 'me', etc.) 2. Possessive ('my', 'mine', etc.) 3. Third person ('them', 'they', etc.)
Nouns	Relative frequency of noun tokens, grouped by synset (1998)
Verbs	Relative frequency of verb tokens, grouped by synset (713)
Adjectives	Relative frequency of adjectives, grouped by synset (644)
Adverbs	Relative frequency of noun tokens, grouped by synset (212)

Table 15: Feature space

7.2.2.3 FEATURE SELECTION

The discriminative power of each feature was assessed using information gain analysis (Yang & Pedersen, 1997). Information gain values represent the weight of information held by a feature regarding a class. It is one indication of how useful features are for discriminating between classes. The general form of information gain for nominal classes is computed as:

$$\text{Information Gain}(\text{Class}, \text{Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute}) \quad \textbf{10}$$

Where H represents the information entropy – a measure of uncertainty about a random variable.

Features were then ranked in order of the information gain values associated with the annotations (i.e. experiential sentence or not). The top 10 features according to this analysis are shown in Table 16 and a summary of the results of the analysis in Table 17.

Rank	Feature (Info. Gain value)	
	Dataset A Agreement 100%	***Dataset B Agreement >50%***
1	First person pronoun (0.48)	First person pronoun (0.358)
2	‘have’ (0.047)	'have' (0.033)
3	‘disease’ (0.039)	'disease' (0.031)
4	‘lung’ (0.032)	'patient' (0.023)
5	‘patient’ (0.031)	'chronic' (0.02)
6	‘get’ (0.027)	'lung' (0.019)
7	‘chronic’(0.027)	'last' (0.018)
8	‘last’ (0.021)	'so' (0.018)
9	‘symptom’ (0.18)	'get' (0.015)
10	‘good’ (0.18)	'move' (0.015)

Table 16: Top 10 features from information gain analysis

	Number of features where Information Gain>0	
	Dataset A Agreement 100%	***Dataset B Agreement >50%***
Nouns	76	80
Verbs	30	33
Adjectives	19	19
Adverbs	14	12
Pronouns	2	2
Total	141	146

Table 17: Number of potentially useful features by feature group

Table 16 shows that the relative frequency of first person pronouns in a sentence exhibits the most discriminative power as intuitively expected. Tokens from other feature groups

(nouns, verbs, adjectives and adverbs) are also highly ranked, but with much lower values, showing that they are not expected to be useful for this classification task.

7.2.2.4 MODEL BUILDING

To evaluate the features chosen to represent sentences, models were trained on various combinations of features from the set described in Section 7.2.2.3. Subsets of our overall feature set (i.e. nouns, verbs, adjectives, adverbs and pronouns) were evaluated separately as well as using them all together. For each subset the ranked set of token-features with non-zero information gain values were used (see Table 7) to train a Naive Bayes classifier. Naive Bayes classifiers (Domingos & Pazzani, 1997) estimate the probability of a hypothesis based on previous evidence and Bayes' theorem:

$$P(h|e)=\frac{P(h)P(e|h)}{P(e)} \quad \textbf{11}$$

Where *h* is some proposition and *e* relates to some evidence. The *posterior* ($P(h|e)$ – the probability of some proposition based on observed evidence) can be calculated based on previously observed data (or *priors*). During the training phase, the training data is used to calculate the prior probabilities (the observed evidence). This can then be used to estimate the probability of a hypothesis on previously unseen data. A Naive Bayes classifier was chosen as it performs well with relatively small training sets (Domingos & Pazzani, 1997) as shown in a previous text classification study (Spasić, Burnap, Greenwood, & Arribas-Ayllon, 2012). Subsequent experiments with Support Vector Machines failed to train a successful model. With each feature set the majority class was chosen in all test cases (see Table 18).

		Automatic label	
		subj	**obj**
Manual label	**subj**	693	4
	obj	542	8

Table 18: Example confusion matrix for SVM experiment test

Classification models were built and evaluated using 10-fold cross-validation using the Weka (Hall et al., 2009) machine learning package, version 3.6.

7.2.3 RESULTS

The classifiers were trained and tested on datasets A and B separately using the selected features in combination and isolation in order to evaluate their appropriateness for this classification task. As the results of the information gain analysis (Table 6) show, the pronouns used in the sentence expected to be most informative. The results of our experiments are shown in Table 19 and Table 20.

	No. of Features	Precision	Recall	F-Score	Kappa
Nouns	76	0.449	0.6	0.513	0.4333
Verbs	30	0.666	0.656	0.661	0.3583
Adjectives	19	0.6	**0.993**	0.748	0.1695
Adverbs	19	0.825	0.311	0.452	0.2102
Pronouns	2	**0.933**	0.803	**0.864**	**0.7174**
All	141	0.719	0.958	0.822	0.5064

Table 19: Naive Bayes classifier results (dataset A - agreement =100%)

No. of Features	Precision	Recall	F-Score	Kappa	
Nouns	80	0.643	0.942	0.764	0.3253
Verbs	33	0.64	0.818	0.718	0.2703
Adjectives	19	0.579	**0.991**	0.731	0.1291
Adverbs	13	0.784	0.288	0.422	0.1804
Pronouns	2	**0.902**	0.728	**0.806**	**0.6201**
All	147	0.672	0.94	0.784	0.4024

Table 20: Naive Bayes classifier results (dataset B - agreement > 50%)

Table 19 and Table 20 show the results achieved by the different feature sets on the two datasets described. With both unanimous and majority agreement datasets, the pronoun

features performed the best when classifying experiential sentences. First person and possessive pronouns were used to achieve over 90% precision in both datasets, and 72-80% recall. The chance-corrected Kappa scores for both the majority and unanimous datasets show that the pronoun features performed much better than the random classifier baseline (0.62 and 0.72 respectively). The Kappa scores attached to each outcome compare the results achieved to picking classes at random. A Kappa score of 0 signifies a result no better than random while a score of 1 signifies perfect agreement.

The combined features achieved greater recall at the expense of precision. Kappa agreement values showed results closer to that expected by chance (0.40 for Dataset B, 0.51 for Dataset A).

In summary our results show that sentences relating to patient experience can be automatically extracted from patient accounts shared online with good confidence. They indicate that the frequency of first-person or possessive pronouns is the most appropriate features of the sets tested. The high precision of pronouns (90-93%), but relatively low recall (72-80%) indicates this feature set should be expanded in order to increase coverage, but the general classes of token used here were largely unsuccessful. Adding more information to the textual data may add value to the features selected. Our future work will focus on augmenting the linguistic information used here with semantic information. Utilising emotional lexicons, such as WordNetAffect (Valitutti & Stock, 2004) and SentiWordNet (Esuli & Sebastiani, 2006) will allow us to generalise tokens in terms of the emotion they express, which could be a useful in separating personal statements from more objective ones. Building on the low-level linguistic information may also help build more accurate models. Exploiting phrase-level features (such as verb or noun phrases), rather than token-level may also increase the information load.

7.3 Conclusion

The emotional component of patient accounts is important in interpreting their experience of illness. As part of the qualitative analysis carried out in Chapter 5, themes around psychological impact, and the emotions that patients encountered as a result of exacerbation of their conditions were highlighted through manual interpretation of the data. In this section, we explored both distributed and automatic approaches to labelling sentiment information in blog posts.

Our Crowdsourced data annotation task focused on three key elements of the patient experience accounts – whether a sentence was objective or subjective (i.e. whether it referred to an *internal state* or not), whether the emotions expressed were positive, negative or neutral in nature and lastly whether a sentence expressed some *need* that might be fulfilled through better treatment or services. In all, just over 280 annotators labeled a subset of the 1770 sentences included from 100 patient blog posts. We evaluated the data in terms of inter-annotator agreement to assess whether or not the consensus derived from agreement between annotators could be depended on, or whether it could have occurred by chance. For the emotional tone and need datasets, unfortunately the confidence in the data was not high enough to say with certainty that agreement between annotators was a good indicator of the sentences' label. In Section 7.1.4, we discuss a number of ways in which subsequent tasks could be designed or carried out in order to try and address the difficulties faced by this data – namely, more annotators and a more focused dataset for such specific and complicated annotation tasks.

The data from the subjectivity question showed good agreement between annotators (α=0.55), resulting in 1600 sentences which had a majority vote in favour of one label between all annotators. We used this data alongside the linguistic information extracted during pre-processing (see Chapter 4) we trained a Naive Bayes classifier to automatically label sentences as relating to some personal experience or more objective facts. The classifier was trained using a number of linguistic features including the part-of-speech tags of the tokens that made up the sentence, and the individual tokens themselves, with the presence or absence of pronouns emerging as the most discriminative. Our classifier, trained and evaluated against the Crowdsourced data, achieved a high accuracy (p 0.90, r 0.80, κ 0.60).

While also potentially increasing performance of subsequent sentiment classifications, identifying subjective statements in blog posts is an important element of highlighting patient experience – or providing easier access to specific information required by researchers. In our task area, looking at experiences of exacerbation for COPD patients, being able to identify sentences relating to the subjective experiences of individual patients could help speed up analysis of the data. In the next chapter, we introduce some ways in which the combination of the subjectivity information with the semantic information added during the work in Chapter 6 could provide faster insights in to what patients are discussing online.

CHAPTER 8

EXPLORING THE DATA – INTEGRATION AND FURTHER ANALYSIS

The ultimate aim of the QuTiP framework, as described in Chapter 3, is to enable qualitative researchers to analyse a much larger set of patient social media data through augmenting the effort of interpretation with automated Text Mining tools. These tools are developed in conjunction with an initial qualitative investigation carried out on a development subset of the overall collected corpus.

We have described the analysis carried out on the development subset. Firstly, we described the qualitative analysis in Chapter 5, where key concepts and themes were drawn out of the data through systematic manual interpretation. We then described extracting both the semantic (Chapter 6) using manually curated terminologies and rule sets and through applying FlexiTerm – the flexible approach to Automatic Term Recognition. Lastly, we introduced the Crowdsourced sentiment labels and subjectivity classifier (Chapter 7).

The next stage in the QuTiP process is to apply the tools developed on the rest of the collected data (all 368 posts) in order to identify interesting patterns that could warrant deeper investigation. The aim here is not to evaluate the methods, but to use them to explore the remaining data for which there are no gold standard annotations.

In this chapter, we describe our approaches to discovering patterns in the patient discourse and the tools we developed to help researchers make use of the semantic and sentiment information which has been added to this dataset using the automated methods described in Chapters 6 and 7.

8.1 EXPLORING THE WHOLE DATASET

The semantic and sentiment annotations (both manual and automatic) as described in Chapters 6 and 7 add important information to the blog posts data we have collected. While the semantic information – the important medical terms used – can inform as to

what patients were discussing – the concepts involved, the sentiment information garnered during Chapter 7 can tell us about the context that was used to describe them.

Now that the methods have been developed and evaluated on the development subset of 100 documents, in this section we describe processing the larger set of 358 blog posts. As well as describing the resultant data, we also explore patterns relating to the concepts involved in sentences and the context in which they are found (i.e. personal or objective).

8.1.1 Semantic Information Extraction

As described in Chapter 6, while standardized lexicons are useful for identifying concept mentions in text, they are brittle when dealing with terminological variation. ATR systems allow the terms to be extracted from text based on linguistic or statistical queues (see Section 6.2). FlexiTerm, a novel ATR approach developed as part of our work, is specifically tuned to overcome the limitations of terminological variation when recognizing terms in text. By grouping variants of the same term together, computing statistics about the representativeness or importance of the term within the corpus can be done more accurately as shown in our evaluation.

In this section, we describe the results of applying FlexiTerm on the whole corpus of 368 blog posts. We also classify a subset of the terms in light of the qualitative framework developed in Chapter 5. Relating terms to this conceptual framework means they could be used to more easily explore the dataset.

Rank	Term variants	Score	Classification
1	breathe easy	45.7862	Support network
	easy breathing		
2	pulmonary rehab	45.0546	Treatment
	pulmanory rehab		
3	lung disease	35.2071	Disease
	lung diseases		
	diseased lungs		
4	chronic disease	32.6683	General concept
5	chronic obstructive pulmonary disease	27.7259	Disease
6	vitamin d	25.5834	Treatment
	d vitamin		
7	lung function	24.6067	Functional concept
8	quality of life	21.4876	Functional concept
9	shortness of breath	20.1013	Symptom
10	chest infection	19.4081	Infection
	chest infections		

11	chronic bronchitis	18.715	Disease
12	breathe easy groups	18.4567	Support network
	breathe easy group		
	breath easy groups		
13	lung transplant	17.6257	Treatment
	lung transplantation		
	lung transplants		
	lung transplantations		
14	blood pressure	17.3287	Tests
14	copd patients	17.3287	Patients
	patients with copd		
	copd patient		
15	chronic lung disease	16.3793	Disease
16	pulmonary disease	13.9785	Disease
17	british lung foundation	13.1833	Support network
18	stem cells	13.0882	Treatment
19	easy group	13.0543	Support network
20	support groups	12.8232	Support network
	support group		

Table 21: Top 20 FlexiTerm terms - whole corpus

Table 21 contains the top 20 grouped terms identified using FlexiTerm. The entire list of over 450 terms is presented in Appendix C. In order to classify the top 100 terms into main topics of conversation, we manually mapped them to the nodes in the coding hierarchy developed during the qualitative analysis phase (see Chapter 5). Naturally, most concepts included in the 100 top-ranked terms relate to the healthcare domain. Interestingly, FlexiTerm revealed some "unexpected" concepts, which provide valuable insight into the self-management of COPD. One such example is ‘laughter yoga’, a term that does not appear in the Unified Medical Language System terminology (accessed June, 2014), and is therefore difficult to discover when relying on traditional methods and supporting resources. This unusual term refers to an alternative therapy, which is discussed in blog posts as a useful way for addressing respiratory symptoms associated with COPD. The following concordances show the term 'laughter yoga' used in context, which paint an interesting picture into everyday life of COPD patients outside the traditional healthcare boundaries.

```
from it, too. I'm thinking of having a "Laughter Yoga Party" if I can locate a quali
should, especially if you're practicing Laughter Yoga. A growing alternative health
    A growing alternative health trend, Laughter Yoga combines laughter and yogic
      As with a cardiovascular workout, Laughter Yoga increases air supply and circu
 particularly jolly during a session of Laughter Yoga? You'll be laughing anyway
     or mood. The total body therapy of Laughter Yoga is available to all. While
While yoga is a centuries-old practice, Laughter Yoga is a modern creation, develope
It has since spread to sixty countries. Laughter Yoga is traditionally a group pract
  as physical challenges. Supporters of Laughter Yoga say that it helps them to live
     you experience during a session of Laughter Yoga? Group members typically stand
     20-30 minutes, ends in meditation. Laughter Yoga practice benefits from an open
    the simha mudra - the lion laugh of Laughter Yoga. To find a Laughter Yoga group
lion laugh of Laughter Yoga. /To find a Laughter Yoga group in your area, check onl
       over the phone with a West Coast Laughter Yoga group called Laughter Yoga
a West Coast Laughter Yoga group called Laughter Yoga On The Phone that has set up
 practice! You'll join the thousands of Laughter Yoga devotees who know that laught
mindbodysanctuary.com/index.php/article/laughter_yoga/35877/It greatly annoys me
```

Similarly, ‘wood stove’ is another example of an "unexpected" highly ranked term extracted automatically by the FlexiTerm method. Discussed in the blog dataset as an irritant (i.e. cause of exacerbation), wood stoves appear in posts written by patients giving advice to others. ‘Wood stove’ does appear as a concept in the UMLS terminology (Concept id: C1268550), but the link between wood stove and exacerbation is not present in the semantic network. Having extracted the term, interpretation of its meaning within the context of the subject matter discussed can highlight why the term is present.

```
Sometimes we rely on space heaters, wood stoves and kerosene heaters. Wood stoves
  wood stoves and kerosene heaters. Wood stoves and kerosene heaters are not
ooking. Avoid using a fireplace or wood stove; wood smoke is a lung irritant
```

These examples highlight the strength of ATR approaches over solely relying on pre-constructed vocabularies. This is especially important for applications such as ours, where we are applying data-driven, inductive qualitative approaches. Were we to solely use standardized vocabularies, it is possible for concepts present in the text which are not included in standard terminology to be completely overlooked. By using an ATR approach as well, important concepts are highlighted in a data-driven manner.

8.1.2 Subjectivity Classification

Finding subjective statements (i.e. relating to the author’s personal experiences, rather than some objective information) is an important component of our investigation in to patients’ experiences of COPD exacerbation. In Chapter 7 we described our approach to utilizing crowd-sourced sentence labels in generating a Naïve Bayes model to automatically classify labels as relating to a personal experience or some objective information. The model we trained performed well on test data, achieving high precision and recall (0.933

and 0.864 respectively) and a Kappa agreement which shows good agreement between manual and automatic labels (0.717).

We applied this classifier to each of the 7955 sentences in our corpus of 368 blog posts. Each had a label of either 'subjective' or 'objective' – relating to the author's personal experience or not.

Label	Number of sentences	%
Subjective	3503	44.0
Objective	4452	56.0

Table 22: Sentence Subjectivity results - whole corpus

Table 22 summarises the labels applied to the contents of our blog post corpus. As with the manual annotations collected through our crowdsourced online exercise (see Section 7.1.1) there is a rough 50% split between both labels. Table 23 and Table 24 contain ten examples of subjective and objective sentences respectively.

1	ho well continue the dream, we will get there, my point is if we were on the ground floor there would be so much more I could do , and take some of the load from lins shoulders, I remain positive, it will happen, it's just when ?
2	An exacerbation of COPD causes severe shortness of breath and that is just the hardest thing to deal with for me.
3	Here is a linkie about pre-exhaust principles Pre = exhausting With the theory links done what I do for COPD is a modified pre-exhausting system.
4	After a LONG separation from my home while I was in the hospital and then in rehab, my hip is still not healing.
5	Hello to all my friends here at `` Welcome To My Little Corner of The World '' well it's been a while since my last post and let me tell you what ride I have been on !
6	My day was just pure torture.
7	I have seen our first snowdrops springing up the garden is coming back to life again.
8	I have taken a time away from this blog to give others a chance to write themselves.
9	Thank goodness I have a reserve stock of these prescriptions as we are still snowed in and I wouldn't be able to get to our surgery or chemist although my husband would be able to walk there for me.
10	This may mean we'll get the new years resolution people here today...if you've quit smoking as your resolution, I 'm going to give you 10 very simple tips to quit.

Table 23: Sentence Subjectivity results - 10 subjective examples

1	The message here to my fellow men, come on guys, a support group will help you.
2	Start out slowly and increase your pace and time as the days go by.
3	How much blood was my heart pumping a minute.
4	Thanks for reading.
5	If heart failure develops, there may be swelling in the ankles, legs, and sometimes the abdomen.
6	So if there is anyone reading this who has influence - could it not be common practice that those diagnosed with COPD are referred to the specalised respiratory team in their area?
7	Even then as long as I look after myself keep away from crowds as crowds are a good way to pick up infections make sure you get your annual flue jabs have a pneumonia jab get plenty of rest eat well sleep well get some exercise know matter how little stay warm in winter - and stay positive you can still look forward to quite a few years of good life.
8	This week's issue of the COPD International Newsletter brings Featured Articles - Energy-Saving Holiday Tips TIPS FOR ENJOYING HOLIDAYS WITHOUT GETTING BREATHLESS The holidays are a special time of the year , however they can very stressful.
9	Orlan is reaching out and asking for help in the fight to save his life.
10	Less air intake means less oxygen for the muscles, making even the most simple tasks difficult .

Table 24: Sentence Subjectivity results - 10 objective examples

The examples given illustrate the different parts of the narrative extracted. In the first table, comments are about the authors' own experiences – their difficulties (sentence 1, 2, 6, 9), treatment progress (sentence 4), management techniques (sentence 3) and general conversation or about the online community as a whole (sentences 5, 7, 8 and 10). Table 24, however, contains more general information or advice (sentences 1, 2, 5, 7, 8, 9, 10) and, interestingly, suggestions and open questions (e.g. sentence 6).

While the subjective information was more useful for our inquiry about experiences of COPD exacerbation, the objective may also prove useful for other investigations carried out on patient narrative data.

8.1.3 PATTERN ANALYSIS - SEMANTIC-SUBJECTIVE CONTEXT CORRELATION

The semantic and subjectivity information found within this blog post dataset using the techniques developed as part of our work can give us certain insight in to what patients are discussing. Firstly, the concepts that are involved, and secondly whether they are sharing information or personal experiences. This information can also help researchers summarise and navigate the data – finding sentences that discuss certain things, for instance. However, the information added to the plain text data can also be used to discover patterns in the data that may yield new information or knowledge. In this context, finding new patterns within discourse of patients could help identify new hypotheses or questions that could be answered through further manual analysis of the blog post content.

In order to discover more about patient experience, we designed an experiment to explore the context in which medical concepts were mentioned by patients. We analysed correlations between individual concept mentions (at the phrase-level) and the subjective or objective context in which they appeared (the sentence level). We used the occurrences of the terms discovered by FlexiTerm in the whole dataset of 368 blog posts, and the subjective/objective labels assigned by the subjectivity classifier to the 7955 sentences. The patterns uncovered could help us explore the elements of their conditions that patients discuss more frequently in accounts of their personal experiences. Likewise, this information could unlock new knowledge about the concepts about which patients share more objective information.

In order to find relationships between these two different sets of annotations, we used Pointwise Mutual Information (PMI) analysis (Church & Hanks, 1990). Given the two sets of annotations, PMI measures the probability of two outcomes co-occurring and the probability of them occurring separately, and gives an overall measure of 'relatedness' (see Equation 12).

$$pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad \textbf{12}$$

We computed the PMI for each term and both objective and subjective contexts. This meant that we could identify concepts which were strongly correlated with being discussed about in personal experiences shared by patients online. Figure 24 shows the terms correlated with sentences classified as subjective, while Figure 25 summarises the terms correlated with objective sentences.

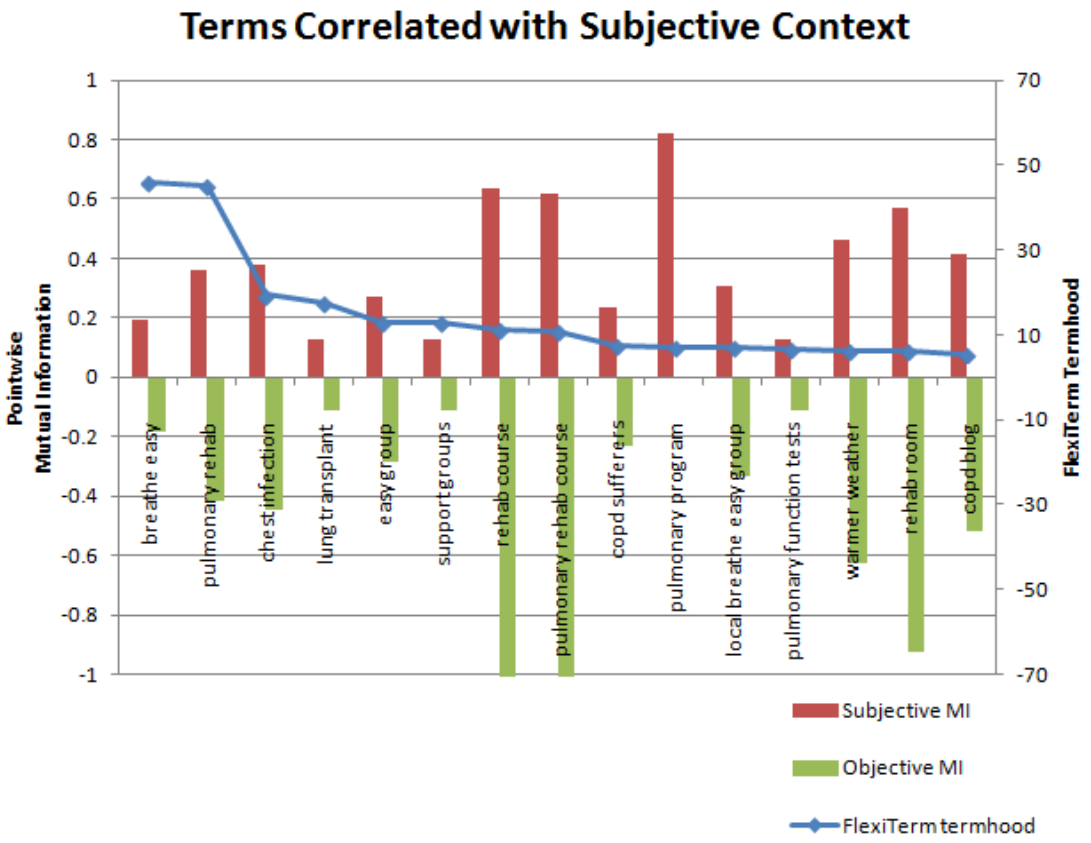


Figure 24: FlexiTerm terms correlated with subjective context

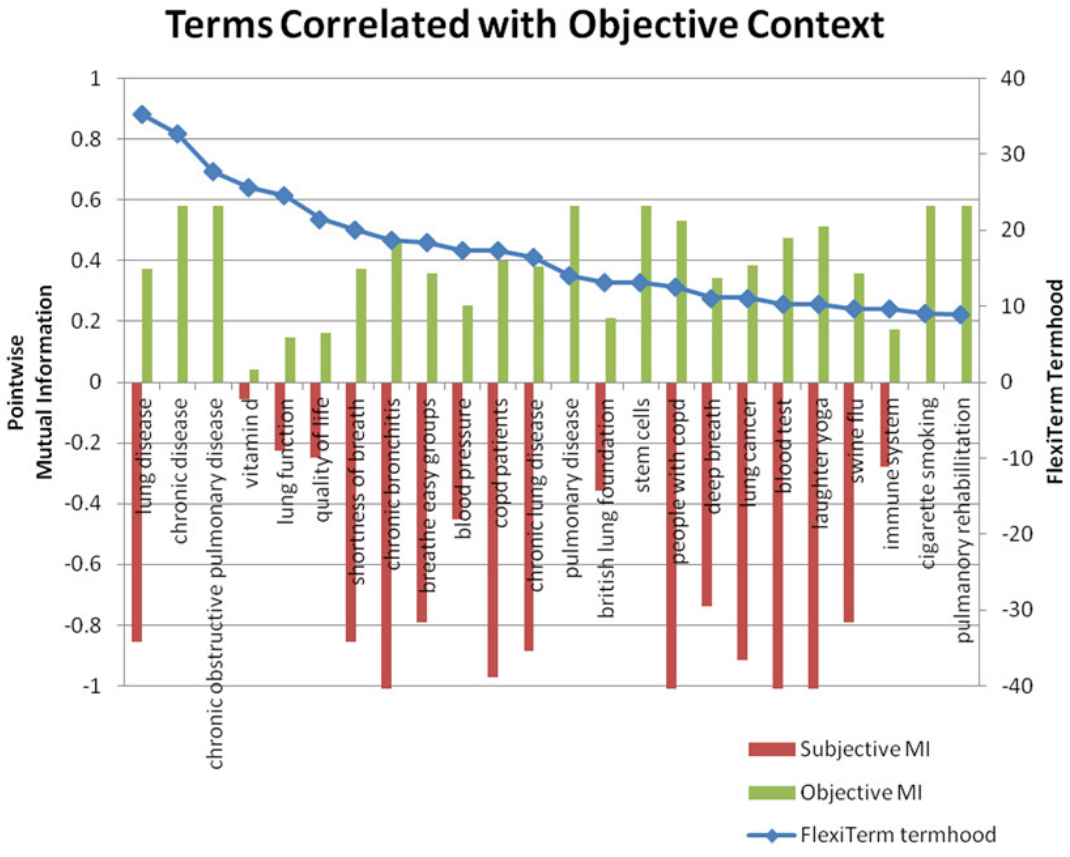


Figure 25: FlexiTerm terms correlated with objective context

From this analysis, we can explore which concepts are involved in patients' discussions about their experience of illness. For instance, in Figure 24, we see that the term 'chest infection' is correlated strongly with subjective sentences.

"I tried to explain that he was just ill and on top of that COPD and the anti-biotics would hopefully stop the chest infection developing."

"I got a very bad chest infection in January and had the normal antibiotics and steriods to no effect. "

"And this can give me an almost foolproof method of determining if a suspected coming chest infection is in fact one."

While the term is most strongly associated with subjective context, it also occurs in objective contexts, sharing information, for example:

"A chest infection at best , breathlessness a certainly , a hospital stay for some , and for those with severe chronic copd , even death a possibility ."

"When a chest infection starts to set in it is harder for the heart to pump blood through the lungs , and the pulse rate shoots up and it will stay high while resting ."

Having mapped the nodes to the node hierarchy (see Section 5.3, Table 5 and Appendix B) we can also group terms, based on the concepts they related to. The terms 'pulmonary rehab', 'pulmonary rehab course', 'rehab course' and 'rehab room' all map to the same concept (pulmonary rehabilitation courses) and are all correlated with a subjective context. There are 52 subjective sentences relating to pulmonary rehabilitation within the blog data, and these sentences have been extracted and highlighted through a Text Mining approach.

"I had to ASK HIM about Pulmonary Rehab (I had been reading about it) ."

"I had reservations about the long walk from the parking lot through the hospital to the rehab room, but I did it with just one stop by the Gift Shop to catch my breath."

"After several years of stubbornly refusing Pulmonary Rehab, and being miserable with my declining health, I finally surrendered (sic) and let go of the denial."

The data analysis performed allows us to summarise the contents of the thousands of sentences that make up the collected blog posts, and find interesting trends without needing to read each individual post. Having analysed a subset of those document manually, we can use what we learned in conjunction with automatic analysis to unlock other trends and information from a much larger dataset. In the next section, we describe tools we have made to help make this information more usable as part of a larger qualitative analysis.

8.2 Exporting Data for Use in Qualitative Research

We've described how information can be extracted automatically from blog posts. The labels gathered and generated take the form of spans of text, coupled with labels (i.e. subjective/objective) and scores (such as the termhood measure used by FlexiTerm). To be usable, this information must be presented in a form that is interpretable by qualitative researchers in the context of a much wider qualitative effort. To that end, we investigated and developed tools to export custom annotations such as those created by the methods detailed here, and export them in to a format that can be interpreted by existing qualitative analysis software packages.

Research towards standardising an open qualitative data exchange model has been undertaken by the Qualitative Data Exchange project (QuDEX)[18] and a schema for representing qualitative data (including coding hierarchies and annotations) is available. Support for this open format is built in to Atlas.ti[19], one of the current market leading qualitative research packages. Given the open-nature of the standard, more packages should hopefully adopt the format as this will help improve data sharing.

[18] http://data-archive.ac.uk/create-manage/projects/qudex
[19] http://atlasti.com/

The output is an XML-style document, which encodes the documents (including meta-data such as the title, and creation date), the contents of the documents and the annotations and labels generated by tools we have described in the preceding chapters.

Figure 26: QuDEX format output

Figure 26 shows some example QuDEX-format output from our export tool. The documents are listed and the annotations take the form of segments, which are a character offset and length, along with a corresponding identifier and plain-text description (see Figure 27). This information will be displayed as the annotation in the interpreting qualitative package.


▼<qudex xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.data-archive.ac.uk/dext/schema/draft"
xsi:schemaLocation="http://www.data-archive.ac.uk/dext/schema/draft http://www.data-archive.ac.uk/dext/schema/draft/
08-23T20:34:58.192000" mdate="2015-08-23T20:34:58.192000" creator="user1" label="Test Collection" displayLabel="Qude
the elements and attributes present in schema model " status="auto" language="en" id="Test_Collection">
▶<resourceCollection id="rc_1">...</resourceCollection>
▼<segmentCollection id="scol_1">
▼<segment id="MMTerm_C1314677_Maintained" cdate="2015-08-23T20:35:00.041000" mdate="2015-08-23T20:35:00.041000" l
Concept: Maintained" displayLabel="MetaMap Term CUI: C1314677 Concept: Maintained" creator="user1_auto" language
▼<text id="text_doc5_sent9" src="document_3">
<lineParam id="doc_5_line_9_char_61_70" startOffset="61" startLineOffset="8" endOffset="70" endLineOffset="8"
</text>

▼<segment id="MMTerm_C0003787_Arizona" cdate="2015-08-23T20:34:58.720000" mdate="2015-08-23T20:34:58.720000" labe
Concept: Arizona" displayLabel="MetaMap Term CUI: C0003787 Concept: Arizona" creator="user1_auto" language="en">
▼<text id="text_doc4_sent18" src="document_2">
<lineParam id="doc_4_line_18_char_97_103" startOffset="97" startLineOffset="17" endOffset="103" endLineOffset
</text>

▼<segment id="MMTerm_C0087009_Specialist" cdate="2015-08-23T20:34:58.720000" mdate="2015-08-23T20:34:58.720000" l
Concept: Specialist" displayLabel="MetaMap Term CUI: C0087009 Concept: Specialist" creator="user1_auto" language
▼<text id="text_doc4_sent2" src="document_2">
<lineParam id="doc_4_line_2_char_18_27" startOffset="18" startLineOffset="1" endOffset="27" endLineOffset="1"
</text>

▼<segment id="MMTerm_C1706852_Article" cdate="2015-08-23T20:34:58.720000" mdate="2015-08-23T20:34:58.720000" labe
Concept: Article" displayLabel="MetaMap Term CUI: C1706852 Concept: Article" creator="user1_auto" language="en">
▼<text id="text_doc4_sent21" src="document_2">
<lineParam id="doc_4_line_21_char_14_20" startOffset="14" startLineOffset="20" endOffset="20" endLineOffset="
</text>

Figure 27: QuDEX Annotations as Segments in XML

By supporting exporting the data created through applying the tools developed as part of this work, and described in this thesis, researchers can interact with the data further and add their own interpretation.

8.3 CONCLUSION

The objective of the QuTiP framework was to enable researchers faster, easier access to the information contained in patient blog posts by applying Text Mining tools developed alongside manual analysis. In Chapters 4 to 7, we introduced tools to gather and automatically extract information from patient blog posts related to COPD. While accomplishing these initial annotations offer benefits on their own (for instance, faster access to statements relating to certain concepts, or sentimental content) finding patterns within that data could help yield more insight in to patient discourse. Also, presenting the information in a format compatible with existing qualitative analysis packages is an important step to unlocking the potential of this data.

In this chapter, we described our approach to delving in to patterns within the data – relating concepts to the context in which they are most often found. Our analysis shows that terms related to pulmonary rehabilitation programmes correlate highly with subjective context. This may be indicative that patients are using their blogs to communicate to others how their treatment is progressing. This information could help form hypotheses about patients' accounts, or help guide further analysis of the data.

Likewise, we described a tool that can be used to export annotations to an open-format for utilization within well-established qualitative research software. Integration with standard qualitative packages is an important step to enable researchers to fully utilize the information added through automatic analysis.

Our analysis highlights a potentially important relationship between certain concepts and subjective or objective contexts. Our investigation centers on patient experience, and so the subjective accounts relate more strongly to our aims. Pulmonary rehabilitation emerged as an important concept in patients' accounts of their illness. This is supported somewhat by the initial qualitative analysis of the development subset – where 'rehab' and 'exercise' were added to the initial coding hierarchy (see Appendix B) representing a total of 45 references in 100 blog posts. The intended outcome of this analysis is some hypothesis that might be explored by further analysis. Our hypotheses and aims were driven by the research questions we had about patient experiences of COPD, but analysis

of the data as described here could form part of the hypothesis-development phase of qualitative research – where initial automated analysis of the dataset guides what is explored further. This final phase of our work offers the potential to give researchers quick descriptions of the data and describing patient discourse and guiding as to what might be found.

CHAPTER 9

CONCLUSION

The Web has been embraced by patients as a channel through which they can share information, their experiences and support each other through living with illness. While these functions give immediate benefits to participants in this online health discussion, these self-reported, first-hand accounts of illness also have the potential to inform on a larger scale with longer-term benefits to healthcare. Understanding how patients experience illness can help set policies, research priorities and to generate hypotheses that can be researched further. While classic approaches towards this kind of data gathering (e.g. interviews or questionnaires) can help sample a small population, analysis of online health discussions could include experiences of tens of thousands of people.

Manually analysing such large volumes of data will be time consuming, expensive and will ultimately increase the amount of time between information about patient experience becoming available and being utilised by healthcare professionals to improve treatments, management and outcomes. It is here that we have proposed that Text Mining may help augment the manual effort required to utilise patients' accounts shared through social media through qualitative analysis.

As stated in Chapter 1, it is our hypothesis that

a) text mining tools can automatically extract and classify information from patient-generated narratives
b) text mining and qualitative analysis can effectively combined to inform healthcare researchers about patients' opinions

In order to test this hypothesis we developed QuTiP, a framework which enables qualitative research and Text Mining of patients' social media content to be carried out in unison – allowing the results of qualitative analysis to influence the automated tools that will help scale up further analysis.

We applied QuTiP to investigate patient experiences of COPD shared through blog posts. We manually curated a relevant blog post collection based on the results of specialised online search tools focused on discovering blogs. As we set out in QuTiP, we carried out an initial qualitative analysis of a subset of the overall blog post collection. While primary outcomes of this phase are the themes found within the data under investigation, we also produce familiarity with the dataset and resources (such as the coding hierarchy) which can be used in the next stage of analysis – creating tools to automate some parts of the interpretation of text data. We created approaches to automatically extract both semantic (i.e. medical concepts) and sentiment (i.e. subjectivity) information. Lastly, we investigated ways in which we could find patterns within the automatically added information in order to support further investigation of the data.

In this chapter, we discuss how our work addressed each element of our hypothesis. We also summarise the contributions made, including methodological and reusable resources. Lastly, we set out important research issues facing the utilisation of patient accounts shared over social media and where our work can help in addressing some of them.

9.1 Text Mining Tools to Extract Patient Experiences

We hypothesised that utilising text mining tools, we could help researchers access information about patient experiences. This involved not only the semantic information (i.e. *what* they were discussing – the medical concepts involved), but also sentimental information.

Automated extraction of medical concepts mentioned in text has been applied to medical notes (written by physicians) and medical literature (see Section 2.2.4). The language used and the variance present in informal text, such as blog posts, offer new challenges to automated text classification (see Chapter 6). Standardised vocabularies, for instance, used to identify medical concept mentions may not cover slang terms used by patients to refer to elements of their condition or treatment. In order to address this, we produced two separate tools – a vocabulary built utilising informal sources of text to increase coverage and FlexiTerm, an automatic term recognition tool which doesn't depend on external knowledge sources and is tuned to deal with high terminological variance. Both these approaches showed promising results. The rule set based on vocabularies built using the outcome of literature review, external vocabularies and the coding hierarchy developed during qualitative analysis showed good coverage including terms to describe symptoms

that were not included in formal vocabularies (e.g. ‘gunk’ or ‘goo’). FlexiTerm performed better than the base-line method on our blog post subset and similarly identified concepts which were not present or not linked to respiratory conditions in more formal vocabularies (e.g. ‘wood burning stove’).

While sentiment classification has been investigated in a social media context (see Section 2.2.5), it has largely been related to consumer experience. We addressed sentiment analysis in a medical context, which, rather than identifying whether the author ‘liked’ or ‘disliked’ something related to a potentially wider range of emotions. We developed a task whereby participants could label sentences found in the collected blog posts as relating to three classes of sentiment information. The website created to enable people to complete this task was distributed over a number of channels on the internet allowing many participants to take part – a method known as crowdsourcing. The resultant data provided training and test data for applying machine learning to generate models for automatic labelling of previously unseen data. The model we produced based on linguistic features of the sentence under consideration can label sentences as relating to personal experience or some more objective information. Our evaluation of the model showed that it performed with high precision and recall (90% and 80% respectively) and a kappa score, (showing agreement between automatic and manual labels) of 0.62 - a good correlation, much higher than would be expected through chance.

These tools – the vocabulary and rule set and FlexiTerm for semantic information extraction and the subjectivity classification tool for sentiment information – can enable researchers to more easily access relevant information from within large collections of text. Moreover, this stand-off information, the labels given, can be mined for other interesting patterns. In Chapter 8 we described our approach to finding correlations between the semantic information identified using FlexiTerm and the subjectivity information added using the model we generated. The aim was to see which concepts correlated with subjective contexts and which correlated with objective contexts. Essentially, we aimed to identify the elements of COPD and COPD treatment that patients most readily share experiences of in their blogs. We identified pulmonary rehabilitation terms as most strongly related to a subjective context. This information could help guide further investigation of this data source and potentially through other channels (such as patient interviews).

The text mining tools described in this thesis – dictionary-based approaches developed in conjunction with a qualitative investigation, automatic term recognition for informal sources of text and subjectivity classification trained on crowdsourced labels – along with the application of data mining to explore relationships in the augmented blog post data show the potential for utilising text mining to explore patient experiences shared through social media data. While our application was specific to COPD, the methods and approaches developed, with the exception of the structured vocabulary generated, were not domain-specific. While the vocabulary we create is specific to COPD; the approach used to create it, involving qualitative coding hierarchies, could be repeated relating to other conditions. FlexiTerm has also been shown to perform well on data relating to other medical domains, but not yet on blog post data specifically (Spasić et al., 2013). Future work should include the evaluation of our approaches on blogs relating to other conditions. While the crowdsource task was successful in creating a good set of labels for subjectivity – a core concept in our investigation – the data created from the other questions relating to emotion and need were less successful. Classification of these potentially important facets of sentiment could provide other tools to interact with patient social media data. Other methods of gathering an authoritative set of training and test data for these types of data should be explored.

9.2 Text Mining and Qualitative Analysis

The QuTiP framework was developed to help integrate Text Mining and qualitative analysis to enable scalable investigation of patients' experiences using social media content (see Chapter 3). Rather than one approach supporting the other, the aim is for both to work in tandem. The initial qualitative outcomes help to inform the automated approaches which are then used to help inform further qualitative analysis. We used the outcomes of our qualitative investigation of patient experiences of COPD exacerbation to help develop information extraction tools, such as the structured vocabulary detailed in Chapter 6. While our qualitative analysis focused on a subset of just 100 documents, we used the resultant coding hierarchy in further automatically supported analysis of the rest of the dataset in Chapter 8. We related important terms which were automatically identified from the wider blog corpus using FlexiTerm to nodes within the coding hierarchy which allowed us to relate them to our thematic analysis. This mapping of terms allowed us to relate the patterns found within the data to the manually constructed hierarchy relating to the themes present in COPD exacerbation accounts.

We have proposed an approach to working with qualitative and Text Mining in tandem to aid the scalable analysis of patient accounts of illness shared online. The results in our study have yielded interesting areas for further inquiry within our chosen medical domain. The next step in our work would be to carry out a wider qualitative analysis of data we have collected and validate the themes found through automatic analysis. Likewise, similar studies should be carried out in order to establishing repeatability of our approach in other medical domains.

9.3 Patient Blogs and Health Care Experiences

Patient blogs contain the unedited, un-prompted accounts of people living with conditions, going through treatment and managing their health both in and out of the direct care setting. Traditionally, information about patients' experiences would be gathered through consulting with them through a structured interview or discussion group. However, this source of information can be difficult to arrange and represents a relatively small portion of the overall patient community. The collective content of the hundreds of thousands of blogs set up by patients looking to share their stories with others is a potentially valuable source of information for health care researchers to make sure that research addresses the most pressing issues affecting patients.

Focusing on COPD, we showed that patient communities are sharing their experiences of living with health care. In Chapter 5, we described a qualitative study, investigating the experiences of COPD patients living with exacerbation – focusing on how they identify and manage these events. We identified themes around what patients go through when exacerbations occur, the physical and psychological impact of exacerbations and the threat of exacerbations has on their day-to-day lives and how they try to address them. This qualitative study is amongst a growing number of qualitative investigations in to patient experiences which have been shared over social media (see Section 2.1.2), adding further weight to the argument that social media content in online health communities is an important source of information in modern health research.

In Chapter 7 we described an online crowdsourcing task which asked participants about the sentiment expressed in sentences taken from the collected blog posts. Questions focused on whether the sentence under consideration was related to a personal experience, whether it expressed positive or negative emotions and whether the author communicated some unmet need through this particular account. While the latter questions yielded

unusable annotations, the question relating to subjectivity provided good quality information about the experiential content in blogs. The agreement between annotators about whether a sentence was subjective or objective was rated using Krippendorff's alpha, computed as α=0.55 – a good correlation. A majority vote in favour of one label over another was achieved between annotators for 1635 of the 1770 sentences used in the task. The outcome showed that roughly 55% of the 1635 sentences were labelled as 'subjective', or relating to the personal experiences of the author. This provides some evidence towards how patients utilise blogs – sharing objective information and subjective accounts of their illnesses. As well as the qualitative information garnered through our analysis of the data, the Crowdsourced labels also provide some quantitative evidence that experiences are a major theme within patient blog posts.

We have shown that patients are sharing experiences of illness through blogs. Our qualitative analysis of the data we have collected as part of this work has shown that themes about patients' day-to-day lives can be highlighted within blog post content and that those insights could help to drive research priorities. The quantitative information adds weight to the assertion that patients are describing their experiences of health care through their blogs. These conclusions are limited by the focus of our study on COPD patients – however, the methods and tools developed as part of this research were designed to be neutral in terms of the target condition. Future research in to this hypothesis should involve a wider sample of the content created by this dynamic online health community.

9.4 Contributions of Our Work

The primary contribution of our work is the QuTiP framework – which can help integrate the manual qualitative analysis with automated Text Mining approaches. However, we also make a number of other contributions.

9.4.1 Methods

Our main contribution is the QuTiP framework for augmenting qualitative analysis with text mining and vice-versa. By integrating these two approaches, the QuTiP framework enables researchers to carry out scalable investigations on an extremely valuable source of information about patients' experiences outside of the direct care setting. By utilising the output of an initial qualitative analysis to drive further investigation and to build tools to automate parts of the interpretation of blog post content, QuTiP allows large collections of

text to be analysed. Likewise, the development of text mining approaches and models alongside the qualitative analysis allows for focused, high quality tools to be created.

QuTiP was intended to be generalisable to other medical domains beyond COPD, and part of our future work will be to apply this approach in other investigations of patient social media content.

9.4.2 DATASETS

Each phase of the work set out in this thesis has created data which could be reused for future work. Firstly, the blog post data itself, collected in a largely manual method supported by online search tools (see Chapter 4). This dataset contains 368 patient-authored blog posts and contain a wealth of information about those authors' experiences living with their condition. While our investigation focused on COPD exacerbations, there is potential for further investigation using this data source.

A subset of 100 of the blog posts collected has been coded and thematically analysed in terms of content related to COPD exacerbation experiences already (see Chapter 5). This marked up information could be utilised in subsequent qualitative analyses to discover more information about how these patients live with exacerbations. Similarly, the coding hierarchy produced as part of our qualitative analysis can be used in similar analyses of other datasets.

A subset of the blogs have also been annotated for semantic and sentiment information. Stand-off annotations of both exacerbation sentences and medical terms (Chapter 6) were added by subject-matter experts. Similarly, authoritative labels have been gathered for subjectivity or objectivity of the 1770 sentences that make up those 100 blogs through our crowdsourcing task (see Chapter 7). These labelled datasets offer potential for subsequent analysis of the blogs (e.g. easier access to relevant information, or further pattern mining – see Chapter 8) and refinement of methods to automatically garner this information from new unseen blog posts.

Lastly, we produced a vocabulary of over 3000 terms relating to COPD exacerbation curated from a combination of literature review, existing medical terminologies and the outcome of our qualitative analysis. This highly focused terminology can be used to drive further investigation in to COPD exacerbations and support text mining in future studies of patient experiences shared over social media.

9.4.3 TOOLS

As part of our work, we have developed tools to enable the gathering of information, the automatic extraction of information and to support its utility in other software packages. Firstly, as detailed in Chapter 4, we developed a program that would use RSS feeds to collect posts from pre-defined blogs before pre-processing and storing them in a relational database. This tool requires the URL of the RSS feed to be known before hand. We also produced a tool that would export blog posts and the additional information added through automatic tools to a standardised format which can be used in compatible software packages for use in further qualitative work.

Most importantly, we also developed tools which are used to automatically add the medical and sentiment information to patient blog posts. FlexiTerm, our novel approach to flexible automatic term recognition has been released as an open-source tool and is available for download[20]. FlexiTerm, although evaluated here on COPD blog posts, has been evaluated on a number of sources of medical text and has been shown to perform well compared to a baseline in each area. While useful for further work on patient social media, it also has the potential to be useful in other sources of informal or 'noisy' textual data.

We also produced our Naive Bayes model for labelling sentences as either subjective or objective according to linguistic properties (relative frequency of pronouns). This model can be used with the Weka package to help analyse new datasets for experience-related content.

9.5 LIMITATIONS OF OUR WORK

We have described how we addressed our research questions and the contributions that our work represents. Some limitations have been discussed, but it is important to set them out here. One main limitation is that our implementation of the QuTiP frameworks was focused towards one area of the medical domain – COPD. While many of our approaches have been developed to be applicable in other areas, further work would be required to show that this is the case. One exception to this, which has already been discussed, is FlexiTerm, which has been evaluated on documents pertaining to other areas of medicine.

Scalability is an important factor in automating this sort of analysis. However, access to authoritative data was a limiting factor in our case. We created annotated datasets of 100

[20] http://users.cs.cf.ac.uk/I.Spasic/flexiterm/

posts to develop and evaluate our methods. Wider evaluation on larger datasets would be advantageous in establishing their suitability, but beyond the scope of this project.

Lastly, our focus was on developing the QuTiP framework and tools in order to support the analysis of COPD blogs. The next step for this work would be to use the information garnered through automated analysis in order to carry out a wider analysis of the whole blog corpus. This should also be considered for future work relating to COPD experiences.

9.6 Future Research

Along with extensions of the work carried out in this thesis summarised in earlier sections, future research in the area of patient social media must focus on three questions. Firstly, what information can be gathered from patient blog posts? While we know that experiences and information are shared by patients, questions remain about the proportion of content those elements make up.

Experience is an important element and was particularly important to our investigation, but establishing what information is shared by patients is also an important area of enquiry. The second question relates to the quality of information shared in blog posts – are patients informing each other well? As mentioned at the beginning of this thesis – other patients' accounts of illness are an increasingly important source of information for individuals making health decisions of their own. That places added responsibility on health bloggers to make sure that their content is responsibly written. If content online cannot be trusted, then this poses a risk for others receiving the most up-to-date, recommended treatments. Issues relating to the quality of patient-authored information have already been asked by health informatics communities (Hesse et al., 2010) and work is underway to assess what is being shared. Greene et al. (2011), for instance, analysed the advice on management strategies shared by patients on diabetes groups on Facebook, finding that advice usually represented best practice, except where specific product recommendations were made. This may not always be the case in all health communities or on all media, and further work is required to establish this.

Lastly, we must establish how this information can be used. We can investigate the experiences shared through social media, but this will likely represent a small portion of the overall patient community. In our work, we have placed our investigation of social media content in terms of hypothesis generation, assuming that the questions raised will be investigated further through more traditional (but often more expensive) means. If the

representativeness of blogging communities can be shown to be much wider, then the implications of social media research can be made much broader.

Our work contributes to these three questions by giving a platform on which easier access and better summaries of the information contained within blogs can be provided. The next steps would be to widen the scope of the conditions and communities under investigation and perhaps focus not only on condition-specific information (i.e. patients' experiences) but other facets of bloggers and blogging. Social media has had an impact on many aspects of modern life, and is having an impact on how patients engage with health care. This form of communication and of cataloguing patients' self-prompted experiences offers a great opportunity to health researchers to discover new information about how patients live with and manage their conditions. It is only through investigating what is being shared and who exactly is sharing that this potential can be realised.

REFERENCES

Adams, A., Lomax, G., & Santarini, A. (2011). Social media & stem cell science: examining the discourse. *Regenerative medicine*, *6*(6 Suppl), 121–4. doi:10.2217/rme.11.82

Adams, R., Chavannes, N., Jones, K., Ostergaard, M. S., & Price, D. (2006). Exacerbations of chronic obstructive pulmonary disease--a patients' perspective. *Primary care respiratory journal : journal of the General Practice Airways Group*, *15*(2), 102–9. doi:10.1016/j.pcrj.2006.01.003

Adams, S., Pill, R., & Jones, A. (1997). Medication, chronic illness and identity: The perspective of people with asthma. *Social Science & Medicine*, *45*(2), 189–201. doi:10.1016/S0277-9536(96)00333-4

Atkinson, K. (2011). GNU Aspell. Retrieved February 09, 2014, from http://www.aspell.net

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co.

Barak, A., Boniel-Nissim, M., & Suler, J. (2008). Fostering empowerment in online support groups. *Computers in Human Behavior*, *24*(5), 1867–83. doi:10.1016/j.chb.2008.02.004

Barker, K. (2008). Electronic support groups, patient-consumers, and medicalization: the case of contested illness. *J Health Soc Behav*, *49*(1), 20–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18418983

Bashyam, V., & Taira, R. K. (2005). Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 26–30. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560562&tool=pmcentrez&rendertype=abstract

Bauman, A. E., Fardy, H. J., & Harris, P. G. (2003). Getting it right: why bother with patient-centred care? *The Medical journal of Australia*, *179*(5), 253–6. Retrieved from http://europepmc.org/abstract/MED/12924973

Berry, M. W., & Browne, M. (2005). Understanding Search Engines: Mathematical Modeling and Text Retrieval (Software, Environments, Tools), Second Edition. Retrieved from http://dl.acm.org/citation.cfm?id=1076286

Boote, J., Telford, R., & Cooper, C. (2002). Consumer involvement in health research: a review and research agenda. *Health policy (Amsterdam, Netherlands)*, *61*(2), 213–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12088893

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, *14*(1), 75–90. doi:10.1177/1354856507084420

Bradley, E. H., Curry, L. A., & Devers, K. J. (2007). Qualitative data analysis for health services research: developing taxonomy, themes, and theory. *Health services research*, *42*(4), 1758–72. doi:10.1111/j.1475-6773.2006.00684.x

Brédart, A., Marrel, A., Abetz-Webb, L., Lasch, K., & Acquadro, C. (2014). Interviewing to develop Patient-Reported Outcome (PRO) measures for clinical research: eliciting patients' experience. *Health and quality of life outcomes*, *12*, 15. doi:10.1186/1477-7525-12-15

British Thoracic Society. (2006). *Burden of Lung Disease* (2nd ed.). London: The British Thoracic Society. Retrieved from https://www.brit-thoracic.org.uk/document-library/delivery-of-respiratory-care/burden-of-lung-disease/burden-of-lung-disease-2006/

Bruce, R. F., & Wiebe, J. M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, *5*(2), 187–205. doi:10.1017/S1351324999002181

Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. *… Data with Amazon's Mechanical Turk*. Retrieved from http://dl.acm.org/citation.cfm?id=1866697

Cambria, E., Benson, T., Eckl, C., & Hussain, A. (2012). Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications*, *39*(12), 10533–10543. doi:10.1016/j.eswa.2012.02.120

Carnegie Mellon University. (2009). MinorThird. Retrieved from http://sourceforge.net/projects/minorthird/

Chapman, W. W., Dowling, J. N., & Wagner, M. M. (2004). Fever detection from free-text clinical records for biosurveillance. *Journal of biomedical informatics*, *37*(2), 120–7. doi:10.1016/j.jbi.2004.03.002

Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: What does it mean? (or it takes at least two to tango). *Social Science & Medicine*, *44*(5), 681–692. doi:10.1016/S0277-9536(96)00221-3

Chou, W.-Y. S., Hunt, Y., Folkers, A., & Augustson, E. (2011). Cancer survivorship in the age of YouTube and social media: a narrative analysis. *Journal of medical Internet research*, *13*(1), e7. doi:10.2196/jmir.1569

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29. Retrieved from http://dl.acm.org/citation.cfm?id=89086.89095

Clarke, A., Sohanpal, R., Wilson, G., & Taylor, S. (2010). Patients' perceptions of early supported discharge for chronic obstructive pulmonary disease: a qualitative study. *Quality & safety in health care*, *19*(2), 95–8. doi:10.1136/qshc.2007.025668

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

comScore. (2007). Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior. Retrieved September 08, 2013, from http://www.comscore.com/Insights/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior

Coulter, A., Peto, V., & Doll, H. (1994). Patients' Preferences and General Practitioners' Decisions in the Treatment of Menstrual Disorders. *Family Practice*, *11*(1), 67–74. doi:10.1093/fampra/11.1.67

Cowan, K. (2010). The James Lind alliance: tackling treatment uncertainties together. *The Journal of ambulatory care management*, *33*(3), 241–8. doi:10.1097/JAC.0b013e3181e62cda

Dafni, U., Tsiodras, S., Panagiotakos, D., Gkolfinopoulou, K., Kouvatseas, G., Tsourti, Z., & Saroglou, G. (2004). Algorithm for Statistical Detection of Peaks --- Syndromic Surveillance System for the Athens 2004 Olympic Games. *Morbidity and Mortality Weekly Report*, *53*, 86–94. Retrieved from http://origin.glb.cdc.gov/mmwr/preview/mmwrhtml/su5301a19.htm

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176. doi:10.1145/363958.363994

De Bra, P. M. E., & Post, R. D. J. (1994). Information retrieval in the World-Wide Web: Making client-based searching feasible. *Computer Networks and ISDN Systems*, *27*(2), 183–192. doi:10.1016/0169-7552(94)90132-5

Denecke, K. (2008). Accessing medical experiences and information. *… on Artificial Intelligence, Workshop on Mining Social …*. Retrieved from http://www.researchgate.net/publication/228497496_Accessing_Medical_Experiences_and_Information/file/50463526769e0e257c.pdf

Department of Health. (2004). Getting over the wall: How the NHS is improving the patient's experience. Department of Health, Richmond House, 79 Whitehall, London SW1A 2NJ, UK, dhmail@dh.gsi.gov.uk. Retrieved from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4090841

Department of Health. (2006). Our health, our care, our say: a new direction for community services (White Paper). Department of Health, Richmond House, 79 Whitehall, London SW1A 2NJ, UK, dhmail@dh.gsi.gov.uk. Retrieved from http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4127453

Department of Health. (2010). The NHS Constitution for England (2010 edition). Department of Health, Richmond House, 79 Whitehall, London SW1A 2NJ, UK, dhmail@dh.gsi.gov.uk. Retrieved from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_113613

Department of Health. (2011). Consultation on a strategy for services for chronic obstructive pulmonary disease (COPD) in England. Government response to the consultation. Department of Health, Richmond House, 79 Whitehall, London SW1A 2NJ, UK, dhmail@dh.gsi.gov.uk. Retrieved from http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Consultations/Responsestoconsultations/DH_127970

Dolan, J. C., Bordley, D. R., & Miller, H. (1993). Diagnostic strategies in the management of acute upper gastrointestinal bleeding. *Journal of General Internal Medicine*, *8*(10), 525–529. doi:10.1007/BF02599632

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2-3), 103–130. doi:10.1023/A:1007413511361

Doran, C., Griffith, J., & Henderson, J. (2006). Highlights from 12 months of collected blogs. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*.

Downe-Wamboldt, B. (1992). Content analysis: method, applications, and issues. *Health care for women international*, *13*(3), 313–21. doi:10.1080/07399339209516006

Duggan, M., & Smith, A. (2013). *Social Media Update 2013*. Retrieved from http://pewinternet.org/Reports/2013/Social-Media-Update.aspx

Durand, M.-A., Boivin, J., & Elwyn, G. (2012, September 11). Stakeholder field-testing of amnioDex, a person-centered decision support intervention for amniocentesis. *International Journal of Person Centered Medicine*. doi:10.5750/ijpcm.v2i3.271

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of advanced nursing*, *62*(1), 107–15. doi:10.1111/j.1365-2648.2007.04569.x

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC*.

Eysenbach, G. (2008). Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of medical Internet research2*, *10*(3), e22.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, March 15). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. doi:10.1609/aimag.v17i3.1230

Feldman, R., & Sanger, J. The Text Mining Handbook (2007). Cambridge University Press.

Fisher, J., & Clayton, M. (2012). Who gives a tweet: assessing patients' interest in the use of social media for health care. *Worldviews on evidence-based nursing / Sigma Theta*

Tau International, Honor Society of Nursing, *9*(2), 100–8. doi:10.1111/j.1741-6787.2012.00243.x

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Fox, S. (2011). The Social Life of Health Information, 2011. Retrieved June 14, 2011, from http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx

Fox, S., & Duggan, M. (2013). *Health Online 2013*. Retrieved from http://www.pewinternet.org/Reports/2013/Health-online.aspx

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, *3*(2), 115–130. doi:10.1007/s007999900023

Friedman, C. (2000). A broad-coverage natural language processing system. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 270–4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243979&tool=pmcentrez&rendertype=abstract

Friedman, C., Alderson, P. O., Austin, J. H. M., Cimino, J. J., & Johnson, S. B. (1994). A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, *1*(2), 161–174. doi:10.1136/jamia.1994.95236146

Friedman, Carol, Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association : JAMIA*, *11*(5), 392–402. doi:10.1197/jamia.M1552

Gadsby, R., Snow, R., Daly, A. C., Crowe, S., Matyka, K., Hall, B., & Petrie, J. (2012). Setting research priorities for Type 1 diabetes. *Diabetic medicine : a journal of the British Diabetic Association*, *29*(10), 1321–6. doi:10.1111/j.1464-5491.2012.03755.x

Gesteland, P. H., Wagner, M. M., Chapman, W. W., Espino, J. U., Tsui, F.-C., Gardner, R. M., … Haug, P. J. (2002). Rapid deployment of an electronic disease surveillance system in the state of Utah for the 2002 Olympic Winter Games. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 285–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2244330&tool=pmcentrez&rendertype=abstract

GOLD. (2014). *Pocket Guide to COPD Diagnosis, Management and Prevention: A Guide for Health Care Professionals*. Retrieved from http://www.goldcopd.org/uploads/users/files/GOLD_Pocket_2014_Jun11.pdf

Goldberg, D. (1978). *Manual of the General Health Questionnaire*. NFER Publishing, Windsor, England.

Green, J., & Britten, N. (1998). Qualitative research and evidence based medicine. *BMJ*, *316*(7139), 1230–1232. doi:10.1136/bmj.316.7139.1230

Greene, J. A., Choudhry, N. K., Kilabuk, E., & Shrank, W. H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of general internal medicine*, *26*(3), 287–92. doi:10.1007/s11606-010-1526-3

Greenwood, M., Elwyn, G., Francis, N., Preece, A., & Spasic, I. (2013). Automatic Extraction of Personal Experiences from Patients' Blogs: A Case Study in Chronic Obstructive Pulmonary Disease. In *2013 International Conference on Social Computing and Its Applications (SCA)* (pp. 377–382). Karlsruhe, Germany: IEEE. doi:10.1109/CGC.2013.66

Gruffydd-Jones, K., Langley-Johnson, C., Dyer, C., Badlan, K., & Ward, S. (2007). What are the needs of patients following discharge from hospital after an acute exacerbation of chronic obstructive pulmonary disease (COPD)? *Primary care respiratory journal : journal of the General Practice Airways Group*, *16*(6), 363–8. doi:10.3132/pcrj.2007.00075

Gysels, M., & Higginson, I. J. (2008). Access to services for patients with chronic obstructive pulmonary disease: the invisibility of breathlessness. *Journal of pain and symptom management*, *36*(5), 451–60. doi:10.1016/j.jpainsymman.2007.11.008

Habraken, J. M., Pols, J., Bindels, P. J. E., & Willems, D. L. (2008). The silence of patients with end-stage COPD: a qualitative study. *The British journal of general practice : the journal of the Royal College of General Practitioners*, *58*(557), 844–9. doi:10.3399/bjgp08X376186

Hafeez, R., Wagner, C. V, Smith, S., Boulos, P., Halligan, S., Bloom, S., & Taylor, S. A. (2012). Patient experiences of MR colonography and colonoscopy: a qualitative study. *The British journal of radiology*, *85*(1014), 765–9. doi:10.1259/bjr/36231529

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1).

Hand, D. J. (1981). Discrimination and classification. *Wiley Series in Probability and Mathematical Statistics*. Retrieved from http://adsabs.harvard.edu/abs/1981dicl.book

Hares, T., Spencer, J., Gallagher, M., Bradshaw, C., & Webb, I. (1992). Diabetes care: who are the experts? *Quality and Safety in Health Care*, *1*(4), 219–224. doi:10.1136/qshc.1.4.219

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). *Bridging the gap: a genre analysis of Weblogs*. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (p. 11 pp.). IEEE. doi:10.1109/HICSS.2004.1265271

Hesse, B. W., Hansen, D., Finholt, T., Munson, S., Kellogg, W., & Thomas, J. C. (2010). Social Participation in Health 2.0. *Computer*, *43*(11), 45–52. doi:10.1109/MC.2010.326

Hewitt-Taylor, J., & Bond, C. S. (2012). What E-patients Want From the Doctor-Patient Relationship: Content Analysis of Posts on Discussion Boards. *Journal of medical Internet research*, *14*(6), e155. doi:10.2196/jmir.2068

Hibbard, J. H. (2003). Engaging health care consumers to improve the quality of care. *Medical care*, *41*(1 Suppl), I61–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12544817

Higgins, J., & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. Retrieved from www.cochrane-handbook.org

Hoch, D., & Ferguson, T. (2005). What I've learned from E-patients. *PLoS Med*, *2*(8), e206. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16060721

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15*(9), 1277–88. doi:10.1177/1049732305276687

Hughes, B., Joshi, I., & Wareham, J. (2008). Health 2.0 and Medicine 2.0: tensions and controversies in the field. *Journal of medical Internet research*, *10*(3), e23. doi:10.2196/jmir.1056

Jansen, J. (2010). Online Product Research - Pew Research Center's Internet & American Life Project. Retrieved from http://www.pewinternet.org/Reports/2010/Online-Product-Research.aspx

Jazzy. (2013). Jazzy. Retrieved February 09, 2014, from http://sourceforge.net/projects/jazzy/

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification, 151–160. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002492

Jones, P W, Harding, G., Berry, P., Wiklund, I., Chen, W.-H., & Kline Leidy, N. (2009). Development and first validation of the COPD Assessment Test. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*, *34*(3), 648–54. doi:10.1183/09031936.00102509

Jones, Paul W, Chen, W.-H., Wilcox, T. K., Sethi, S., & Leidy, N. K. (2011). Characterizing and quantifying the symptomatic features of COPD exacerbations. *Chest*, *139*(6), 1388–94. doi:10.1378/chest.10-1240

Justeson, J. S., & Katz, S. M. (2008). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(01), 9–27. doi:10.1017/S1351324900000048

Kageura, K., & Umino, B. (1996). Methods of Automatic Term Recognition - A Review. *Terminology*, *3*(2), 259–289.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. Retrieved from http://www.sciencedirect.com/science/article/pii/S0007681309001232

Keim-Malpass, J., & Steeves, R. H. (2012). Talking with death at a diner: young women's online narratives of cancer. *Oncology nursing forum*, *39*(4), 373–8, 406. doi:10.1188/12.ONF.373-378

Kessler, R., Ståhl, E., Vogelmeier, C., Haughney, J., Trudeau, E., Löfdahl, C.-G., & Partridge, M. R. (2006). Patient understanding, detection, and experience of COPD exacerbations: an observational, interview-based study. *Chest*, *130*(1), 133–42. doi:10.1378/chest.130.1.133

Kita, K., Kato, Y., Omoto, T., & Yano, Y. (1994). A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, *1*, 21–33.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 453). New York, New York, USA: ACM Press. doi:10.1145/1357054.1357127

Kitzinger, J. (1995). Qualitative Research: Introducing focus groups. *BMJ*, *311*(7000), 299–302. doi:10.1136/bmj.311.7000.299

Kondracki, N. L., Wellman, N. S., & Amundson, D. R. (2002). Content Analysis: Review of Methods and Their Applications in Nutrition Education. *Journal of Nutrition Education and Behavior*, *34*(4), 224–230. doi:10.1016/S1499-4046(06)60097-3

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of biomedical informatics*, *37*(6), 512–26. doi:10.1016/j.jbi.2004.08.004

Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, *30*(1), 61–70. doi:10.1177/001316447003000105

Krippendorff, Klaus. (2013). Computing Krippendorff's Alpha-Reliability. Retrieved from http://www.asc.upenn.edu/usr/krippendorff/mwebreliability5.pdf

Krumholz, H. M., Bradley, E. H., & Curry, L. A. (2013). Promoting publication of rigorous qualitative research. *Circulation. Cardiovascular quality and outcomes*, *6*(2), 133–4. doi:10.1161/CIRCOUTCOMES.113.000186

Lakshmanan, G. T., & Oberhofer, M. A. (2010). Knowledge Discovery in the Blogosphere: Approaches and Challenges. *IEEE Internet Computing*, *14*(2), 24–32. doi:10.1109/MIC.2010.26

Lieberman, M. (2007). The role of insightful disclosure in outcomes for women in peer-directed breast cancer groups: a replication study. *Psycho-oncology*, *16*(10), 961–4. doi:10.1002/pon.1181

Lippincott, T. (2008). agreement.py. Retrieved from http://cswww.essex.ac.uk/Research/nle/arrau/Lippincott/

Lombardo, J., Sniegoski, C., Loschen, W., Westercamp, M., Wade, M., Dearth, S., & Zhang, and G. (2008). Public Health Surveillance for Mass Gatherings. *Johns Hopkins APL Technical Digest2*, 347–355. Retrieved from http://www.jhuapl.edu/techdigest/TD/td2704/LombardoMassGatherings.pdf

Lophatananon, A., Tyndale-Biscoe, S., Malcolm, E., Rippon, H. J., Holmes, K., Firkins, L. A., … Muir, K. R. (2011). The James Lind Alliance approach to priority setting for prostate cancer research: an integrative methodology based on patient and clinician participation. *BJU international*, *108*(7), 1040–3. doi:10.1111/j.1464-410X.2011.10609.x

Lowney, A., & O'Brien, T. (2011). The landscape of blogging in palliative care. *Palliative Medicine*, *0*(0), 1–2.

M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, M. G. (2000). Focused crawling using context graphs. In *Proc. of the 26th International Conference on Very Large Databases (VLDB)* (pp. 527–534). Cairo, Egypt. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.9226

MacNee, W., & Rennard, S. (2004). *Fast Facts: Chronic Obstructive Pulmonary Disease*. Health Press Limited.

Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: a structured review. *Journal of evaluation in clinical practice*, *12*(5), 559–68. doi:10.1111/j.1365-2753.2006.00650.x

McCray, A. T., Srinivasan, S., & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 235–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247735&tool=pmcentrez&rendertype=abstract

Meadows, K. A. (2011). Patient-reported outcome measures: an overview. *British journal of community nursing*, *16*(3), 146–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21378658

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 128–44. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18660887

Michaels, C., & Meek, P. M. (2004). The language of breathing among individuals with chronic obstructive pulmonary disease. *Heart & lung : the journal of critical care*, *33*(6), 390–400. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15597293

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon, 26–34. Retrieved from http://portal.acm.org/citation.cfm?id=1860631.1860635

Murthy, D. (2012). Towards a Sociological Understanding of Social Media: Theorizing Twitter. *Sociology*, *46*(6), 1059–1073. doi:10.1177/0038038511422553

Nenadić, G., Spasić, I., & Ananiadou, S. (2002). Automatic acronym acquisition and term variation management within domain specific texts. In *Proceedings of the 3rd International Conference on Language, Resources, and Evaluation (LREC-3)* (pp. 2155–

2162). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.137.3864

NICE. (2004). CHRONIC OBSTRUCTIVE PULMONARY DISEASE - National clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and secondary care. *Thorax*, *59*((Suppl I)), 1–232.

NLM. (2012). Unified Medical Language System. Retrieved July 17, 2012, from http://www.nlm.nih.gov/research/umls/

NM Incite. (2011). Healthcare Social Media by the Numbers. Retrieved July 16, 2012, from http://nmincite.com/healthcare-social-media-by-the-numbers/

O'Neil, M., Payne, C., & Read, J. (1995). Read codes version 3: a user led terminology. *Methods of Information in Medicine1*, *34*, 187–192.

O'Neill, E. S. (2002). Illness representations and coping of women with chronic obstructive pulmonary disease: a pilot study. *Heart & lung : the journal of critical care*, *31*(4), 295–302. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12122393

O'Reilly, T. (2005). What is Web 2.0? Retrieved July 17, 2012, from http://oreilly.com/web2/archive/what-is-web-20.html

Ojoo, J. C., Moon, T., McGlone, S., Martin, K., Gardiner, E. D., Greenstone, M. A., & Morice, A. H. (2002). Patients' and carers' preferences in two models of care for acute exacerbations of COPD: results of a randomised controlled trial. *Thorax*, *57*(2), 167–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1746235&tool=pmcentrez&rendertype=abstract

Ong, B. N. (1996). The lay perspective in health technology assessment. *International journal of technology assessment in health care*, *12*(3), 511–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8840670

Pang, B., & Lee, L. (2004). A sentimental education. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04* (p. 271–es). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1218955.1218990

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. doi:10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02* (Vol. 10, pp. 79–86). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1118693.1118704

PatientsLikeMe. (2015). Background | PatientsLikeMe. Retrieved May 24, 2014, from https://www.patientslikeme.com/

Patton, M. Q. (2014). *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (Vol. 18, p. 832). SAGE Publications. Retrieved from https://books.google.co.uk/books/about/Qualitative_Research_Evaluation_Methods.html?id=-CM9BQAAQBAJ&pgis=1

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., … Brew, C. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, *2012*(Suppl. 1), 3–16. Retrieved from http://www.la-press.com/sentiment-analysis-of-suicide-notes-a-shared-task-article-a3016

Petit-Zeman, S., & Uhm, S. (2012). Identifying research priorities in preterm birth. *Infant*, *8*(3), 71–72.

Philips, L. (1990). Hanging on the Metaphone. *Computer Language*, *7*(12 (December)).

Pilling, A., Bassett, C., & Wolstenholme, R. J. (2003). A nurse-led service for acute exacerbation of COPD. *Nursing times*, *99*(26), 32–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12875115

Pinnock, H., Kendall, M., Murray, S. A., Worth, A., Levack, P., Porter, M., … Sheikh, A. (2011). Living and dying with severe chronic obstructive pulmonary disease: multi-perspective longitudinal qualitative study. *BMJ (Clinical research ed.)*, *342*, d142. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3025692&tool=pmcentrez&rendertype=abstract

Pollock, A., St George, B., Fenton, M., & Firkins, L. (2012). Top ten research priorities relating to life after stroke. *Lancet neurology*, *11*(3), 209. doi:10.1016/S1474-4422(12)70029-7

Pollock, A., St George, B., Fenton, M., & Firkins, L. (2014). Top 10 research priorities relating to life after stroke--consensus from stroke survivors, caregivers, and health professionals. *International journal of stroke : official journal of the International Stroke Society*, *9*(3), 313–20. doi:10.1111/j.1747-4949.2012.00942.x

Poses, R. M., & Isen, A. M. (1998). Qualitative research in medicine and health care: questions and controversy. *Journal of general internal medicine*, *13*(1), 32–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1496891&tool=pmcentrez&rendertype=abstract

Princeton University. (2013). About Wordnet. Retrieved December 28, 2013, from http://wordnet.princeton.edu/

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. doi:10.1109/MASSP.1986.1165342

Rankin, D., Barnard, K., Elliott, J., Cooke, D., Heller, S., Gianfrancesco, C., … Lawton, J. (2014). Type 1 diabetes patients' experiences of, and need for, social support after attending a structured education programme: a qualitative longitudinal investigation. *Journal of clinical nursing*. doi:10.1111/jocn.12539

Rashtchian, C., & Young, P. (2010). Collecting image annotations using Amazon's Mechanical Turk. … *Mechanical Turk*. Retrieved from http://dl.acm.org/citation.cfm?id=1866717

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning From Crowds. *The Journal of Machine Learning Research*, *11*, 1297–1297–1322–1322. Retrieved from http://dl.acm.org/citation.cfm?id=1756006.1859894

Reinke, L. F., Engelberg, R. A., Shannon, S. E., Wenrich, M. D., Vig, E. K., Back, A. L., & Curtis, J. R. (2008). Transitions regarding palliative and end-of-life care in severe chronic obstructive pulmonary disease or advanced cancer: themes identified by patients, families, and clinicians. *Journal of palliative medicine*, *11*(4), 601–9. doi:10.1089/jpm.2007.0236

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* - (Vol. 10, pp. 105–112). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1119355.1119369

Riloff, E., Wiebe, J., & Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction, 1106–1111. Retrieved from http://dl.acm.org/citation.cfm?id=1619499.1619511

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27*(3), 129–146. doi:10.1002/asi.4630270302

Rodgers, S., Dyas, J., Molyneux, A. W. P., Ward, M. J., & Revill, S. M. (2007). Evaluation of the information needs of patients with chronic obstructive pulmonary disease following pulmonary rehabilitation: a focus group study. *Chronic respiratory disease*, *4*(4), 195–203. doi:10.1177/1479972307080698

Rogers, F. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, *51*, 114–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=197951&tool=pmcentrez&rendertype=abstract

Rozmovits, L., & Ziebland, S. (2004). What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs. *Patient Educ Couns.*, *53*(1), 57–64.

Salton, G. (1971). The SMART Retrieval System—Experiments in Automatic Document Processing. Retrieved from http://dl.acm.org/citation.cfm?id=1102022

Salton, G., & Mcgill, M. (1986). Introduction to Modern Information Retrieval.

Sarkar, I. N. (2010). Biomedical informatics and translational medicine. *Journal of translational medicine*, *8*(1), 22. doi:10.1186/1479-5876-8-22

Schler, J. (2005). The importance of neutral examples for learning sentiment. In *Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*

(FINEXIN). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.6586

Schofield, I., Knussen, C., & Tolson, D. (2006). A mixed method study to compare use and experience of hospital care and a nurse-led acute respiratory assessment service offering home care to people with an acute exacerbation of chronic obstructive pulmonary disease. *International journal of nursing studies*, *43*(4), 465–76. doi:10.1016/j.ijnurstu.2005.07.002

Schölkopf, B., Platt, J., & Hofmann, T. (2006). Isotonic Conditional Random Fields and Local Sentiment Flow. *Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference*, 961–968. Retrieved from http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6287464

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.

Shahab, L., Jarvis, M. J., Britton, J., & West, R. (2006). Prevalence, diagnosis and relation to tobacco dependence of chronic obstructive pulmonary disease in a nationally representative population sample. *Thorax*, *61*(12), 1043–7. doi:10.1136/thx.2006.064410

Shuval, K., Harker, K., Roudsari, B., Groce, N. E., Mills, B., Siddiqi, Z., & Shachak, A. (2011). Is qualitative research second class science? A quantitative longitudinal examination of qualitative research in medical journals. *PloS one*, *6*(2), e16937. doi:10.1371/journal.pone.0016937

Sibanda, T., He, T., Szolovits, P., & Uzuner, O. (2006). Syntactically-informed semantic category recognition in discharge summaries. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 714–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839398&tool=pmcentrez&rendertype=abstract

Sillence, E., & Mo, P. K. H. (2012). Communicating health decisions: an analysis of messages posted to online prostate cancer forums. *Health expectations : an international journal of public participation in health care and health policy*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22296292

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *Urbana*. Retrieved from http://luci.ics.uci.edu/websiteContent/weAreLuci/biographies/faculty/djp3/LocalCopy/IEEEXplore(3)0.pdf

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, *6*(3), 239–251. doi:10.1093/bib/6.3.239

Spasić, I., Burnap, P., Greenwood, M., & Arribas-Ayllon, M. (2012). A naïve bayes approach to classifying topics in suicide notes. *Biomedical informatics insights*, *5*(Suppl. 1), 87–97. doi:10.4137/BII.S8945

Spasić, I., Greenwood, M., Preece, A., Francis, N., & Elwyn, G. (2013). FlexiTerm: a flexible term recognition method. *Journal of Biomedical Semantics*, *4*(1), 27. doi:10.1186/2041-1480-4-27

Spyns, P. (1996). Natural language processing in medicine: an overview. *Methods of Information in Medicine*, *35*, 285–301.

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association : JAMIA*, *17*(6), 646–51. doi:10.1136/jamia.2009.001024

Stewart, M. (2001). Towards a global definition of patient centred care. *BMJ*, *322*(7284), 444–445. doi:10.1136/bmj.322.7284.444

Su, F., & Markert, K. (2009). Subjectivity recognition on word senses via semi-supervised mincuts, 1–9. Retrieved from http://dl.acm.org/citation.cfm?id=1620754.1620756

Swanson, D., & Smalheiser, N. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, *91*(2), 183–203. doi:10.1016/S0004-3702(97)00008-8

Tan, A. (1999). Text Mining: the state of the art and the challenges. In *Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD '99 Workshop on Knowledge Discovery from Advanced Databases* (pp. 65–70).

The Nielsen Company. (2012). CONSUMER TRUST IN ONLINE, SOCIAL AND MOBILE ADVERTISING GROWS. Retrieved from http://www.nielsen.com/us/en/newswire/2012/consumer-trust-in-online-social-and-mobile-advertising-grows.html

Thomas, L. A. (2009). Effective dyspnea management strategies identified by elders with end-stage chronic obstructive pulmonary disease. *Applied nursing research : ANR*, *22*(2), 79–85. doi:10.1016/j.apnr.2007.04.010

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* (Vol. 1, pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1073445.1073478

Tsui, F.-C., Espino, J. U., Dato, V. M., Gesteland, P. H., Hutman, J., & Wagner, M. M. (2003). Technical description of RODS: a real-time public health surveillance system. *Journal of the American Medical Informatics Association : JAMIA*, *10*(5), 399–408. doi:10.1197/jamia.M1345

Tyndale-Biscoe, S., Malcolm, E., & Gnanapragasam, V. J. (2012). Setting priorities for prostate cancer research. *Trends in Urology & Men's Health*, *3*(1), 31–33. doi:10.1002/tre.244

Ulicny, B., Baclawski, K., & Magnus, A. (2007). New metrics for blog mining. In *SPIE Defense & Security Symposium*.

Valitutti, R., & Stock, O. (2004). Developing Affective Lexical Resources. *Psychology*, *2*, 61–83. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.2710

Van De Belt, T. H., Engelen, L. J. L. P. G., Berben, S. A. A., & Schoonhoven, L. (2010). Definition of Health 2.0 and Medicine 2.0: a systematic review. *Journal of medical Internet research*, *12*(2), e18. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2956229&tool=pmcentrez&rendertype=abstract

Vestbo, J., Hurd, S. S., Agustí, A. G., Jones, P. W., Vogelmeier, C., Anzueto, A., … Rodriguez-Roisin, R. (2013). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Retrieved from http://www.goldcopd.org/uploads/users/files/GOLD_Report_2013_Feb20.pdf#26

Vuurens, J. B. P., & de Vries, A. P. (2012). Obtaining High-Quality Relevance Judgments Using Crowdsourcing. *IEEE Internet Computing*, *16*(5), 20–27. doi:10.1109/MIC.2012.71

Walters, J. A., Hansen, E. C., Walters, E. H., & Wood-Baker, R. (2008). Under-diagnosis of chronic obstructive pulmonary disease: a qualitative study in primary care. *Respiratory medicine*, *102*(5), 738–43. doi:10.1016/j.rmed.2007.12.008

Wang, B., Spencer, B., Ling, C. X., & Zhang, H. (2008). Semi-supervised Self-training for Sentence Subjectivity Classification. *Advances in Artificial Intelligence*, *5032*, 344–355. doi:10.1007/978-3-540-68825-9_32

White, T. (2004). Can't beat Jazzy. Retrieved February 09, 2014, from http://web.archive.org/web/20130509121337/http://www.ibm.com/developerworks/java/library/j-jazzy/

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., … Heywood, J. (2010). Sharing health data for better outcomes on PatientsLikeMe. *Journal of medical Internet research*, *12*(2), e19. doi:10.2196/jmir.1549

Wiebe, J. (2000). Learning Subjective Adjectives from Corpora. In *AAAI/{IAAI}* (pp. 735–740). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.2615

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, *20*(2), 233–287. Retrieved from http://dl.acm.org/citation.cfm?id=972525.972529

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - (pp. 246–253). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1034678.1034721

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2006). Learning Subjective Language. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/0891201041850885

Wilson, J. (1999). Acknowledging the expertise of patients and their organisations. *BMJ (Clinical research ed.)*, *319*(7212), 771–4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1116603&tool=pmcentrez&rendertype=abstract

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses, 761–767. Retrieved from http://dl.acm.org/citation.cfm?id=1597148.1597270

World Health Organization. (2010). *International Statistical Classification of Diseases and Related Health Problems* (4th ed.). Geneva, Switzerland: WHO Press. Retrieved from http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf?ua=1

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization, 412–420. Retrieved from http://dl.acm.org/citation.cfm?id=645526.657137

Yin, R. K. (2003). *Case Study Research: Design and Methods*. Retrieved from https://books.google.co.id/books/about/Case_study_research.html?id=BWea_9ZGQMwC&pgis=1

Zhao, Y., & Zhu, Q. (2012). Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*. doi:10.1007/s10796-012-9350-4

APPENDIX A – DATA SOURCES

Blog	Included	No RSS/ access	Doctor/ company	Marketing/ SEO
http://talk.nhs.uk/blogs/copd/default.aspx	x			
http://livingwithcopd.org/	x			
http://kwrenbscopdnewsoftheday.blogspot.com/	x			
http://livingoutloudwithcopd.blogspot.com/	x			
http://www.copdlighthouse.blogspot.com/	x			
http://desperatedanandco.blogspot.com/	x			
http://copdandcheryl.blogspot.com/	x			
http://quitterswin.wordpress.com/	x			
http://ifitistobeitsuptomelivingwithcopd.blogspot.com/	x			
http://breathingbetterlivingwell.blogspot.com/	x			
http://sandyellen.blogspot.com/	x			
http://www.mylifewithemphysema.blogspot.com/	x			
http://copdnewsoftheday.com/		x		
http://copdandlvrs.blogspot.com/		x		
http://bitz-and-bobz.co.uk/copd/		x		
http://copd-ourconcerns.blogspot.com/		x		
http://pugetsoundblogs.com/copd-and-other-stuff/		x		
http://www.copdcoalition.eu/page/12			x	
http://coalhealthstudy.org/			x	
http://www.copd-international.com/COPD-Info/			x	
http://crowtrees.wordpress.com/			x	
http://allergydocblog.com/			x	
http://cascadestudy.org/			x	
http://blog.ctnews.com/bridgeporthospital/			x	
http://copddoclondon.blogspot.co.uk/			x	
http://www.discoverymedicine.com/Edwin-K-Silverman/			x	
http://www.alltreatment.com/blog/2011/dr-mintz-on-e-cigarettes/			x	
http://blog.careandcompliance.com/			x	
http://www.copdicd9.com/			x	
http://www.healthoxygen.com/blog/			x	
http://drwakde.wordpress.com/			x	
http://dontforgettobreathe.org.nz/				x
http://www.chronicdiseasepreparedness.org/				x

http://copdtreatmentguidelines.blogspot.co.uk/	x
http://www.airpurifiersource.com/blog/page/3/	x
http://copdexacerbationinfo.com/	x
http://lungdiseases.co.uk/	x
http://copdprognosis.net/	x
http://erickclarksville.blogspot.co.uk/	x
http://stevensjohnsonsyndrome.clarislaw.com/	x
http://spiropd.com [*** WARNING: THIS WEBSITE NOW CONTAINS MALWARE ***]	x
http://healthylungs.info/	x
http://www.alltreatment.com/blog/	x
http://copdlungdisease.com/	x
http://respiratory-care-sleep-medicine.advanceweb.com/	x
http://www.emphysemastages.com/	x
http://mriser.wordpress.com/	x
http://findhomeremediesnow.com/12/	x
http://emphysemasymptoms0.wordpress.com/	x
http://www.smokingsideeffects.net/	x

APPENDIX B – COMPLETE INITIAL CODING HIERARCHY

				docs	ref
Admission				1	1
Advice				25	43
Antibiotics				9	11
Anxiety				4	5
Appointment				19	33
Behaviour modification				1	1
Blogging				14	15
Carers				1	2
Communication				3	4
Comorbidities				6	8
Copied external content				17	17
Decision				1	2
Diagnosis				7	7
Difficulty				22	39
Disability				10	19
Exacerbation				42	74
	General comment			2	3
	Physical exertion			6	6
		Impact		2	2
	Prevention			0	0
		Medication		3	4
		Avoiding physical activity		3	3
		Slowing down		1	1
		Exercise		5	5
		Improve air quality		1	2
		Avoid risks		2	2
	Infection			4	4
		Discovering exacerbation		11	15
			Patient notices change in condition	7	11
			Diagnosis at doctor's appointment	3	3
			Emergency hospital (ambulance)	1	1
		Treating Exacerbation		8	15
			Vitamin D for exacerbation prevention	2	2

			Sleeping in a chair for comfortable breathing	1	2
			Prescribed antibiotics	2	2
			Home supply antibiotics	3	3
			Prescribed steroids	3	3
			Prescribed nebulizer capsules	1	1
			Decline steroids or antibiotics	1	1
			Over the counter cold remedies	1	1
		Monitoring Symptoms		11	39
			Sleep disturbance sign of exacerbation	2	2
			Phlegm	2	2
			Vigilance of Infections Close by	2	3
			Tiredness	1	2
			Pain	2	3
			Temperature	1	1
			Waiting for chest symptoms before treating	5	8
			Breathing	5	8
			Weather as cause of SOB	1	1
			Heart rate	1	3
			Oxygen levels	1	3
			Waiting too long for treatment	1	1
			Upper respiratory symptoms	1	1
			Cough	1	1
		Impact of exacerbation		5	8
			Disturbance to work	2	3
			Hospitalization	3	4
			Miserable	1	1
		Issues		3	3
			Availability of medication	2	2
			Medication	1	1

			side-effects		
		Self-diagnosis		1	1
	Personal limitations			4	4
	Learning techniques			3	4
	Weather_air quality			10	14
	Sharing information			6	7
	Unpredictable			1	1
Exercise				8	11
Fear				1	1
Fit to fly				2	2
Frustration				3	6
History				1	1
Impact				8	13
Information seeking				16	29
Interesting				1	1
Lessons				2	2
Medical professionals				13	24
Needs				2	2
Negativity				8	9
Positivity				23	46
Public perception				11	15
Reassurance				1	3
Recommendation				2	2
Rehab				19	34
Results				1	2
Sadness				1	1
Self-care				5	5
Side effects				6	6
Smoking				8	16
Steroids				5	5
Support				5	8
Support network				8	10
Symptoms				3	3
Tests				17	24
Treatment				23	32
Treatment choice				8	14
Trials				1	1
Uncertainty				1	2
Vitamin D				2	3
Weather				14	27

APPENDIX C - FLEXITERM OUTPUT FOR PATIENT BLOG POST CORPUS

Rank	Term variants	Score	Classification
1	breathe easy	45.7862	Support network
	easy breathing		
2	pulmonary rehab	45.0546	Treatment
	pulmonary rehab		
3	lung disease	35.2071	Disease
	lung diseases		
	diseased lungs		
4	chronic disease	32.6683	General concept
5	chronic obstructive pulmonary disease	27.7259	Disease
6	vitamin d	25.5834	Treatment
	d vitamin		
7	lung function	24.6067	Functional concept
8	quality of life	21.4876	Functional concept
9	shortness of breath	20.1013	Symptom
10	chest infection	19.4081	Infection
	chest infections		
11	chronic bronchitis	18.715	Disease
12	breathe easy groups	18.4567	Support network
	breathe easy group		
	breathe easy groups		
13	lung transplant	17.6257	Treatment
	lung transplantation		
	lung transplants		
	lung transplantations		
14	blood pressure	17.3287	Tests
14	copd patients	17.3287	Patients
	patients with copd		
	copd patient		
15	chronic lung disease	16.3793	Disease
16	pulmonary disease	13.9785	Disease
17	british lung foundation	13.1833	Support network
18	stem cells	13.0882	Treatment
19	easy group	13.0543	Support network
20	support groups	12.8232	Support network
	support group		
21	people with copd	12.4766	Patients
22	deep breath	11.0904	Exercise
	deep breaths		
	deep breathing		
22	lung cancer	11.0904	Comorbidities
	lung cancers		
22	rehab course	11.0904	Rehab

	rehab courses		
	course of rehab		
23	pulmonary rehab course	10.9861	Rehab
	pulmonary rehab courses		
24	blood test	10.3202	Tests
	blood tests		
25	laughter yoga	10.3106	Treatment
26	swine flu	9.7041	Infection
26	immune system	9.7041	General concept
	immune systems		
27	cigarette smoking	9.0109	Smoking
	cigarette smoke		
28	pulmanory rehabillitation	8.8723	Rehab
	pulmonary rehabilitation		
29	copd exacerbation	8.7799	Exacerbation
	copd exacerbations		
	exacerbation of copd		
	exacerbations of copd		
30	copd sufferers	7.6246	Patients
	copd sufferer		
	consideration for copd sufferers		
	suffering from copd		
30	flu shot	7.6246	Treatment
	flu shots		
31	health care	7.4513	General concept
32	respiratory infections	7.3936	Infection
	respiratory infection		
33	pulmonary program	7.047	Treatment
34	local breathe easy group	6.9315	Support network
	local breath easy group		
	local breathe easy groups		
34	breathing exercises	6.9315	Exercise
	breathing excercise		
	breathing exercise		
34	vitamin c	6.9315	Treatment
	amounts of vitamin c		
	vitamins c		
34	symptoms of copd	6.9315	Symptom
	copd symptoms		
35	cases of swine flu	6.5917	Infection
	case of swine flu		
	swine flu case		
35	pulmonary function tests	6.5917	Tests
	pulmonary function test		
	pulmonary functions tests		

36	air pollution	6.2383	Exacerbation
	air of pollutants		
	air pollutants		
36	lung capacity	6.2383	Functional concept
36	world copd	6.2383	
36	heart disease	6.2383	Comorbidities
36	vitamin e	6.2383	Treatment
	e. levels of vitamins		
36	warmer weather	6.2383	Exacerbation
36	rehab room	6.2383	Treatment
37	clinical trial	6.1393	General concept
38	puppy mills	6.0073	Noise
	puppy mill		
39	copd blog	5.5452	Support network
39	lung damage	5.5452	
39	exercise program	5.5452	
	exercise programs		
39	healthcentral network	5.5452	
40	american lung association	5.4931	
40	breathe easy meeting	5.4931	
	breathe easy meetings		
40	lung function test	5.4931	
	lung function tests		
40	pulmonary rehab program	5.4931	
	pulmonary rehab programs		
41	rehabilitation program	5.4065	
42	fresh air	5.3141	
43	obstructive lung disease	5.2184	
44	immune response	5.0253	
45	antitrypsin deficiency	4.852	
	antitrysin deficiency		
45	copd awareness	4.852	
	awareness for copd		
45	blue badge	4.852	
45	lung condition	4.852	
	lung conditions		
45	copd nurse	4.852	
45	heart failure	4.852	
45	family members	4.852	
	family member		
45	lung health	4.852	
45	health professional	4.852	
	health professionals		
45	heart rate	4.852	
45	practice nurse	4.852	

	practise nurse	
45	rehab team	4.852
46	viral infection	4.5055
	viral infections	
47	faster heart beat	4.3944
47	bone density test	4.3944
47	double lung transplant	4.3944
47	pulmonary rehabilitation program	4.3944
	pulmonary rehabilitation programs	
48	action plan	4.1589
	action planning	
48	hospital admission	4.1589
	hospital admissions	
48	bad weather	4.1589
48	sweetest blessings	4.1589
48	breathe easy support group	4.1589
	breath easy support group	
48	flu cases	4.1589
48	chest x-ray	4.1589
	chest x-rays	
48	finger clubbing	4.1589
48	person with copd	4.1589
	copd person	
	personal copd	
48	risk of copd	4.1589
	risk for copd	
48	disease progression	4.1589
	progressive disease	
	progress of disease	
48	flu jab	4.1589
	flu jabs	
48	local group	4.1589
	local groups	
48	heart problem	4.1589
	heart problems	
48	local hospital	4.1589
48	rescue inhaler	4.1589
	rescue inhalers	
48	millions of people	4.1589
48	telehealth program	4.1589
49	asthma patient	3.9856
	patients with asthma	
50	lung tissue	3.9278
51	bone density	3.8123
51	respiratory team	3.8123

52	physical activities	3.4657
	physical activity	
52	air passages	3.4657
	air passage	
52	air purifier	3.4657
	air purifiers	
52	alpha-1 antitrypsin	3.4657
52	course of antibiotics	3.4657
	courses of antibiotics	
	courses of different antibiotics	
52	metro atlanta	3.4657
	metro atlantans	
52	heart attacks	3.4657
	heart attack	
52	blood work	3.4657
52	difficulty breathing	3.4657
	breathing difficulties	
	breathing difficulty	
52	breathing problems	3.4657
52	breathing test	3.4657
	breathing tests	
52	spiriva capsules	3.4657
	spiriva capsule	
52	chronic obstructive lung disease	3.4657
52	cold weather	3.4657
52	nerve damage	3.4657
52	toronto general	3.4657
52	kasper dl	3.4657
52	generic drugs	3.4657
52	emergency room	3.4657
52	strength endurance	3.4657
52	physical exam	3.4657
52	physical examination	3.4657
52	facebook friends	3.4657
	facebook friend	
	friend on facebook	
	friends on facebook	
52	risk factors	3.4657
	risk factor	
52	flu season	3.4657
52	healthy person	3.4657
52	hospital stay	3.4657
	stay in hospital	
52	lung infections	3.4657
	lung infection	

52	lungs of people	3.4657
52	lung surgery	3.4657
52	o2 sats	3.4657
52	oxygen therapy	3.4657
	oxygen oxygen therapy	
52	respiratory therapist	3.4657
	respiratory therapists	
53	omega-3 fatty acids	3.2958
53	emergency supply of antibiotics	3.2958
	emergency weeks supply of antibiotics	
53	emergency information card	3.2958
	emergency information cards	
53	intensive care unit	3.2958
53	stem cell network	3.2958
53	centers for disease control	3.2958
53	pulmonary rehab class	3.2958
	pulmonary rehab classes	
53	mucus clearing device	3.2958
	mucus clearing devices	
53	toronto general hospital	3.2958
53	jewish general hospital	3.2958
53	world health organization	3.2958
53	type of immune response	3.2958
53	nicotine replacement therapy	3.2958
53	specialist respiratory team	3.2958
	specalised respiratory team	
54	inflammatory disease	3.2924
54	public health	3.2924
55	fatty acid	3.0036
56	retinoic acid	2.7726
56	beta-2 agonist	2.7726
	beta-2 agonists	
56	cold air	2.7726
56	air sacs	2.7726
56	health alert	2.7726
56	right arm	2.7726
56	breathing techniques	2.7726
	techniques on breathing	
56	responsible breeder	2.7726
	responsible breeders	
56	long-acting bronchodilator	2.7726
	long-acting bronchodilators	
56	short-acting bronchodilators	2.7726
	short-acting bronchodilator	
56	insurance company	2.7726

	insurance companies	
56	carbon dioxide	2.7726
56	chest problems	2.7726
56	nhs choices	2.7726
56	chronic cough	2.7726
56	coenzyme q10	2.7726
	co-enzyme q10	
56	medical conditions	2.7726
	medical condition	
56	control group	2.7726
56	life with copd	2.7726
	life copd	
56	copd support	2.7726
56	stages of copd	2.7726
	stage of copd	
56	course of steroids	2.7726
	courses of steroids	
56	respiratory diseases	2.7726
	respiratory disease	
56	dry mouth	2.7726
56	internet explorer	2.7726
56	family history	2.7726
56	online forums	2.7726
56	gastrointestinal symptoms	2.7726
56	general health	2.7726
56	ill health	2.7726
56	health information	2.7726
56	reuters health	2.7726
56	space heaters	2.7726
56	holiday season	2.7726
56	wealth of information	2.7726
56	inhaled steroids	2.7726
	steroid inhaler	
56	landscape photography	2.7726
56	sick lungs	2.7726
56	medical treatments	2.7726
	medical treatment	
56	rescue medicine	2.7726
	rescue medicines	
56	respiratory nurse	2.7726
	respiratory nurses	
56	pet owners	2.7726
56	pulse oximeter	2.7726
56	oxygen treatment	2.7726
56	respiratory problems	2.7726

56	surgical procedure	2.7726
56	rehab program	2.7726
56	research team	2.7726
	team of researchers	
56	spiriva respimat	2.7726
	spiriva restimat	
56	respitory team	2.7726
56	right shoulder	2.7726
56	steam room	2.7726
56	social worker	2.7726
	social workers	
56	spiro test	2.7726
	spiro tests	
57	blood gas	2.5415
57	vitamin supplement	2.5415
58	clean air	2.426
58	diagnostic testing	2.426
59	johns hopkins news alert	2.1972
59	alpha-1 antitrypsin deficiency	2.1972
59	american cancer society	2.1972
59	health care provider	2.1972
	health care providers	
59	stem cell therapy	2.1972
59	chronic inflammatory disease	2.1972
59	vitamin d supplement	2.1972
59	social security disability	2.1972
59	inflammatory lung disease	2.1972
59	team relay event	2.1972
59	harvard medical school	2.1972
60	physical abilities	2.0794
	physical ability	
60	acute exacerbation	2.0794
	acute exacerbations	
60	email address	2.0794
60	medical advice	2.0794
60	aerobic exercise	2.0794
	aerobic exercises	
60	chronic ailments	2.0794
	chronic ailment	
60	air flow	2.0794
	flow of air	
60	hope air	2.0794
60	air quality	2.0794
60	deficiency alpha-1	2.0794
60	alternate legs	2.0794

60	amateur meteorologist	2.0794
60	swollen ankles	2.0794
60	antioxidant capacity	2.0794
60	housing ass	2.0794
60	asthma nurse	2.0794
60	panick attacks	2.0794
60	right attitude	2.0794
60	bad luck	2.0794
60	regular basis	2.0794
60	bblw team	2.0794
60	brecon beacons	2.0794
60	bill justice	2.0794
60	god bless	2.0794
60	blood flow	2.0794
60	blood gases	2.0794
60	blood oxygen	2.0794
	oxygenation of blood	
60	blood vessels	2.0794
60	upper body	2.0794
60	shallow breathing	2.0794
60	breathing treatment	2.0794
	breathing treatments	
60	shallow breather	2.0794
	shallow breathers	
60	transplant coordinator	2.0794
60	cart for shopping	2.0794
	shopping cart	
	shopping carts	
60	inflammatory cells	2.0794
60	type of challenge	2.0794
	types of challenges	
60	christmas dinner	2.0794
60	merry christmas	2.0794
60	chronic condition	2.0794
60	rehab classes	2.0794
60	combination products	2.0794
	combination product	
60	copd forums	2.0794
60	copd site	2.0794
60	free copy	2.0794
60	costa rica	2.0794
60	past couple	2.0794
60	cystic fibrosis	2.0794
60	damaged tissues	2.0794
	tissue damage	

60	front door	2.0794
60	emory university	2.0794
60	form of emphysema	2.0794
60	lungs with emphysema	2.0794
60	symptoms of emphysema	2.0794
	emphysema symptoms	
60	oxygen equipment	2.0794
60	life expectancy	2.0794
60	extracellular matrix	2.0794
60	support family	2.0794
60	festive season	2.0794
60	flight of stairs	2.0794
	flights of stairs	
60	ground floor	2.0794
60	peak flow	2.0794
60	flu vaccine	2.0794
	flu vaccines flu vaccines	
60	flue jab	2.0794
60	weight gain	2.0794
60	general public	2.0794
60	gift shop	2.0794
	shop for gifts	
60	safe graham	2.0794
60	great idea	2.0794
	great ideas	
60	great interest	2.0794
60	group of patients	2.0794
60	gym timer	2.0794
	gym timers	
60	happy holiday	2.0794
	happy holidays	
	holiday happiness	
60	heart tests	2.0794
60	transplant hospital	2.0794
60	university hospital	2.0794
60	pulmonary hypertension	2.0794
60	lung illness	2.0794
60	inhaled medicines	2.0794
60	metered-dose inhaler	2.0794
	metered-dose inhalers	
60	kind thoughts	2.0794
60	light smokers	2.0794
	light smoker	
60	weight loss	2.0794
60	love tad	2.0794

60	right lung	2.0794
60	right medication	2.0794
	right dose of medication	
60	medical school	2.0794
60	medical test	2.0794
	medical tests	
60	nice people	2.0794
60	nurse practitioner	2.0794
60	olive oil	2.0794
60	oral steroids	2.0794
60	oxidative stress	2.0794
	oxidant stress	
60	oxygen saturation	2.0794
	oxygen saturations	
60	supplemental oxygen	2.0794
60	oxygen supply	2.0794
60	rural patients	2.0794
60	salt pipe	2.0794
60	treatment plan	2.0794
60	pneumonia shot	2.0794
60	pr program	2.0794
60	sputum production	2.0794
60	treatment program	2.0794
	programme for treatment	
60	pulmonary specialist	2.0794
60	pulse rate	2.0794
60	respiratory system	2.0794
60	type of response	2.0794
	types of response	
60	white rice	2.0794
60	right side	2.0794
60	web site	2.0794
	web sites	
60	trouble sleeping	2.0794
60	sore throat	2.0794
	sore throat sore throat	
60	true southerner	2.0794
	true southerners	
60	spirometry test	2.0794
60	wood stoves	2.0794
	wood stove	
60	transplant team	2.0794
60	walk test	2.0794
60	wide world	2.0794
61	d supplement	1.7329

62	global initiative for obstructive lung disease	1.6094
63	amounts of vitamin a	1.3863
	vitamin a	
63	active life	1.3863
63	air con	1.3863
63	bad chest	1.3863
63	blf members	1.3863
63	boehringer ingelheim	1.3863
63	brand-name drugs	1.3863
63	lip breathing	1.3863
63	normal breathing	1.3863
63	phone call	1.3863
	phone calls	
63	pancreatic cancer	1.3863
63	risk of cancer	1.3863
63	car park	1.3863
	parking car	
63	careful planning	1.3863
63	primary care	1.3863
63	cbt course	1.3863
63	medical center	1.3863
63	smoking cessation	1.3863
63	chest clinic	1.3863
63	chest pain	1.3863
63	exercise class	1.3863
63	cleveland clinic	1.3863
63	community interventions	1.3863
63	health condition	1.3863
63	respiratory conditions	1.3863
63	conventional medication	1.3863
	medical content	
63	diagnoses of copd	1.3863
	diagnosis of copd	
63	copd friends	1.3863
	friends with copd	
63	copd group	1.3863
63	copd information	1.3863
63	copd management	1.3863
	effective copd management	
63	copd reading	1.3863
63	copd treatments	1.3863
	treatment of copd	
63	website copd	1.3863
63	oxygen cylinder	1.3863

63	risk of death	1.3863
63	lung decline	1.3863
63	types of diseases	1.3863
63	family doctor	1.3863
63	lung doctor	1.3863
63	patient education	1.3863
63	pharmacy errors	1.3863
63	simple exercise	1.3863
	simple exercises	
63	exercise tolerance	1.3863
63	excess mucus	1.3863
63	exhaust fan	1.3863
	exhaust fans	
63	genetic predisposition	1.3863
	genetic predispositions	
63	southern girls	1.3863
63	group members	1.3863
63	treatment group	1.3863
63	spiriva handihaler	1.3863
63	health insurance	1.3863
	health insurers	
63	healthy lungs	1.3863
63	holiday stress	1.3863
63	hospital radio	1.3863
63	respiratory illnesses	1.3863
63	immune pathways	1.3863
63	medical information	1.3863
63	inhalers as new types	1.3863
	types of inhalers	
63	later stages	1.3863
63	warm weather	1.3863
63	life for patients	1.3863
63	relay for life	1.3863
63	pursed lips	1.3863
63	lori palermo	1.3863
63	pub lunches	1.3863
63	lung volume	1.3863
63	steroid medication	1.3863
63	nursing staff	1.3863
	wonderful nursing staff	
63	oxygen tanks	1.3863
63	pet store	1.3863
63	spiritual program	1.3863
63	publix supermarket	1.3863
63	thanksgiving quotes	1.3863

63	resistance training	1.3863	
63	warning signs	1.3863	
63	south wales	1.3863	
63	tobacco surcharge	1.3863	
64	arterial gas test	1.0986	
64	blood gas test	1.0986	
65	emergency antibiotics	0.6931	
65	tiotropium bromide	0.6931	
65	clear mucus	0.6931	
65	prince county	0.6931	
65	great deal	0.6931	
65	handihaler device	0.6931	
65	dry powder	0.6931	
65	expert patient	0.6931	
65	health organization	0.6931	
65	hip replacement	0.6931	
65	maximal inspiration	0.6931	
65	pulmonary tests	0.6931	

APPENDIX D - EXACERBATION SENTENCE MIXUP RULES

```
// Pronoun Dictionaries
defDict first_person_prn =
i,me,we,us,myself,ourself,ourselves;
defDict second_person_prn = you,yourself,yourselves;
defDict third_person_prn =
she,her,he,him,they,them,himself,herself,itself,themself,thems
elves;

defDict demonstrative_prn = this,these,that,those;
defDict indefinite_prn =
anybody,anyone,anything,each,either,everybody,everyone,everyth
ing,neither,nobody,nothing,somebody,someone,something;
defDict possessive_prn =
my,your,his,her,its,our,your,their,mine,yours,his,hers,ours,yo
urs,theirs;


defDict oralDelivery = pill, pills, tablet, tablets, capsule,
capsules, oral;
defDict inhaledDelivery = inhaled, inhaler, inhalers, puffer,
puffers;
defDict intravenousDelivery = iv, inject, injected, needle,
drip,intravenous,intravenously;
defDict nebulizerDelivery = nebulizer, nebuliser,nebulizers,
nebulisers,nebulized,
aerosol,atomizer,atomiser,atomizers,atomisers;

defSpanType delivery =:
      ... [ ai(oralDelivery) ] ...||
      ... [ ai(inhaledDelivery) ] ...||
      ... [ ai(nebulizerDelivery) ] ...||
      ... [ ai(intravenousDelivery) ] ...||
      ... [ eqi('intra') eqi('-'){0,1} re('^venous') ] ...;

defSpanType noun =sentence:
      ... [ @NN ] ...||
      ... [ @NNP ] ...||
      ... [ @NNS ] ...||
      ... [ @NNPS ] ...;

defSpanType adj =:
      ... [ @JJ ] ...||
      ... [ @JJR ] ...||
      ... [ @JJS ] ...;

defSpanType adv =:
      ... [ @RB ] ...||
      ... [ @RP ] ...||
      ... [ @RBR ] ...||
      ... [ @RBS ] ...||
      ... [ @RRB ] ...;
```

```
defSpanType pronoun =:
      ... [ @PRP ] ...||
      ... [ @PRPS ] ...||
      ... [ @WP ] ...;

defSpanType verb =:
      ... [ @VB ] ...||
      ... [ @VBP ] ...||
      ... [ @VBZ ] ...||
      ... [ @VBD ] ...||
      ... [ @VBG ] ...||
      ... [ @VBN ] ...;

defSpanType first_prn=pronoun:
      ... [ ai(first_person_prn) ] ...;

defSpanType second_prn=pronoun:
      ... [ ai(second_person_prn) ] ...;

defSpanType third_prn=pronoun:
      ... [ ai(third_person_prn) ] ...;

defSpanType demonstrative_prn=pronoun:
      ... [ ai(demonstrative_prn) ] ...;

defSpanType indefinite_prn=pronoun:
      ... [ ai(indefinite_prn) ] ...;

defSpanType poss_prn=pronoun:
      ... [ ai(possessive_prn) ] ...;

defSpanType explicit_mention=:

//         .\causes\infection.mixup

defDict infect_common = bug,bugs, nasties;
defDict urtiWords = rhinitis, sinusitis, rhinorrhea,
pharyngitis;
defDict lrtiWords = bronchitis;
defDict genRespInf = flu, influenza;


//Upper Respiratory Tract Infection
//*********************************

defSpanType _cold =:
      ... [ @adj? @noun? re('^cold')] ...;

defSpanType cold =_cold:
      ... [ any* @noun ] ...;

defSpanType URTI =:
      ... [ eqi('acute')? re('^up') re('^resp') re('^tract')?
re('^infect')  ] ...||
      ... [ eqi('urti')  ] ...||
      ... [ eqi('uri')  ] ...||
```

```
      ... [ @cold ] ...||
      ... [ ai(urtiWords) ] ...;


//Lower Respiratory Tract Inection
//*******************************

defSpanType LRTI =:
      ... [ eqi('acute')? re('^low') re('^resp') re('^tract')?
re('^infect')  ] ...||
      ... [ eqi('lrti')  ] ...||
      ... [ eqi('lri')  ] ...||
      ... [ re('^chest') re('^infect') ] ...||
      ... [ re('^chest') re('^cold') ] ...||
      ... [ ai(lrtiWords) ] ...;


// Pneumonia
//*********************************

defSpanType pneumonia =noun:
      ... [ re('^pneum') ] ... ;

// General - grouping
//*********************************

defSpanType infection =:
      ... [ ai(infect_common) ]...||
      ... [ ai(genRespInf) ] ...||
      ... [ @URTI ] ...||
      ... [ @LRTI ] ...||
      ... [ @pneumonia ] ...||
      ... [ @adj? @noun? re('^infect') ] ...;

//         .\causes\irritants.mixup

defDict gen_allergen = allergen,allergens,aeroallergen;
defDict spores = spore, spores, mold;
defDict pollen = pollen, pollens;
defDict smoke = smoke, cigarettes, smoking, smokers,
smoker,cigarette;
defDict weather = cold, colder,cool, cooling,
cooler,hot,hotter,warm,warmer,warming,weather,temperature,temp
eratures,temp,temps;
defDict pollution = smog, pollution, exhaust, gas;
defDict dust = dust;
defDict pets = fur, pet,pets,animal,animals, dog,dogs,
cat,cats;


defSpanType irritants =:
      ... [ ai(gen_allergen) ] ...||
      ... [ ai(spores) ] ...||
      ... [ ai(pollen) ] ...||
      ... [ ai(smoke) ] ...||
      ... [ ai(weather){1,2} ] ...||
```

```
      ... [ ai(pollution) ] ...||
      ... [ ai(dust) ] ...||
      ... [ ai(pets) ] ...;


//         .\exacerbations.mixup

//         .\prevention\prevention.mixup

defSpanType avoid=:
      ... [ re('^avoid') @noun ] ...||
      ... [ re('^avoid') @verb ] ...||
      ... [ re('^stay') eqi('away') any{0,1} @noun ] ...
      ;

defSpanType air_quality_devices=:
      ... [ eqi('air')? re('^purif') ] ...||
      ... [ eqi('de')? eqi('-')? re('^humidif') ] ...||
      ... [ re('^dehumidif') ] ...||
      ... [ re('^de-humidif') ] ...||
      ... [ re('^aircon') ] ...||
      ... [ re('^air-con') ] ...||
      ... [ eqi('air') eqi('-')? re('^con') ] ...;


defSpanType prevention =:
      ... [ @air_quality_devices ] ...||
      ... [ @avoid ] ...;


//         .\symptoms\symptoms.mixup

// Symptom dictionaries
defDict mucus_words = phlegm, sputum, gunk,goo,sludge,ooze,
mucus,phlem,phlegmy,phlemy;
defDict dyspnoea_words = sob, dyspnea, dyspnoea;
defDict worse =
worse,worsen,worsening,deteriorate,weaken,decscend,decline,poo
r,bad,uncomfortable,discomfort, issue, issues,
problem,problems,difficult,difficulty,difficulties;
defDict better =
better,bettered,improve,improved,good,great,nice,positive,sati
sfactory,satisfy,superb,super,wonderful,comfortable,comfort;
defDict feel = feel,feels,feeling,felt;

// Exacerbation symptoms
// *********************

// Pain symptoms(i.e. chest,throat or back pain)
defSpanType pain =:
      ... [ 'sore' @noun ] ...||
      ... [ 'pain' 'in' any{0,1} @noun ] ...||
      ... [ @noun @PUN? 'pain' ] ... ||
      ... [ @adj @PUN? 'pain' ] ... ||
      ... [ @noun @PUN? re('^ache') ] ... ||
      ... [ @adj @PUN? re('^ache') ] ...
```

```
      ... [ @noun re('^achi') ] ... ||
      ... [ @adj re('^achi') ] ...||
      ... [ 'aching' @noun ] ...;


// Breathing difficulties
defSpanType breath=:
      ... [ re('^breath') ]...;

defSpanType dyspnoea=:
... [ ai(dyspnoea_words) ] ...||
... [ re('^breathless') ] ...||
... [ re('short') eqi('of') eqi('breath') ] ...||
... [ @breath <any,!re('[\W]')>{0,2} ai(worse) ] ... ||
... [ ai(worse) <any,!re('[\W]')>{0,2} @breath  ] ...;

// Cough/sputum

defSpanType cough=:
      ... [ @adj? re('^cough') ] ...;

defSpanType mucus=:
      ... [ @adj? ai(mucus_words) ] ...;

// General
defSpanType general_symptoms=:
      ... [ re('^feel') @adj @adj? @adj? ] ...;

// Grouping

defSpanType symptom =:
      ... [ @pain ] ...||
      ... [ @dyspnoea ] ...||
      ... [ @cough ] ...||
      ... [ @mucus ] ...||
      ... [ @general_symptoms] ...;

//         .\treatment\antibiotics.mixup

defDict antibiotic_brands = Acarexx, Achromycin, Achromycin,
Actisite, Actisite, Adoxa, AgriMectin, Agrimectin, Aknemin,
Aknemycin, AlaTet, AlaTet, Almodan, Alodox, Alodox, Amficot,
Amix, amopen, Amoram, AMOXI, Amoxicillinan, Amoxicot,
Amoxidin, Amoxil, Amoxymed, Ampitrin, Amrit, Arestin,
Arpimycin, Atridox, Atridox, Atridox, Augmentin, Aureomycin,
Avidoxy, AzaSite, Azinthromycin, Benzamycin, Biaxin, Bimectin,
BioMycin, BioMycin, BioTab, BioTab, BioTab, Biomox, Blemix,
Brodspec, Brodspec, Chemocycline, Chemocycline, Clavamox,
CleeravueM, Cyclodox, Cyclodox, Cyclomin, Declomycin,
Declomycin, Declomycin, Declomycin, DelMycin, Demix, Demix,
Dentomycin, Diabecline, Dificid, Dispermox, Doryx, Doryx,
Doxal, Doxal, Doxatet, Doxatet, Doxy, Doxy, DoxyCaps,
DoxyCaps, Doxylar, Doxylar, Dynabac, Dynacin, EMycin, ESolve,
Economycin, Economycin, Emgel, Emtet499, Emtet500, EndoMectin,
Equell, Equimax, Equimectrin, Eqvalan, Ery, ErySol, EryTab,
Eryc, Erycette, Eryderm, Erygel, Erymax, Erymin, EryPed,
```

```
ErythraDerm, Erythrocot, Erythrolar, Erythromicin, Erythroped,
Eryzole, Eyemycin, Galenamox, Galenomycin, Galenomycin,
Geomycin, Geomycin, Heartgard, Hexasol, Hexasol, Ilosone,
Ilotycin, Ivermax, Ivomec, Ketek, Liquamycin, Liquamycin,
Maracyn, Marcillin, Meclan, Meclan, Mectizan, Micotil,
Micotil, Minocin, Minogal, Monodox, Monodox, Morgidox,
Morgidox, Moxatag, Moxilin, MYE, Myrac, Nordox, Nordox,
Noromectin, Noromycin, Noromycin, Ocudox, Ocudox, Omnipen,
OmnipenN, Oracea, Oracea, Oraxyl, Oraxyl, Ornacycline,
Ornacycline, Ornacycline, Ornacyn, Panmycin, Panmycin, PCE,
Pediazole, Pelodis, Pelodis, Pelodis, Penbritin, Periostat,
Periostat, Polycillin, PolycillinPRB, Polyflex, Principen,
Pylera, Pylera, Ramysis, Ramysis, Rapamune, Respillin,
Rimacillin, Rimoxallin, Robimycin, Rommix, Romycin, Senox,
Solodyn, Spectrobid, Staticin, Stiemycin, Stromectol,
Sulfimycin, Sumox, Sumycin, Sustamycin, Sustamycin, TStat,
Terak, Terak, TerraCortril, TerraCortril, Terramycin,
Terramycin, TetrabidOrganon, TetrabidOrganon, Tetracap,
Tetracap, Tetrachel, Tetrachel, Tetracon, Tetracon, Tetradure,
Tetradure, Tetralysal, Tetrex, Tetrex, Tetroxy, Tetroxy,
Theramycin, Tija, Tija, Topicycline, Topicycline, Totacillin,
TotacillinN, Trimox, Tylan, Unasyn, Urobiotic, Urobiotic,
Vectrin, Vetrimec, Vetrimec, VibraTabs, VibraTabs, VibraTabs,
Vibramycin, Vibramycin, VibramycinD, VibramycinD, Vidopen,
Wymox, Zimecterin, Zithromax, Zmax, Zoxycil, ZPAK;

// Common name
defSpanType antibioticCommName =:
... [ re('carbef$') ] ... ||
... [ re('CARBEF$') ] ... ||
... [ re('^cef') ] ... ||
... [ re('^CEF') ] ... ||
... [ re('cidin$') ] ... ||
... [ re('CIDIN$') ] ... ||
... [ re('cillide$') ] ... ||
... [ re('CILLIDE$') ] ... ||
... [ re('CILLIN$') ] ... ||
... [ re('cillin$') ] ... ||
... [ re('cillinam$') ] ... ||
... [ re('CILLINAM$') ] ... ||
... [ re('CYCLINE$') ] ... ||
... [ re('cycline$') ] ... ||
... [ re('fungin$') ] ... ||
... [ re('FUNGIN$') ] ... ||
... [ re('gillin$') ] ... ||
... [ re('GILLIN$') ] ... ||
... [ re('kacin$') ] ... ||
... [ re('KACIN$') ] ... ||
... [ re('CIDIN$') ] ... ||
... [ re('micin$') ] ... ||
... [ re('MICIN$') ] ... ||
... [ re('mycin$') ] ... ||
... [ re('MYCIN$') ] ... ||
... [ re('monam$') ] ... ||
... [ re('MONAM$') ] ... ||
... [ re('oxef$') ] ... ||
```

```
... [ re('OXEF$') ] ... ||
... [ re('parcin$') ] ... ||
... [ re('PARCIN$') ] ... ||
... [ re('penem$') ] ... ||
... [ re('PENEM$') ] ... ||
... [ re('rifa$') ] ... ||
... [ re('RIFA$') ] ... ||
... [ re('rubicin$') ] ... ||
... [ re('RUBICIN$') ] ... ||
... [ re('tricin$') ] ... ||
... [ re('TRICIN$') ] ... ;

// Generic mentions
defSpanType antibioticGeneric=:
... [ re('^antibio') ] ... ;

// Brand name mentions
defSpanType antibioticBrandName =:
... [ ai(antibiotic_brands) ] ... ;

// Grouping
defSpanType antibiotic =:
... [ @delivery? @antibioticCommName @delivery? ] ... ||
... [ @delivery? @antibioticGeneric @delivery? ] ... ||
... [ @delivery? @antibioticBrandName @delivery? ] ...;

//         .\treatment\home_remedies.mixup

defDict hot= hot,warm;

defSpanType hot_drink =:
     // brand names?  i.e. Lemsip
     ... [ re('^lemon') @PUN? re('^flavour')? re('^flavor')?
@PUN? ai(hot)? @PUN? re('^drink') ] ...||
     ... [ ai(hot) @PUN? re('^lemon')? @PUN? 'flavoured'?
'flavored'? re('^drink') ] ...;

defSpanType home_remedies=:
     ... [ @hot_drink ] ...;

//         .\treatment\oxygen.mixup

defDict oxygenWords = oxygen, o2;
defDict nose = nose,noses,nasal,nostril,nostrils;
defDict cannula = cannula, cannulae, tube, tubes;
defDict therapy = therapy, treatment,supplement,supplemental;
defDict preMaskWords = face,venturi;
defDict o2exclude =
saturation,level,levels,sats,sat,exchange,intake;

defSpanType nasal =sentence:
     ... [ ai(nose) ] ...;

defSpanType o2delivery =sentence:
     ... [ @nasal any{0,2} ai(cannula) ] ...||   //e.g. 'nose
tubes', 'nasal cannula'
```

```
     ... [ ai(cannula) any{0,2} @nasal ] ...||   // e.g.
'tubes  in my nose', 'cannula for my nose'
     ... [ ai(oxygenWords) ai(preMaskWords){0,1} eqi('mask')
] ...|| // e.g. 'oxygen mask', 'o2 face mask'
     ... [ ai(preMaskWords) eqi('mask') ] ...; // 'face
mask', 'venturi mask'

defSpanType oxygen=sentence:
     ... [ ai(therapy){0,1} ai(oxygenWords) ] !ai(o2exclude)
...|| // e.g. 'supplemental oxygen', 'oxygen', but not 'oxygen
level' or 'oxygen exchange'
     ... [ ai(oxygenWords) any{0,2} ai(therapy) ]
!ai(o2exclude) ...||
     ... [ @o2delivery ]...;

//         .\treatment\steroids.mixup

//*** Generic Steroid Dictionaries ***

defDict sterGeneric = steroid, steroids, corticosteroid,
corticosteroids;
defDict sterMisSpell = steriod, steriods, corticosteriod,
corticosteriods;


// Oral Steroids

//  **** Prednisolone Dictionaries *****

defDict prednisoloneComm = prednisolone, pred, prdl;
// abbreviated - AsmalPred Plus,  Econopred Plus, Inflamase
Forte, Inflamase Mild, Isolone Forte, Key-Pred SP, Minims
Prednisolone, Ocu-Pred Forte, Poly Pred, Pred Forte, Pred
Mild, Predacort 50, Predalone 50, Predicort RP, Pri-Cortin 50,
//defDict prednisoloneBrand = AK-Cide, AK-Pred, AsmalPred,
Blephamide, Bubbli-Pred, Cetapred, Codelsol, Codelson,
Cotolone, Depo-Predate, Econopred, Flo-Pred, Hydeltra-T.B.A.,
Hydeltrasol, Inflamase, Inflamase, Isolone, Key-Pred, Key-
Pred, Medasulf, Medicort, Metimyd, MILLIPRED, Minims, Ocu-
Pred, Ocu-Pred, Ocu-Pred-A, Omnipred, Optimyd, Orapred,
Pediapred, Poly, Pred, Pred-G, Pred-Ject-50, Pred-Phosphate,
Predacort, Predaject-50, Predalone, Predate-50, Predcor,
Predenema, Predfoam, Predicort, Predicort-50, Prednesol,
Prednisol, Prednoral, Predsol, Predsulfair, Prelone, Pri-
Cortin, Pricortin, Stintisone, Sulster, Supred, Surolan,
Temaril-P, Vasocidin, Veripred;
defDict prednisoloneBrand = AsmalPred, Blephamide, Cetapred,
Codelsol, Codelson, Cotolone, Econopred, Hydeltrasol,
Inflamase, Inflamase, Isolone, Medasulf, Medicort, Metimyd,
MILLIPRED, Minims, Omnipred, Optimyd, Orapred, Pediapred,
Poly, Pred, Predacort, Predalone, Predcor, Predenema,
Predfoam, Predicort, Prednesol, Prednisol, Prednoral, Predsol,
Predsulfair, Prelone, Pricortin, Stintisone, Sulster, Supred,
Surolan, Vasocidin, Veripred;
defDict prednisoloneMod = plus, forte, mild;
```

```
//  **** Methylprednisolone Dictionaries ***

defDict methylprednisoloneComm = methylprednisolone, meprdl,
metipred;
defDict methylprednisoloneBrand = Cortimed, Dep, Depmedalone,
Depopred, Duralone, Duro, HybriSil, Medipred, Medralone,
Medralone, Medrol, Methacort, Methylcotol, Methylcotolone,
Methylone, Methylpred, Predacorten;

// Inhaled Steroids

//  ********* Beclomethasone Dictionaries ***

defDict beclomethasoneComm = Beclomethasone, beclometh;
defDict beclomethasoneBrand = Beclazone, Beclovent, Beconase,
Propaderm, Qvar, Vancenase, Vanceril;
defDict beclomethasoneBrandMod = A, C;

//  ***** Fluticasone Dictionaries ****


defDict fluticasoneComm = Fluticasone;
defDict fluticasoneBrand = Advair, Cutivate, Flixotide,
Flonase, Flovent, foxair, seretide, viani,Veramyst;
//foxair,seretide,viani missing originally, added 26-02-2013
defDict fluticasoneBrandMod = Nebule;

//  ***** Budesonide Dictionaries ***

defDict budesonideComm = Budesonide;
defDict budesonideBrand = Entocort, Preferid, Pulmicort,
Rhinocort, Symbicort;
defDict budesonideBrandMod = LS, Respules, Turbohaler, Aqua;


// Steroids
// ********

// Generic
// ******************************************************
defSpanType steroidGeneric=:
... [ ai(sterGeneric) ] ... ||
... [ ai(sterMisSpell) ] ...;

// Oral
// ******************************************************

defSpanType prednisolone=:
... [ ai(prednisoloneComm) ] ... ||
... [ ai(prednisoloneBrand) ] ... ||
... [ ai(prednisoloneBrand) ai(prednisoloneBrand) ] ...;

defSpanType methylprednisolone=:
... [ ai(methylprednisoloneComm) ] ...||
... [ ai(methylprednisoloneBrand) ] ...;
```

```
defSpanType oralSteroids =:
... [ @prednisolone ] ...||
... [ @methylprednisolone ] ...;

// Inhaled
// ********************************************************

defSpanType beclomethasone=:
... [ ai(beclomethasoneComm) ] ... ||
... [ ai(beclomethasoneBrand) ] ... ||
... [ ai(beclomethasoneBrand) ai(beclomethasoneBrandMod) ] ...
;

defSpanType fluticasone=:
... [ ai(fluticasoneComm) ] ... ||
... [ ai(fluticasoneBrand) ] ... ||
... [ ai(fluticasoneBrand) ai(fluticasoneBrandMod) ] ... ;

defSpanType budesonide=:
... [ ai(budesonideComm) ] ... ||
... [ ai(budesonideBrand) ] ... ||
... [ ai(budesonideBrand) ai(budesonideBrandMod) ] ... ;

defSpanType inhaledSteroids =:
... [ @beclomethasone ] ... ||
... [ @fluticasone ] ... ||
... [ @budesonide ] ... ;

// Combine to General Steroids Span
// ********************************************************

defSpanType _steroids=:
...[ @steroidGeneric ]...||
...[ @oralSteroids ]...||
...[ @inhaledSteroids ]...;

defSpanType sterDelivery =:
... [ @_steroids any{0,1} @delivery   ] ...||
... [ @delivery any{0,1} @_steroids  ] ...;

defSpanType steroids=:
      ... [ @_steroids ] ...||
      ... [ @sterDelivery ] ...;


//********** Combine ****

defSpanType auto_exacerbation=subj:
      [ ... @explicit_mention ...]||
      [ ... @infection ...]||
      [ ... @symptom ...]||
      [ ... @antibiotic ...]||
      [ ... @home_remedies ...]||
      [ ... @steroids ...]||
      [ ... @oxygen ...];
```