CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Classification of socially generated medical data
Authors	Alnashwan, Rana
Publication date	2019-09
Original Citation	Alnashwan, R. 2019. Classification of socially generated medical data. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2019, Rana Alnashwan. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-15 04:25:35
Item downloaded from	https://hdl.handle.net/10468/9842



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Classification of Socially Generated Medical Data

Rana Othman Alnashwan

MSc

Thesis submitted for the degree of

Doctor of Philosophy



NATIONAL UNIVERSITY OF IRELAND, CORK

SCHOOL OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

September 2019

Supervisors: Humphrey Sorensen

Adrian O'Riordan

Research supported by Princess Nourah bint Abdulrahman University

Contents

List of Figures vii
List of Tables ix
Abstract xiii
Acknowledgements xv

1 Introduction **1**
1.1 Introduction 1
1.2 Research Questions and Contributions 2
1.3 Thesis Structure 4
1.4 Publications 6

2 Background **7**
2.1 The Problem of Sentiment Analysis 7
2.2 Primary Definitions 9
2.3 Sentiment Analysis Research 10
2.3.1 Sentiment Analysis Tasks 11
2.3.2 Sentiment Analysis Levels 11
2.3.3 Sentiment Analysis Approaches 12
2.3.4 Data Domains in Sentiment Analysis 12
2.4 Mining Social Media for Sentiment 13
2.4.1 Challenges of Social Media 13
2.4.2 Social Media Sub-Genres 14
2.4.2.1 Twitter 14
2.4.2.2 Online Forums 15
2.5 Summary 16

3 Literature Review **18**
3.1 Sentiment Classification: Prominent General Approaches 18
3.1.1 Machine Learning 18
3.1.1.1 Feature Engineering 19
3.1.1.2 Supervised Learning Approach 21

3.1.1.3 Semi-Supervised Learning Approach 32
3.1.1.4 Deep Learning Approaches 33
3.1.2 Lexicon-Based Approach 33
3.1.3 Sentiment Lexicon Generation 34
3.2 Applications of Social Media Opinions 35
3.2.1 General Applications in SA Using Social Media 36
3.2.2 Applications of Socially Generated Medical Data 37
3.2.2.1 Mining Personal Health Information Polarity and Opinion 37
3.2.2.2 Analysing Emotions and Studying Emotional Affects 38
3.2.2.3 Measuring the Quality of Document Content or Health Interaction 39
3.2.2.4 Analysing Drug Medical Data 39
3.2.2.5 Detecting Health Care Quality 40
3.3 Summary 42

4 Short-form Sentiment Analysis Using Ensemble Learning 43
4.1 Introduction 43
4.2 Improving Sentiment Analysis Through Ensemble Learning of Meta-Level Features 43
4.2.1 Feature Engineering 44
4.2.1.1 Feature Hashing 44
4.2.1.2 Meta-Level Features 45
4.3 Experimental Evaluation 50
4.3.1 Classifiers 50
4.3.1.1 Two-Class Support Vector Machine 51
4.3.1.2 Two-Class Bayes Point Machine 52
4.3.1.3 Two-Class Logistic Regression 52
4.3.1.4 Two-Class Decision Forest 53
4.3.2 Classifier Ensemble for Tweet Sentiment Analysis 54

4.3.3 Datasets 55
4.3.3.1 Stanford Twitter Sentiment (STS) 55
4.3.3.2 SemEval-2016 56
4.3.3.3 Health Care Reform (HCR) 56
4.3.4 Experimental Setup 56
4.4 Results 59
4.5 Summary 61

5 Long-form Sentiment Analysis 62
5.1 Introduction 62
5.2 Motivation 62
5.2.1 Why Lyme Disease? 63
5.3 Previous Studies 64
5.4 Design and Implementation of Automated Sentiment Classification Recognition 66
5.4.1 Domain-Dependent Categories Identification 66
5.4.1.1 Dataset 66
5.4.1.2 Identifying Categories 66
5.4.1.3 Validating Category Selection 68
5.4.2 Feature Engineering 70
5.4.2.1 The Baseline 71
5.4.2.2 Content-Free Features 72
5.4.2.3 Meta-Level Features 72
5.4.2.4 Content-Specific Features 76
5.4.3 Classification Approach 76
5.4.3.1 Automatic Feature Selection/Reduction 77
5.5 Experimental Evaluation 78
5.5.1 Dataset 78
5.5.2 Gold Labelling 80
5.5.3 Data Pre-processing 81
5.5.4 Machine Learning Techniques 82

5.5.5 Evaluation Metrics 83
5.6 Experiment Results 84
5.7 Summary 87

6 Application to an Alternative Disease Dataset 89
6.1 Introduction 89
6.1.1 Why Analyse Lupus Disease Posts? 90
6.2 Design and Implementation of Automated Sentiment Classification Recognition 90
6.2.1 Domain-Dependent Categories Identification 90
6.2.1.1 Dataset 90
6.2.1.2 Identifying Categories 90
6.2.1.3 Evaluating Categories 92
6.2.2 Feature Engineering 93
6.2.3 Classification Approach 93
6.3 Experimental Evaluation 93
6.3.1 Dataset 93
6.3.2 Gold Labelling 93
6.3.3 Machine Learning Techniques 94
6.4 Experiment Results 95
6.5 Summary 97

7 Evaluation and Discussion 98
7.1 Introduction 98
7.2 Performance Evaluation 98
7.3 Evaluation of Content-free, Meta-level and Content-specific Feature Sets 103
7.3.1 Statistical Tests 103
7.3.2 Goodness of Fit 105
7.4 Error Analysis 108
7.5 Main Findings 109
7.6 Summary 110

8 Semi-supervised Approach - Modified Co-training **111**

8.1 Introduction . . . 111

8.2 Previous Studies . . . 112

8.3 Methods and Experiment Design . . . 114

8.3.1 Co-training Model . . . 114

8.3.2 Determining Auto-Labelled Data . . . 117

8.3.3 Baseline Runs . . . 119

8.4 Experiment Setup . . . 119

8.4.1 Number of Iterations . . . 119

8.4.2 Unlabelled Data Available at Each Iteration . . . 120

8.4.3 Platform . . . 121

8.5 Experiment Results . . . 122

8.6 Summary . . . 123

9 Long-form Content and Topic Analysis **125**

9.1 Introduction . . . 125

9.2 Stakeholder Perspectives . . . 126

9.2.1 Individual Perspective . . . 126

9.2.2 Professional Perspective . . . 127

9.2.3 Organizational Perspective . . . 127

9.2.4 Related work in stakeholder identification of medical social data . . . 128

9.3 Content Analysis . . . 128

9.4 Topic Analysis . . . 132

9.5 Discussion . . . 135

9.5.1 Benefits to Various Stakeholders . . . 136

9.5.2 Health Information Leaflet and Expert Comparison . . . 137

9.6 Summary . . . 139

10 Conclusion and Future Work **141**

10.1 Main Conclusions of the Empirical Investigations . . . 141

10.2 Future work . . . 145

10.2.1 Generalizing the Feature-based Model to Other Diseases and Platforms 145
10.2.2 Improving and Exploring Semi-supervised Learning Approaches 145
10.2.3 Conducting more Experiments on Extracted Features . . 146
10.2.4 Data Visualisation 146

A Expert Opinion **162**

List of Figures

2.1 Example of the hierarchy structure of DailyStrength forums . . 16

3.1 Feature engineering position in ML work flow 19

4.1 Schematic of the approach 44
4.2 Sample of features vector space 48
4.3 Meta-level feature Algorithm 50
4.4 A general representation of SVM algorithm [1] 51
4.5 S-shaped curve in Logistic regression [2] 53

5.1 Process for identifying sentiment categories 68
5.2 Flow diagram for annotation prosses 70
5.3 Statistics for the distribution of the categories 71
5.4 Feature set 77
5.5 Flow diagram of Lyme data collected for multi-class sentiment classification 79
5.6 Vote distribution 80
5.7 Example of a post related to Lyme infection confusion 82
5.8 Example of a post related to the Depressed and frustrated class related to Lyme Disease 82

6.1 Statistics for the distribution of the categories related to Lupus posts 92
6.2 Vote distribution for Lupus posts data 94
6.3 Flow diagram of Lupus data collected for multi-class sentiment classification 94
6.4 Example of a post related to Asking about treatment class related to Lupus disease 95

7.1 The learning curve for the feature-based model 103
7.2 Confusion matrix for Lyme and Lupus data 109

8.1 Procedure followed by the modified co-training model 115
8.2 Pseudocode for enhanced co-training 116
8.3 Average accuracy of different threshold values. 118
8.4 Data split for semi-supervised learning runs, baseline supervised learning runs, and fully supervised learning runs. 119
8.5 Average accuracy of different proportions of N to create an unlabelled pool (U') for the two classifiers in each iteration process. 121

9.1 Content analysis process 131
9.2 Presentation of extracted process 132

A.1 Email survey for expert opinion 162

List of Tables

3.1 Summary of studies using a supervised learning approach for short-form data . . . 26
3.2 Summary of studies using a supervised learning approach for long-form data . . . 30
3.3 List of emoticons used in [3] . . . 32
3.4 Summary of examples of studies on the application of sentiment analysis work in the medical domain . . . 41

4.1 Features from lexicon and polarity resources . . . 49
4.2 Examples of annotated Tweets from the datasets . . . 56
4.3 The Twitter Dataset statistics . . . 57
4.4 Polarity classification performance . . . 60

5.1 Description of categories and their subcategories . . . 69
5.2 Adopted lexical feature in our model . . . 73
5.3 Description of the extracted lexicon feature . . . 75
5.4 Categorizing different feature sets used for multiclass classification . . . 78
5.5 Fleiss' Kappa interpretation [4] . . . 80
5.6 Performance in terms of accuracy for three supervised learning algorithms . . . 83
5.7 Experimental result for different feature set on Lyme Disease . 86

6.1 Description of Lupus discussions categories and their subcategories . . . 91
6.2 Experimental result for different feature set on Lupus disease . 96

7.1 Feature-based model performance on a Lyme Disease dataset using cross validation . . . 101
7.2 Feature-based model performance on a Lupus disease dataset using cross validation . . . 102

7.3 Accuracy of the classification model with three different types of feature sets on the Lyme Disease dataset. 104
7.4 Accuracy of the classification model with three different types of feature set on the Lupus dataset. 105
7.5 Kappa performance of the classification model with three different types of feature set on the Lyme Disease dataset. 107
7.6 Kappa performance of the classification model with three different types of feature set on the Lupus dataset. 107

8.1 Ratios of the class distribution for both datasets 118
8.2 Classification accuracy (%) of co-training with DI and DD views 123

9.1 Topic modelling results 134
9.2 Comparison of topics contained in leaflets and those identified on social media 138
9.3 Benefits of our model from the physician participants' point of view 139

I, Rana Othman Alnashwan, certify that this thesis is my own work and has not been submitted for another degree at University College Cork or elsewhere.

_____

Rana Othman Alnashwan

Abstract

The growth of online health communities, particularly those involving socially generated content, can provide considerable value for society. Participants can gain knowledge of medical information or interact with peers on medical forum platforms. However, the sheer volume of information so generated – and the consequent 'noise' associated with large data volumes – can create difficulties for information consumers. We propose a solution to this problem by applying high-level analytics to the data – primarily *sentiment analysis*, but also *content* and *topic* analysis - for accurate classification. We believe that such analysis can be of significant value to data users, such as identifying a particular aspect of an information space, determining themes that predominate among a large dataset, and allowing people to summarize topics within a big dataset.

In this thesis, we apply *machine learning* strategies to identify sentiments expressed in online medical forums that discuss Lyme Disease. As part of this process, we distinguish a complete and relevant set of categories that can be used to characterize Lyme Disease discourse. We present a feature-based model that employs supervised learning algorithms and assess the feasibility and accuracy of this sentiment classification model. We further evaluate our model by assessing its ability to adapt to an online medical forum discussing a disease with similar characteristics, Lupus. The experimental results demonstrate the effectiveness of our approach.

In many sentiment analysis applications, the labelled training datasets are expensive to obtain, whereas unlabelled datasets are readily available. Therefore, we present an adaptation of a well-known semi-supervised learning technique, in which *co-training* is implemented by combining labelled and unlabelled data. Our results would suggest the ability to learn even with limited labelled data. In addition, we investigate complementary analytic techniques – content and topic analysis – to leverage best used of the data for various consumer groups.

Within the work described in this thesis, some particular research issues are addressed, specifically when applied to socially generated medical/health datasets:

- When applying binary sentiment analysis to short-form text data (e.g. Twitter), could meta-level features improve performance of classification?
- When applying more complex multi-class sentiment analysis to classification of long-form content-rich text data, would meta-level features be a

useful addition to more conventional features?

- Can this multi-class analysis approach be generalised to other medical/health domains?
- How would alternative classification strategies benefit different groups of information consumers?

Acknowledgements

Accomplishing this thesis would not have been possible without the great support and encouragement of several individuals. I would like to take this opportunity to express my gratitude to all the people who directly or indirectly contributed and extended their helpful suggestions to the completion of this work.

First and foremost, I would sincerely like to acknowledge the tireless and prompt help of my supervisors, Humphrey Sorensen and Adrian O'Riordan, for their dedicated support, guidance, continual encouragement, invaluable advice, time and effort. The wealth of their knowledge and experience of research and broad research vision have always inspired me and guided me when tackling challenges.

To Cathal Hoare, who shared his knowledge and experience during my study. To my friends in the PhD lab, who made the work environment both exciting and entertaining throughout my PhD. With a particular word of thanks to Tamara Vagg, who gave me such support, encouragement and precious friendship. To my friends in Cork, who made the journey enjoyable and were also there to offer support.

I would also like to thank my sponsor in Saudi Arabia, Princess Nourah Bint Abdulrahman University, for their scholarship and continuous support throughout the years of my study in Ireland.

I also wish to thank the examiners of this thesis, Dr James Doherty and Dr Jennifer Foster, for their valuable comments.

The endless love, encouragement and support of my family cannot be expressed by words. To my father, Prof. Othman Alnashwan, I am grateful for your unconditional support and for inspiring me to follow my dreams. To the kindest person in my life, my mother Helah, for your incredible encouragement, patience and sacrifice and for making me who I am today. To my sisters and brothers, who have given me positive energy during this journey. You are the most significant support and source of my strength.

My heartfelt thanks to my Husband Ibrahim, for walking step by step with me during the whole of our PhD studies together. I will always sincerely appreciate your constant encouragement, patience and love. To my beloved daughter Rand, thank you for creating your unique charm during my journey with your kindness, beauty and hugs. To my new arrival, my son Mohammed, who was born during this journey and added so much joy to it. Thank you all for everything.

Chapter 1

Introduction

1.1 Introduction

The last decade has seen the rapid expansion of social media [5] as a means of interaction, communication and sharing of information among interested individuals and groups. One phenomenon is the emergence of topic-specific forums – e.g. addressing a health-related matter [6] – where individuals can submit, share and access information pertinent to the topic of interest. Through the use of such forums, or groups, individuals can effectively become members of a larger collective – and they can contribute content, queries, answers and discussions to be shared among the collective. Perhaps second to search engines, such forums act as an important online resource for interested parties.

Quite often, a forum might be established for a very specific reason and, in its early days, be quite focused as regards content. However, over time, it can become de-focused, as it must accommodate a diversity of views, of users and of contributions: effectively, the information space broadens out – sometimes to a chaotic level. For any given forum, this presents a problem for its users, and especially any new/potential users, who might be lost in this information space: how does one navigate this information space? how does one identify significant themes and threads within the information space? how does one filter out the noise so that relevant data can be found?

In this thesis, we broadly address the problems just identified. It is our intention to take a large corpus of textual data extracted from health-related forums – first on the topic of Lyme Disease, later on other comparable topics – and identify how it could be analysed in different ways so as to provide a rudimentary

mapping of the information space that should prove beneficial to users. By such an *information classification* approach, we aim to subdivide a large diverse (and complex) information space into a set of smaller, more coherent, sub-collections that are more comprehensible.

One form of analysis/classification which we primarily focus on is *sentiment analysis*, where we map user-contributed content to a range of sentiments (or affects)[1] that might best characterize the theme(s) of the content. To achieve sentiment analysis, we develop and test various *machine learning* approaches to the problem: we vary the essential learning strategies, we vary the exact learning mechanism adopted, we vary the learning parameters, we vary the linguistic processing methods and we vary the resources used – all in an attempt to achieve maximum accuracy in the classification task (and to estimate what the upper limit might be). While we consider short-form text classification (e.g. Twitter data), we are primarily interested in long-form, content-rich, text that is more characteristic of medical forum data. We present various experiments, results and analyses in regard to machine learning for sentiment analysis throughout the thesis.

While the major part of the thesis relates to sentiment analysis, this is just one criterion by which we can classify an information space. There are other, orthogonal, measures by which we might organize the content, e.g., based on topic analysis, on content analysis, etc. These mechanisms, can also provide useful insight into a complex information space – and the combination of methods can have different, but significant, benefits for various classes of data user.

1.2 Research Questions and Contributions

The main research questions investigated in this thesis are as follows:

RQ1: Can meta-level features improve binary sentiment analysis performance on short-form socially generated text, Twitter in particular?

The main contributions of our work under this research question can be summarized as follows:

- The work applies existing sentiment analysis approaches to Twitter data.

[1] The terms 'sentiment' and 'affect' have been used interchangeably in the literature and refer to the extraction of the opinions, emotions or views that may be expressed in text. See Chapter 2 for a note on terminology.

- We investigate the effectiveness of using a combination of existing lexicon resources as meta-level features in Twitter sentiment classification.
- We apply an ensemble learning approach based on the meta-level features of seven existing lexicon resources.

RQ2: Can meta-level features, in conjunction with conventional features, improve multi-class sentiment analysis performance on more content-rich Lyme Disease medical forums?

After investigating the effectiveness of using the meta-level features for binary sentiment analysis in Twitter (i.e., short-form socially generated data), we were motivated to utilize this to enhance multi-class sentiment classification of medical forums (i.e., long-form, medical, socially generated data). The main contributions of our work under this research question can be summarized as follows:

- We identify and present a complete and relevant set of categories that can be used to characterize Lyme Disease discourse.
- We investigate strategies, both individually and collectively, for automating the classification of medical forum posts into the categories identified under the previous contribution.
- We present a feature-based model that consists of three different feature sets: content-free, content-specific, and meta-level features.
- We build and utilize a new gold-standard dataset for evaluating our proposed methods in sentiment analysis tasks.

RQ3: Can the previous feature-based model adapt to an online medical forum discussing a different disease (i.e., Lupus)?

- We evaluate and assess the set of categories identified on medical forum discussions related to Lupus.
- We evaluate the previous proposed feature-based classification model for an automated multi-class classification model in relation to a different disease, i.e., Lupus.
- We build and utilize a new gold-standard dataset for evaluating our proposed methods in sentiment analysis tasks.

RQ4: Can the feature-based model be adapted for semi-supervised learning?

The research in this thesis is motivated by the challenge of multi-class sentiment classification in medical and health-related discourse. A major obstacle arises with respect to the scarcity of labelled data as a basis for machine learning. A potential solution to this, which we address in this thesis, is the use of *co-training* as a semi-supervised approach for dealing with labelled and unlabelled data. The main contributions of our work under this research question can be summarized as follows:

- We adapt a co-training model to our feature-based model, to recognize sentiment automatically in a multi-class classification problem in a semi-supervised setting.
- We evaluate the previously proposed semi-supervised classification model for an automated multi-class classification model for a different disease, i.e., Lupus.

RQ5: Can the medical forums discourse be analysed by different methods to be useful in other ways?

In addition to sentiment analysis, there are several different methods of analysis, either taken individually or collectively, that could benefit and help diverse healthcare stakeholders due to their focusing on different aspects of the information collected. The main contributions of our work under this research question can be summarized as follows:

- We perform a *content analysis* of an online health community to obtain a general view of the medical content available specifically for Lyme Disease forums.
- We perform a *topic analysis* to detect different topics represented by participants or patients in the health community related to Lyme Disease forums utilizing the latent Dirichlet allocation (LDA) model.

1.3 Thesis Structure

The remainder of the thesis is structured as follows:

Chapter 2 (Background): This chapter first describes the problem and provides a definition of sentiment analysis. After that, it provides background knowledge related to sentiment analysis tasks, levels and the most prominent approaches and types of data. Lastly, it identifies some of the challenges faced when mining

social media for sentiment analysis.

Chapter 3 (Literature Review): This chapter presents an overview of the existing work in the area of sentiment analysis. It then outlines some of the potential applications of sentiment analysis, including its application to socially generated medical data.

Chapter 4 (Short-form Sentiment Analysis): This chapter describes a supervised learning approach to sentiment analysis on Twitter for automated polarity sentiment classification, utilizing an ensemble learning approach based on meta-level features. The chapter investigates the effectiveness of using a combination of existing lexicon resources as meta-level features in ensemble learning for sentiment classification.

Chapter 5 (Long-form Sentiment Analysis): This chapter presents and identifies sentiments expressed in online medical forums that discuss Lyme Disease. In this chapter, we first identify a complete and relevant set of categories that can characterize Lyme Disease discourse. Second, we present a feature-based model that consists of three different feature sets: content-free, content-specific and meta-level features. The chapter discusses employing supervised learning algorithms to build a feature-based classification model and assesses the feasibility and accuracy of the automated classification.

Chapter 6 (Application to an Alternative Disease Dataset): This chapter presents and evaluates a feature-based model (as outlined in Chapter 5) by assessing its ability to adapt to an online medical forum discussing Lupus.

Chapter 7 (Evaluation and Discussion): This chapter summarizes and discusses the findings of the empirical investigations presented in Chapters 5 and 6. It also assesses the performance of the proposed supervised feature-based model using statistical analysis.

Chapter 8 (Semi-supervised Approach - Modified Co-training): This chapter introduces an adaptation of a well-known semi-supervised learning technique, necessary because labelled training datasets are fairly expensive to obtain but unlabelled datasets are readily available. It describes the implementation of co-training by combining labelled and unlabelled data. We assess the performance of domain-independent and domain-dependent sentiment features (i.e., similar to the features used in Chapters 5 and 6) as two distinct sets of views of a post and utilize them in order to identify the multi-class sentiment expressed using an adaptive co-training model.

Chapter 9 (Long-form Content and Topic Analysis): This chapter presents a framework that utilizes two alternative techniques – *content analysis* and *topic analysis* – to analyse the same information as dealt with heretofore. We apply a content analysis to obtain a general view of the medical content available. In addition, we employ a latent Dirichlet allocation (LDA) strategy as an unsupervised generative model for detecting different topics represented by participants or patients in the health community.

Chapter 10 (Conclusion and Future Work): This chapter presents the main conclusions of the work of this thesis and outlines and discusses possible future extensions of the work presented.

1.4 Publications

Chapters of this thesis are based upon the following publications – as indicated where appropriate:

- R. Alnashwan, A. P. O'Riordan, H. Sorensen, and C. Hoare, "Improving sentiment analysis through ensemble learning of meta-level features," *in KDWEB 2016: 2nd International Workshop on Knowledge Discovery on the Web*, Sun SITE Central Europe (CEUR)/RWTH Aachen University, 2016.
- R. Alnashwan, H. Sorensen, A. O'Riordan, and C. Hoare, "Multiclass sentiment classification of online health forums using both domain-independent and domain-specific features," *in Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pp. 75–83, ACM, 2017.
- R. Alnashwan, H. Sorensen, A. O'Riordan, and C. Hoare, "Accurate classification of socially generated medical discourse," *International Journal of Data Science and Analytics*, vol.8, no.4, pp. 1–13, 2018.
- R. Alnashwan, H. Sorensen, and A. O'Riordan, "Classification of online medical discourse by modified co-training," in 2019 *IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, 2019.
- R. Alnashwan, H. Sorensen, and A. O'Riordan, "Medical social data mining and stakeholder perspectives,". (under review)

Chapter 2

Background

2.1 The Problem of Sentiment Analysis

Sentiment analysis (SA) is the computational analysis of people's opinions, sentiments, emotions and attitudes expressed towards a concept or entity – a task that presents a large problem space [5]. SA is referred to by many names: *sentiment analysis, opinion mining, review mining* [5][7], *attitude analysis, appraisal extraction* [7], *opinion extraction, sentiment mining, subjectivity analysis, affect analysis and emotion analysis*; all of the above are considered to come under the umbrella of *sentiment analysis* or *opinion mining*, albeit with slightly different emphases [5]. Tasks can be identified as detecting, extracting and classifying sentiments, opinions and attitudes regarding various topics as expressed in written language [7].

A Note on Terminology

Although there is a considerable variety of automated detection terminology in use – such as sentiment, opinion, affect, feelings, and emotions in text – there is little material differentiation between these subjective terms in prior research [8]. Hence, sentiment is attitude or thought prompted by feeling; opinion is a view or judgement about a matter; and affect is the most abstract notion related to feelings [8]. Using a broad interpretation, the term *sentiment analysis* has been used interchangeably with *opinion mining* to denote the same field of study [5][9]. A number of studies refer to *sentiment analysis* or *opinion mining* as a specific application that classifies text units by their polarity (positive, negative or, in some cases, neutral)

> [3][10][11]. However, Munezero et al. [8] provide a more complex definition of *sentiment analysis* and *opinion mining*: in this view, sentiments are emotional characteristics created over a period of time about specific subjects. Within this context, they identify affect as emotions and feelings expressed through the text. Abbasi and Chen [12], on the other hand, define *affect analysis* as a sub-category of *sentiment analysis* that considers the emotion/affect of the text, while they tend to use sentiment and affect interchangeably.
>
> The general view – and the one adopted in this thesis – is that the terms sentiment, affect and opinion have been used interchangeably in the literature, where they refer to the extraction of opinions, emotions or views that may be expressed in text.
>
> Finally, from a computational viewpoint, sentiment analysis, or opinion mining, refers to the application of techniques taken from various research domains, such as natural language processing (NLP) [9], machine learning (ML), information retrieval (IR), and data mining (DM), to extract subjective information from text [13].

Sentiment analysis has become very attractive to the research community due to evidence of its value within a broad range of fields, such as business, politics and healthcare. Furthermore, the vast amount of data available on the social web through user-generated content can have considerable value for society when assessed as part of opinion or sentiment analysis. The spread and richness of such data posted online – in reviews, microblogs, blogs, and forums – has led to an increase in the sharing of knowledge and opinions online and has been shown to influence social, political and economic behaviour worldwide [7]. Significant information can be detected within online data that can have both pragmatic (e.g., a business or organization assessing consumer opinion regarding its goods or services) and political (e.g., attempting to gauge social mores or movement on a policy issue) benefits [5]. Several works have pointed to the ability of socially generated content to influence and reflect, for example, political elections [14], market strategies, and box office revenue [15]. Social media content can also have an impact on individuals' decision making in relation to a product, policy or many other issues in life.

Sentiment and opinion analysis has recently become a major field of study; the factors behind this interest include [9]:

- The availability of well-established techniques, tools and algorithms that can play a significant role in analysing and interpreting data, e.g., machine learning algorithms.
- The vast amount of data available online due to the increasing spread of the Web.
- Recognition of the interesting and attractive intellectual challenges involved.
- How this field can offer smart applications that can have considerable value for society.

2.2 Primary Definitions

As mentioned at the beginning of the chapter, the main task of sentiment analysis is to study the opinions and sentiments contained in a given text. In this section, the problem is formally described in more detail:

Let D be a set of evaluated texts.

Each opinion document d $\in$ D includes three basic components, as defined by Liu [16]:

- Object or entity: a product, person, service, event, topic or organization on which opinions are expressed. Each object/entity can be described using various sets, subsets and so on in a hierarchy of relationships. Each object/entity or set/subset has its own attributes. For example, when a user writes a review about a laptop and its battery, screen, camera, etc., the laptop is the object that has a set of attributes e.g., size and weight, and a set of parts, e.g., battery, screen and camera. Battery, as a set, can have certain attributes e.g., battery life and battery size. For simplicity, all attributes and all set parts are referred to using the term aspect. The hierarchy is reduced to two levels: the root node is the entity and the leaf and sub-leaf levels are the various aspects of the entity.
- Opinion holder: the source that intends to express an opinion about a specific entity/object. The source could be a person or an organization. The opinion holders in reviews or blog posts are usually the authors of the document, while in news articles the opinion holder can be the person or organization explicitly stated and mentioned [17].

- Opinion: this refers to an attitude, view or appraisal of an object/entity expressed by an opinion holder. Opinions can be categorized into positive, negative or neutral sentiment and/or numeric values that represent the strength of the sentiment. These categories are called sentiment/opinion orientation/polarity, depending on the application domain conventions.

After describing the basic components of opinion, an opinion can be defined using a quintuple (e_i, a_{ij}, s_{ijkl}, h_k ,t_l) [5], where e_i denotes the name of an object/entity, a_{ij} is the aspect of the entity e_i, and the third component, s_{ijkl}, is the sentiment expressed about a_{ij} by the holder h_k in a certain period of time t_l.

From the above definition, most general opinion mining (or sentiment analysis) problems can be addressed by identifying the most essential information in the opinion and discovering all or most of the quintuple elements in an opinion document d. Sentiment analysis that depends on this approach can be called aspect-based sentiment analysis (or opinion mining) or feature-based sentiment analysis (or opinion mining) [5]. In this context, and from the opinion document d, many tasks can be expanded using the following approach: recognize and identify entities, extract aspects of an entity, identify the holders of the opinion, and analyse and evaluate sentiments of an opinion in the contextual source. Specifically, each of these components can be an individual task in sentiment analysis, depending on the area of focus of the research or the aim of the application.

2.3 Sentiment Analysis Research

Sentiment analysis is a very rich and topical research problem (as mentioned earlier) and many approaches, techniques and methods have been applied across different tasks to address the problem of sentiment analysis classification. In general, the sentiment analysis problem can be divided into four main dimensions:

- Sentiment analysis can be divided into different *sub-tasks*. Polarity detection (e.g. positive/negative), emotion detection and sentiment strength can all be significant dimensions of the sentiment analysis problem.
- Sentiment analysis *levels* determine the various *granularity* perspectives of the text units analysed (e.g., sentences, paragraphs or documents).
- Sentiment analysis *approaches* may differ, whereby a considerable number

of techniques and approaches are used to address the sentiment analysis problem. The techniques include machine learning, lexicon-based and hybrid approaches.

- A variety of sentiment analysis *domains* may exist, as opinions can be expressed about almost anything and anywhere in contextual forms. Some of the domains that are amenable to sentiment analysis are presented and reviewed later in the chapter.

We elaborate on these SA dimensions below.

2.3.1 Sentiment Analysis Tasks

One of the common tasks in sentiment analysis is polarity detection, in which the opinionated text is assigned a positive or negative sentiment orientation – and thus treated as a binary classification problem. In this task, sentiment orientation can be identified in a text in general or with regard to a specific entity (or aspects of a specific entity). Some studies go beyond binary classification. For example, they might follow the universal emotions from Plutchik's wheel [18]: joy, sadness, anger, fear, surprise, trust, disgust and anticipation [19].

From the previous sections, sentiment analysis has been shown to assume that a given text is opinionated. However, this is not always the case. One text might only have factual information, while another could include feelings, beliefs and/or attitudes. A text can refer to a subject when it expresses a personal feeling or any subjective information, while neutral information or facts can be treated as objective. A relevant task, that of identifying if a text is subjective or not, is known as a *subjectivity analysis* problem. For example, [20] investigate subjectivity analysis in a way that deals with identifying both factual and opinionative information. In addition, a number of studies apply subjectivity analysis first as a sub-task, then consider sentiment analysis for subjective data only; alternatively, subjectivity and polarity might be treated as independent aspects of the same problem, as done by [21]. Another task in sentiment analysis is to combine polarity with subjectivity analysis by adding a neutral category to positive and negative categories, as done by [22].

2.3.2 Sentiment Analysis Levels

Sentiment analysis can be applied to different text units. When implementing sentiment analysis, researchers tend to choose the text units according to the

goal of the study. For example, *document-level* sentiment classification is one of the most widely applied tasks in sentiment analysis problems, which generalizes *message-level* classification [21][23]. In other applications, classifying single sentences, instead of the sentiment expressed from a whole document, is more appropriate for addressing a problem. This task is called *sentence-level* sentiment classification. Some authors propose an effective approach to sentence-level analysis [24][25][26].

2.3.3 Sentiment Analysis Approaches

Various approaches, techniques and methods have been applied across various tasks to address the sentiment analysis classification problem. Approaches to sentiment classification can be broadly categorised into *machine learning, lexicon-based* or *hybrid* approaches [13] (with some overlap between categories).

Machine learning approaches can be based on *supervised* learning, whereby we train an algorithm on a set of a priori labelled data. Alternatively, *unsupervised* learning uses no labelled data; instead, an algorithm might exploit knowledge about words or phrases in order to identify sentiment within text (this is often referred to as a *lexicon-based* approach). Machine learning approach can also be *semi-supervised* – using a mix of labelled and unlabelled data to train a classification function. Previous research that employed these approaches, either individually or in combination, to conduct sentiment analysis is reviewed in Chapter 3.

2.3.4 Data Domains in Sentiment Analysis

The application of sentiment analysis to different kinds of online textual data from various domains has been used to automate or replace more established ways of identifying and summarizing opinions, such as surveying and focus groups. The types of textual data generated from different web sources include news articles, blogs, reviews, social media and micro-blog posts, and forum discussions [16]. Various applications of sentiment analysis have been used with both formal and informal text forms (the informal text genre poses further challenges for natural language processing).

2.4 Mining Social Media for Sentiment

The global phenomenon of social media has allowed individuals (and collectives) to communicate, add new connections, receive information, distribute content, share thoughts and opinions, and enhance their social life inside and outside their networks. Any negative impacts that it might have are largely ignored in the rush to participate. As a result, social media web services have now become an important source for gathering a variety of kinds of information for sentiment analysis [5] – simply because this is where the data is. If one wants to assess peoples' thoughts, opinions and feelings on issues that pertain to modern life, social media (or, more generally, the web) is the place to look.

2.4.1 Challenges of Social Media

In general terms, the content in social media is often written in a non-traditional style, which has presented several challenges to basic natural language processing (NLP) tools developed for traditional/formal text. A considerable amount of literature discusses how the informal nature of social media content impacts on the development of automatic language processes [5][9][16]. These challenges include the following:

- Content in social media can be written in unstructured and informal text, which can include colloquial language. Such content can include many creative forms of slang, the misuse of punctuation, ad hoc abbreviations and symbols, and misspellings – all of which could, for example, increase data sparseness.
- The web is a dynamic phenomenon. Information can change and involve interactions in different directions among different participants. Observing and keeping up to date with these changes are important issues for various NLP applications.
- Sentences in written text can be expected to be poorly structured and include grammatical errors, unlike formal media sources such as online news.
- In some cases, text can present misleading, unclear or mixed sentiment content and might include sarcastic and metaphoric sentences which are hard for NLP to process.

In this context, the above-mentioned characteristics of the text genre in social

media show considerable contrast with edited and traditionally formal text – thus increasing the need to develop linguistic techniques and tools that can specifically target and process the social media text genre.

2.4.2 Social Media Sub-Genres

This thesis focuses on analysing sentiment in socially generated text data. We initially look at short-form sentiment analysis: detecting sentiment contained in Twitter data. We chose Twitter because it is the most-used microblogging service and provides an enormous number of freely-available short messages written by users. While this presents interesting research problems, we came to the conclusion that an increasing amount of what people now produce on social media sites (particularly in medical discourse) is long-form text – and that can require different analysis strategies. We present a study based on a machine learning model to tackle three different problems: to identify a complete and relevant set of categories that can characterize focused disease discourse; to test and investigate strategies, both individually and collectively, for automating the classification of medical forum posts into those categories; and, as labelled training datasets are fairly expensive to obtain and unlabelled datasets are readily available, we propose a semi-supervised learning technique.

2.4.2.1 Twitter

According to [27], "Similar to blogging, Twitter is a real-time network that allows users from across the globe to share information through private and public messages that are organized chronologically on a particular user's account". Other researchers [22] describe Twitter as a microblogging platform that has an enormous number of short messages written by users. Twitter is the most popular microblogging service and has shown significant growth since it was launched in 2006[2]. It has become one of the most widely used platforms among people who use social media, regardless of their age, gender, or where they come from; and it has been the focus of much research into sentiment analysis [22][28]. Various conventions have been adopted by Twitter users, using features such as retweets, replies, and hashtags in their tweets. Among many Twitter features, a user can send a tweet privately or publicly and the Twitter reply feature allows the user to respond to other users, indicated by @username. Retweets are the republished content of a tweet. Hashtags are used to predefine the content of a

[2] $https://about.twitter.com/en_us/company.html$

posted tweet, specified by adding # as the prefix of a word, and enabling users to search, find or follow topics or themes of interest.

As a microblogging service, Twitter has several inherent challenges as a social media data source. While the tweets users posted were initially restricted to 140 characters, they were expanded to 280 characters in 2017 to relax constraints of expressibility (in this thesis, we consider tweets posted before the character limit was extended). Due to the informal text style on Twitter, it is hard to trace specific events, as people can post information about anything and everything [29]. The characteristics of Twitter and the limited length of communication has led to the use of Internet slang and abbreviations [30], which includes several abbreviations, acronyms and emoticons that are not common in traditional media, e.g., omg, lol, 2u2, or :).

2.4.2.2 Online Forums

Sentiment analysis can be used to attack more complex and difficult problems when analysing a longer text unit, such as found on online forums. These forums have found widespread use and have resulted in large repositories of text data, which can include discussions with respect to a wide range of domains and subjects (health, politics, travel, etc.). Within online forum communities, thousands of communications can take place daily, with members of varying educational, knowledge and social backgrounds discussing several issues with others across a large geographical distance. This leads to a considerable range of submissions, including long and short posts, information-providing and information-seeking, general and technical content, structured and unstructured data, etc. Members of such a community participate in order to provide and receive information and support through discussing aspects of their particular issue. Forums can provide information that is unique in nature and which is not available elsewhere (e.g., "Can I mix different forms of magnesium? Or is it better to use only one?" from general health forum) and have become an increasingly popular resource for interested users to share, discuss or ask about life's problems.

In recent years, there has been a growth in the number of online forum communities related to different topics – for example health (such as MedHelp[3] and Dailystrength[4]) and travel (e.g., TripAdvisor[5] and Travellerspoint[6]). The

[3] $https://www.medhelp.org/$
[4] $https://www.dailystrength.org/$
[5] $https://www.tripadvisor.ie/ForumHome$
[6] $https://www.travellerspoint.com/forum.cfm$

characteristics of online forums differ somewhat from those of microblogging and other types of social media. For example, a forum may include a hierarchy structure (forums can comprise several sub-forums), any one of which can be the root of lower-level sub-forums. At leaf level, one can have topics – user-initiated discussions on a particular item of concern (see Figure 2.1). A user post can relate to a new topic or can form a reply to a previous post submitted by another user (referred to as a thread).

This type of online domain presents several challenges, in addition to those mentioned previously: there is no length limit on text, which can include several sentences and/or paragraphs; the text can contain a great deal of irrelevant information; and the text can reflect ambiguous and/or contradictory opinions within different part thereof. The combination of such problems leads to a diversity of research approaches, as will be made clear throughout this thesis.

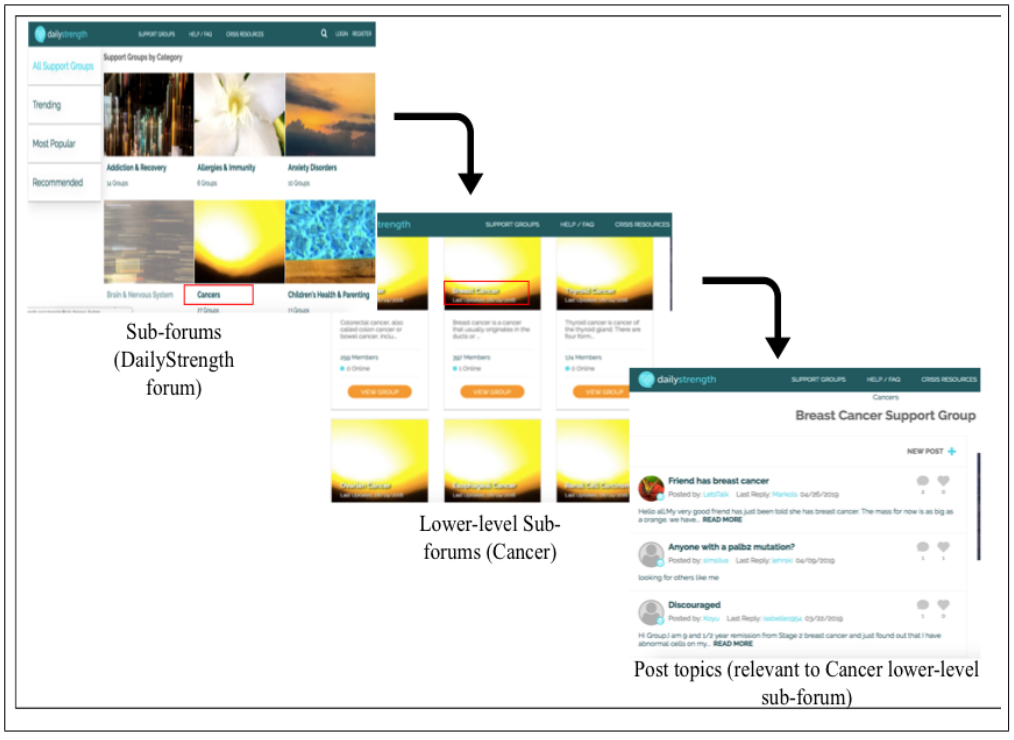


Figure 2.1: Example of the hierarchy structure of DailyStrength forums

2.5 Summary

This chapter defined the concept and problem of sentiment analysis. The general overview of the dimensions of sentiment analysis explored in the chapter included SA tasks, levels, approaches and the domains and types of data in SA. A

number of challenges in SA were explained in relation to socially generated data. Our SA framework focused on analysing sentiment in short and long socially generated text data, as described in this chapter.

In the next chapter, we provide a broad review of research within the field of sentiment analysis. A number of approaches to sentiment classification are described, as well as various applications.

Chapter 3

Literature Review

3.1 Sentiment Classification: Prominent General Approaches

Various approaches, techniques and methods have been applied across different tasks to address the sentiment classification problem. According to Maynard and Funk [31], sentiment classification can largely be divided into two main approaches: *machine learning* (ML) and *lexicon-based*. Machine learning applies classification algorithms along with a feature set derived from the source text, whereas a lexicon-based approach relies on a predefined list of terms annotated by sentiment score (referred to as a *sentiment lexicon*). A lexicon-based method can be implemented either by a dictionary- or corpus-based approach [32]. Furthermore, combining ML and lexicon-based approaches in a hybrid approach – or leveraging the use of a lexicon as a feature in ML – has been shown to be an effective method of sentiment classification; in this regard, a sentiment lexicon is an important factor in most methods in sentiment classification[32].

3.1.1 Machine Learning

Text classification using ML can be categorized as *supervised* (algorithm learns using a large set of trained, labelled data), *unsupervised* (no labelled data) or *semi-supervised* (only a limited set of labelled data is available) learning – the technique adopted clearly depends on the availability of a trained dataset.

Most sentiment analysis depends on a *feature set* to determine sentiment. Feature sets and their combination play an important role in sentiment analysis problems.

Various types of feature sets have been explored in the literature in order to determine the sentiment (or affect) of units of text.

3.1.1.1 Feature Engineering

Generating appropriate features is one of the most important and challenging aspects of applying supervised learning techniques. According to Brownlee, "Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data" [33]. An effective feature set can have a great influence on a model's performance: so, an important factor in a sentiment classification task is to engineer an appropriate set of features [34]. The position of feature engineering in machine learning workflow can be seen in Figure 3.1.

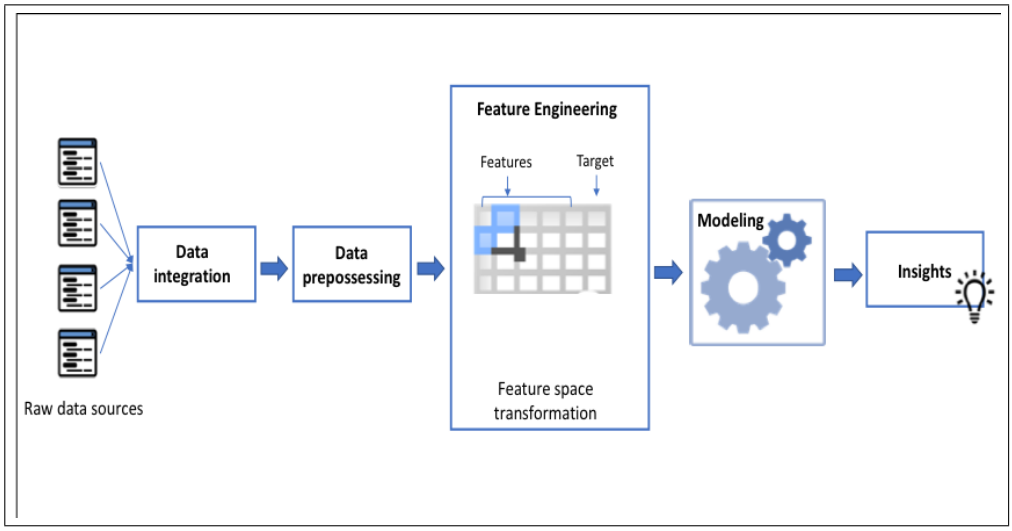


Figure 3.1: Feature engineering position in ML work flow

Three feature categories that are commonly used in sentiment analysis studies are presented in more detail below: *syntactic features, semantic features,* and *stylistic features*, the first two being the most used [35][30].

3.1.1.1.1 Syntactic Features

We consider syntactic features to include n-gram and part-of-speech tags, among other features. With regard to n-grams, it is contentious among researcher as to whether or not these can be considered as syntactic in nature. We follow the approaches taken by Abbasi et al. [35] and Giachanou and Crestani [30], who considered n-grams as one type of syntactic feature.

An n-gram feature: this is one of the most common and effective features. In this, $\{f_1, f_2, f_3, \ldots . f_m\}$ is a representation of m predefined features from a document. These features can be individual words (unigram) (e.g., "nice", "car"), a sequence of two words (bigram) (e.g., "nice car") or a sequence of n words (n-gram). The n-gram features extracted from a document, d, can be represented by the feature vector d=(n_1(d),n_2(d),...), where n_i(d) is the feature value for the f_i word (term) occurring in d [36]. The values can be binary (1 if the gram is present and 0 if not), they can be the frequency values (the number of times the gram occurs in a document) or the values can be derived from a weighting scheme (such as TF-IDF).

For example, Pang et al. [36] applied a supervised binary classification approach to classify film reviews into positive and negative classes. Their approach showed that using unigram features with binary-valued document vectors performed better than frequency values in that vector. Furthermore, they examined unigram features and bigram features and their combination; among all implemented features and their combinations, the best results were achieved by using unigram features.

Part-of-speech (POS) tags: (e.g., noun, verb, adjective) can capture some knowledge opinion expression, mostly by finding adjectives that are considered to express opinions [30]. The extent of the effectiveness of this feature has been controversial. For example, some studies report non-improvement in a model that uses POS tags, e.g., [10] and [29], whereas others have reported a small improvement after applying them, e.g., [22] and [11].

3.1.1.1.2 Semantic Features

One of the key concepts in text classification is detecting semantic information represented in text. One of the best-known semantic features is the opinion or sentiment of words or phrases [30]. This feature incorporates a manual, semi-automatic or fully automatic annotation process to aggregate the values of polarity or affect intensity corresponding to words and phrases, thus generating a sentiment lexicon. Techniques for generating such a lexicon will be addressed in Section 3.1.3. Semantic features can also include contextual features, which rely on predicting sentiments given the semantic orientation of the surrounding words.

3.1.1.1.3 Stylistic Features

Stylistic features have been reported as an important aspect of text analysis [35]. These include lexical features, typically involving word/character-based analysis [37], and structural features, which refer to the text layout, such as the separation into paragraphs of a long social media text. Stylistic features have been used in authorship identification, such as by Abbasi and Chen [38] and Li et al. [39]. However, stylistic features are less common in sentiment analysis research [35].

3.1.1.1.4 Content Specific vs. Content Free

Feature sets can be categorized into content-specific and content-free features. Content-specific features are composed of words that have significance for a specific domain – e.g., the words "treatment" and "rash" have specific meanings in a medical context (including Lyme Disease). Syntactic features can be considered content-specific features, as they are composed of words (i.e., n-grams) that have value for a specific community. Studies have shown higher text-classification performance after including content-specific features [40][38]. In contrast, content-free features include lexical (based on a character and word statistical analysis of the text lexical variation) and structural (based on organisation of the text) features [40]. According to [20], structural features have been among the most informative features to indicate the subjectivity in text.

3.1.1.2 Supervised Learning Approach

One of the most common approaches to sentiment classification is to formulate a problem using supervised learning: apply machine learning algorithms to gain knowledge of the application domain by presenting sufficient previously-evaluated data. Samuel defined machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed" [41]. The idea behind the supervised learning approach is to train machine learning algorithms that are capable of determining the sentiment orientation of unseen data by giving sentiment-labelled data. Thus, both a *training* and *test* dataset are required, the first labelled by a domain expert and the second to be labelled by the algorithm. In a sentiment classification problem, supervised learning algorithms have been applied effectively to various domains, which include film or product reviews [40], political [31] and medical [42].

Most supervised machine learning relies on three fundamental factors: 1) engineering effective feature(s), i.e., identifying syntactic/semantic/stylistic features;

2) implementing a machine learning classifier model (algorithm); and 3) the quality of the annotation process of the training data.

While feature engineering is essential for machine learning and can lead to better performance, the model itself and the data can also affect the result. Sentiment classification can be developed as a supervised learning problem with binary classes (e.g., positive/negative) [3][10][11] or multiple classes [43]. A number of standardized datasets have been developed for testing and comparing classification techniques [29][44]. Other researched extract their own dataset, depending on the problem and the availability of the data [22][10][43].

In sentiment analysis, a number of standard supervised learning algorithms have been applied to determine the classes of the sentiment [34]. The most frequently applied classifiers include support-vector machines (SVM), Naive Bayes (NB), logistic regression (LR), random forest and decision forest techniques.

For the sentiment analysis in this study, we distinguish prior research according to the text-lengths that have been analysed, as follows: 1) *short-form document* corpora, such as Twitter; and 2) *long-form corpora*, such as blogs, reviews and forums.

3.1.1.2.1 Short-form corpora

Table 3.1 summarises important research to date on short-form document corpora. Several sentiment polarity classification studies have focused on the analysis of short-form corpora, such as microblogging. For example, Go et al. [3] were among the first to perform sentiment classification on Twitter data in order to classify tweets into positive or negative classes. They used unigram and bigram words in conjunction with POS tags as features to train NB, MaxEnt (maximum entropy), and SVM classifiers. They found that POS tags were not effective, that unigram features outperformed the use of bigram features among all the trained classifiers, and that combining unigram and bigram features trained by MaxEnt outperformed all the other models. According to Go et al. [3], the performance scores were higher when combining unigram and bigram features, as using a bigram feature alone increased the sparseness of the feature space. Similarly, Bermingham and Smeaton [10] evaluated the sentiments in short- (i.e., Twitter and micro reviews) and long-form reviews using an SVM and NB classifier. Their findings were contrary to those of Go et al. [3], as they reported that using only a unigram feature with NB achieved the best performance for short forms. However, this was not the case for long forms, for which more

complex features, such as POS n-grams, were more effective at distinguishing binary classes.

In a similar vein, n-gram features have been investigated in the sentiment analysis of Twitter data. Unlike the aforementioned, Pak and Paroubek [22] tackled the problem as a three-way classification task (positive, negative and neutral) and compared the performance of NB, SVM and conditional random field (CRF) classifiers. Their evaluation results demonstrate that a bigram feature performed better than a unigram or trigram. This indicates that bigrams represent features with a balance between coverage and sentiment pattern, when compared with unigram and trigram features, respectively. A more sophisticated POS was also investigated, distributing the estimated probability of the presence of POS tags within different sets of text and using it to measure posterior probability. Pak and Paroubek concluded that the NB model in conjunction with an n-gram POS distribution outperformed SVM and CRF classifiers trained only with n-gram words on a Twitter dataset. As they stated, POS tags can present good insight into text that represents sentiment expression [22]. In a multi-class classification context, Miller et al. [43] investigated different classification algorithms: a decision tree (J48), multinomial Naive Bayes (MNB), Bayesian networks (Bayes Net) and sequential minimal optimization (SMO), using an SVM with unigrams extracted from tweets. They compared the performance of the different classifiers in classifying tweets into four different classes. The MNB classifier is one of the simplest classifiers but produced the best results of all those evaluated. The authors suggest that this might be explained by the data evaluated by expert annotators having less noise [43].

Syntactic features, including n-grams, have shown improvements in text sentiment classification. Despite the work done on extracting only syntactic features, classification performance can be improved by adding other types of feature. As an example, Agarwal et al. [11] explored a combination of syntactic, stylistic and source-specific features (microblogging features). Their task included two-way classification (positive versus negative) and three-way classification (positive versus negative versus neutral), which were examined by training an SVM on a Twitter dataset. Two models were investigated: a feature-based model and a tree kernel model. The feature-based model was based on a unigram feature and what they call a senti-feature, the senti-feature integrating several types of feature divided into polar and non-polar features. Polar features refer to those features that use an external resource to calculate prior polarity, whereas a non-polar feature is one that is irrelevant to the prior polarity. Furthermore,

POS was examined, as this can be a polar or a non-polar feature. The tree kernel model was used to measure all possible correlations of each fragment features and their combinations. Integrating unigram with the senti-feature was the most effective model for two-way classifiers, whereas adding the kernel and senti-feature models outperformed the others in the three-way classification task. Of the feature tested in the feature-based model, prior polarity of words with their POS tags proved to be the most important feature. This finding was also reported by Kouloumpis et al. [29], from their investigation of Twitter data trained by an AdaBoost algorithm in a three-way classification problem. They found that the n-grams, together with the lexicon features (which are similar to the polar feature in Agarwal et al. [11]) and the microblogging features were the most important.

In a similar vein, Mohammad et al. [44] won a competition at SemEval-2013[7] (a semantic evaluation workshop/competition), achieving first in performance among 44 participants: they had the best performing model for sentiment analysis based on training an SVM on a large combination of features in a message-level task, using both Twitter and SMS. The combination generated the following features: word n-grams (1, 2, 3, 4-grams), character n-grams (3, 4 and 5 characters), POS tags, lexicon features, all-caps, hashtags, punctuation, emoticons, elongated words (such as "yessss") and clusters (word clusters). The lexical features were extracted from both manually and automatically generated lexicons. The research investigated feature performance by including/excluding them in the full combination in the model. It was concluded that the most effective feature was the sentiment lexicon, which concurred with the finding by Agarwal et al. [11] and Kouloumpis et al. [29]. However, unlike Agarwal [11], removing the hashtags, emoticons, and elongated words features had almost no impact on performance.

Approaches that integrate sentiment lexicon resources and POS as features in supervised classification schemes have also been studied. For example, Bravo-Marquez et al. [21] combined 10 existing sentiment analysis methods and resources as a feature set in a supervised classifier that focused on certain aspects, such as polarity, strength and emotion. Sentiment analysis was performed on Twitter dataset using four different machine learning algorithms: Naive Bayes, logistic regression, multilayer perceptron, and SVM. The results showed that a lexicon-based approach was the most effective for polarity classification, while

[7] $http://www.cs.york.ac.uk/semeval-2013/task2$

POS approaches were more suitable for subjectivity classification.

Table 3.1: Summary of studies using a supervised learning approach for short-form data

Study	Features				Algorithms	Classification task			Type of data	Source of data
	Syntactic	Semantic	Stylistic	Source- specific		Binary (P/N)	3-way (P/N/A)	Multi-class		
[3] Go et al. (2009)	Unigram, bigram, POS	–	–	–	NB, MaxEnt, SVM	✓			Twitter	Own (STS)
[10] Bermingham and Smeaton (2010)	Unigram, bigrams, trigrams	–	–	–	SVM, NB	✓	✓		Twitter and Blippr	Own
[22] Pak et al. (2010)	Unigram, bigrams, trigrams, POS	–	–	–	NB, SVM, CRF		✓		Twitter	Own
[43] Miller et al. (2017)	Unigram	–	–	–	J48, NB, Bayes Net, SVM,			✓	Twitter	Own
[11] Agarwal et al. (2011)	Unigram	Senti-features			SVM	✓	✓		Twitter	Own
[29] Kouloumpis et al. (2011)	Unigram, bigram, POS	Lexicon feature	–	Micro-blogging features	AdaBoost		✓		Twitter	STS, ETC
[44] Mohammad et al. (2013)	Word/character n-grams, POS, negation, clusters	Lexicons	–	Micro-blogging features	SVM		✓		Twitter	SemEval-2013
[21] Bravo-Marquez et al. (2014)	POS	Lexicons	–	Micro-blogging features	NB, LR, multilayer perceptron, SVM	✓			Twitter	STS, Sanders, SemEval-2012

3.1.1.2.2 Long-Form Corpora

Long-form corpora have also been investigated in sentiment analysis tasks, examples of which can be found in Table 3.2. Long-form corpora include online film/product reviews and online discourse in forums and blogs. The text characteristics differ from short-form corpora in terms of the quantity/proportion of sentiment expressed, structure, text style and level of informality. Bermingham and Smeaton [10] performed a study to investigate and compare the performance of sentiment classification for long- and short-form online corpora. The experiment was conducted using an SVM classifier and an MNB classifier for Twitter and blog sentiment analysis. They found that MNB showed greater accuracy than SVM for Twitter, whereas SVM achieved higher performance for blogs. The researchers concluded with the finding that classifying short-form corpora was an easier task than for long-form corpora.

One of the pioneering works in supervised sentiment classification is by Pang et al. [36]. The authors performed binary classification (positive/negative) to classify an online movie review dataset. The syntactic features utilized in their model were unigram, bigram and POS features, these being integrated with NB, maximum entropy and SVM classifiers. The strongest performance was obtained by SVM as a learning algorithm with a unigram feature. Several variations influenced by this approach have been proposed [45][46]. Bermingham and Smeaton [10] reported that the best performance, based on different datasets, was achieved by using a purely unigram feature and, in contrast with Pang et al. [36], trained on MNB classifiers.

Generating appropriate features is one of the most important and challenging aspects of applying supervised learning techniques. Thus, sophisticated models have been proposed for studying the sentiment analysis of long-form corpora with an expanding features type, in addition to syntactic features. For example, Abbasi et al. [35] investigated a multilingual (viz., English and Arabic) sentiment classification model. The proposed model was based on syntactic and stylistic features, used to classify reviews and forum posts into positive and negative classes. In addition to syntactic features (such as n-gram and POS tags), stylistic features were also utilized (such as lexical, structural, and function word-style markers). Efficient features that would enhance the model's performance were identified by developing an entropy weighted genetic algorithm (EWGA). They reported that the highest performance using SVM was reached when both stylistic and syntactic features were applied in conjunction with the EWGA feature selection on either a review dataset [36][46] or online forums.

One of the key concepts in text classification is analysing the semantic information represented in text and a number of different techniques have been developed for lexicon adoption as a feature in online data [42][24][47]. For the polarity classification domain for online forums, Ali et al. [24] studied sentiment expressed in online forum posts by combining syntactic and several lexicon features to understand the semantic observation of sentences in posts. From their experiments, they conclude that lexicon clues of polarity are more effective than counting features (such as the number of adverbs or adjectives) where both are based on lexicon feature; also logistic regression is the best classifier among others for the selected features.

Sentiment analysis can be conducted with a multi-class classification model that can accurately identify sentiments across a large range of long-form unstructured texts. For example, Bobicev et al. [42] performed sentiment analysis of online forums related to reproductive technologies based on a domain-dependent lexicon (HealthAffect), which the authors developed to aid their classification task. The results show that the HealthAffect lexicon outperformed other existing lexicons in a multi-class classification task conducted on forums from an IVF website dedicated to reproductive technologies. Similarly, Lu [47] successfully classified online posts into three general categories. Three classification methods were used: C4.5, Naive Bayes and an SVM, with n-gram-based features which included the traditional bag-of-words (BoW) model, domain-specific features that extract semantic types from posts, and, lastly, a sentiment analysis feature that includes terms with the highest polarity score measured using the SentiWordNet lexicon. The best scoring model reported was when an SVM classifier was trained on a combination of those feature, while utilizing feature reduction techniques to remove irrelevant features. Zhang et al. [48] performed a hierarchal multi-class classification with two cascading SVM algorithms using pattern and word features to identify user intent in online forums.

Another model in sentiment analysis is based on sequential pattern algorithms. For example, Melzi et al. [26] conducted a comprehensive sentiment analysis at the sentence-level. Their Sentiment analysis was based on the following: 1) binary classification to determine polarity, 2) multi-class classification based on six different emoticons, and 3) multi-labelled classification based on several associated classes that text unit may have. They used the following: syntactic features (such as unigrams, unigrams and bigrams), emoticon features (such as smiley emoticons being classified according to six emotions; emotion context based on neighbours' emoticons and overall emoticons; emotion words, by count-

ing the number of words that corresponded to each emoticon), and frequent patterns in online posts. The feature component was enhanced by identifying patterns based on a sequence pattern algorithm that used statistical means to determine the most frequent patterns. From the results, it was concluded that the best feature set, when integrated with an SVM classifier, was a combination of unigrams and bigrams with emotion words only, whereby the patterns feature were, according to their justification, insufficient and inappropriate due to their being extremely broad.

Table 3.2: Summary of studies using a supervised learning approach for long-form data

Study	Features				Algorithms	Classification task			Type of data	Source of data
	Syntactic	Semantic/ Lexical	Stylistic	Source-specific		Binary (P/N)	3-way (P/N/A)	Multi-class		
[10] Bermingham and Smeaton (2010)	Unigram, bigrams, trigrams	–	–	–	SVM, NB	✓	✓		Blog posts and online review	TREC Blogs06 collection[8], Pang and Lee's movie review [46]
[36] Pang et al. (2002)	Unigram, bigrams, POS	–	–	–	NB, maximum entropy, SVM	✓			Online review	Own
[35] Abbasi et al. (2008)	Unigram, bigrams, trigrams, POS	Character level, Word level	Structural features	–	SVM	✓			Online review, forum posts	Pang and Lee's movie review [46], own
[24]]Ali et al. (2013)	–	Lexicon-based features	–	Rule-based features	NB, SVM, LR		✓		Forum posts	Own
[42] Bobicev et al. (2015)	–	Lexicon-based features	–	–	NB, SVM, Decision Trees, KNN			✓	Forum posts	Own
[47] Lu (2013)	Unigram, bigrams, trigrams	Lexicon-based features	–	Domain-specific features	SVM, C4.5, NB			✓	Forum posts	Own
[48] Zhang et al. (2014)	Unigram	Pattern Features	–	–	LIBSVM			✓	Forum posts	Own
[26] Melzi et al (2014)	Unigram, bigrams	Emotion words, motion context, patterns	–	Smileys, Amplifiers	SVM	✓		✓	Forum posts	Own

[8] C. MacDonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical report, University of Glasgow, Department of Computing Science, 2006.

3.1.1.2.3 Annotation Process for the Trained Data

Data classification approaches that utilize supervised machine learning techniques require an annotated dataset to train the algorithms. For example, product review data can be considered as training data and testing data, whereby each review contains reviewers' free text comments and a star rating. The data can be transformed into input text and target value, respectively. Specifically, if the rating scheme runs from one to five stars, the high-value star reviews indicate positive sentiment, low-rated reviews indicate negative sentiment and reviews rated with three stars can receive a neutral sentiment label. Several studies have followed this process to annotate film review [35][36] and product review datasets [40].

However, it is a different case for data other than review data – and this has become one of the limitations of sentiment classification in supervised learning. Sentiments in user-generated data can be subjective and the ground-truth sentiment categories of text data are mostly identified by human evaluators. Several text classification studies have manually labelled the evaluated datasets – for example, by training algorithms from manually labelled forum post data (long-form) [24][26] for a polarity classification task and a multi-class classification task [42][48]. Others [11][44] have followed the manual annotation process for the polarity classification of Twitter data (short-form). Manual (human) annotation is needed due to label sparsity. However, manual annotation can be costly, time-consuming and labour-intensive. To tackle this problem, researchers have proposed a range of methods for obtaining sentiment annotation with a minimum of human engagement, such as the *distant supervision* approach.

Distance supervision makes use of features extracted from data to label the training data. One method is to leverage the marked-up text as a noisy indicator for labelling the training data, e.g., the use of emoticons. The idea of using emoticons to label the training data used to train supervised learning was proposed by Read [45], whereby each text passage that includes a smile emoticon is labelled with a positive orientation and the existence of a frown emoticon is labelled as a negative orientation. In the Twitter sentiment context, Go et al. [3] followed this approach and used emoticons to label training data. From the Twitter API, they collected 1.6M tweets containing emoticons. The presence of a positive or negative emoticon indicates a positive or negative tweet, respectively. Table 3.3 presents a list of positive and negative emoticons. The authors argue that building automatically annotated data using emoticons as polarity indicators results in a performance comparable to labelling polarity according to a star-

scale. Similarly, Pak and Paroubek [22] adopted an emoticon procedure to generate labelling data. In addition to using positive and negative emoticons to label data, they included a neutral class derived from objective Twitter messages posted by news accounts.

Table 3.3: List of emoticons used in [3]

Emoticons mapped to :)	Emoticons mapped to :(
:)	:(
:-)	:-(
:)	: (
:D	
=)	

Hashtags in social media can also be indications of positive and negative attribution expressed by the message writer. For example, Kouloumpis et al. [29] relied on hashtags to automatically annotate a Twitter dataset with the most-related hashtags, indicating positive, negative and neutral for three-way polarity classification.

However, when emoticons are used to label data as sentiment, indicators have a certain level of noise, which makes it very hard to achieve high performance in text classification [49]. Hence, due to noise and other issues, such as sparsity of label data, manual labelling of data can be required to overcome these issues [24][11][44][26].

3.1.1.3 Semi-Supervised Learning Approach

Supervised learning has several benefits in sentiment detection: it has the ability to deal with large and noisy data; and fully automated development and implementation can achieve a high performance in comparison with complex linguistic or other approaches. However, one of the big limitations in supervised learning is its sensitivity to the quality and quantity of the training data.

As an alternative to supervised learning, semi-supervised learning can constitute an alternative approach to machine learning, as it learns from both labelled and unlabelled data in the training stage. Most existing semi-supervised learning approaches are derived from the following techniques:

(1) Expectation maximization (EM), which deals with incomplete data by calcu-

lating maximum likelihood estimates in an iterative manner [50];

(2) Graph-based algorithms, which utilize labelled data to regularize the learning process. For example, Pang and Lee [46] successfully applied the process to sentiment analysis; and

(3) Co-training [51], which requires datasets to be described by features that have a natural separation into two distinct sets.

According to Yu [23], EM in a semi-supervised setting is limited by the mixed-model assumption, while graph-based semi-supervised learning is not ideal when processing large-scale data. Consequently, in this thesis, the focus is on co-training for a more generalizable model.

3.1.1.4 Deep Learning Approaches

With the rapid progression in text classification problems, various approaches have emerged in recent years. For example, deep learning has emerged as a subsidiary field of machine learning that uses multilayers of artificial neural networks with hidden computational power [52]. For text classification and representation, word embedding has been leveraged to produce the representation of text in deep learning. Recently, D'Andrea et al. [53] used deep learning to detect stances on Twitter towards the topic of vaccination. A convolutional neural network (CNN) and a long short-term memory (LSTM) network were used for classifying tweets into the following: in favour, not in favour and neutral towards vaccination [53]. Wang et al. [54] proposed a capsule tree-LSTM model for classifying sentences. The authors employed a dynamic routing algorithm to develop sentence representation by automatically learning the different weights of each node.

Hybrid deep learning approaches have also been explored. For example, Ajao et al. [55] leveraged two such models to detect fake news on Twitter. In their study, the authors employed a CNN–LSTM hybrid and compared its performance with simple LSTM. Simple LSTM demonstrated the best results, although the authors argue that the performance of the hybrid approach will be improved by enriching the size of the training data.

3.1.2 Lexicon-Based Approach

The lexicon-based approach relies on the knowledge embedded in an opinion-lexicon as the main indicator in sentiment classification. Positive opinion words

indicate a desirable and pleasant view, whereas, negative opinion words express an undesired view. An opinion lexicon can include opinion words, phrases and idioms [32].

Generally, lexicon-based approach can be unsupervised (does not require a training process from labelled data, relying purely on the lexicon) or a hybrid (combining the use of a lexicon with the training processes of the labelled data).

The simplest, and most common, way of classification using an opinion lexicon involves identifying the presence of one or more known opinion words. For example, in subjectivity classification at the sentence level [25], the sentence is identified as subjective when it includes one or more of a predefined set of adjectives; if not, the sentence is determined as objective. In addition to identifying binary classes (e.g., subjective/objective or positive/negative), an opinion lexicon can be used to define an opinion score, either by counting the number of words/phrases that exist or by summing the values from an opinion lexicon, such as in Bobicev et al. [42].

In the hybrid approach, where sentiments are classified by aggregating labelled data with the lexicon knowledge, two different strategies have been adopted. The first utilises the opinion lexicon as a feature in a supervised setting – specifically, each sentiment analysis lexicon with their outcomes are included as a feature vector for each text being classified. Examples of this can be found in [21][29], as mentioned in Section 3.1.1.2. The second type of hybrid approach classifies the document from the knowledge in the opinion lexicon or from the training classifier on annotated data – or by combining both their predictions. For example, Balage Filho et al. [56] apply a hybrid classification approach that has two emoticon lexicons as their rule-based classifier and SentiStrength as a lexicon-based classifier; a third classifier is a Support Vector Machine (SVM), a standard machine learning classifier. The study assigns a confidence threshold in each of the classifiers to achieve the overall confidence level required.

3.1.3 Sentiment Lexicon Generation

In the research literature, opinion words are also referred to as sentiment words [34]. An opinion lexicon is a dictionary that includes a combination of opinion words and their associated sentiment categories or a value that corresponds to the polarity intensity of a word or phrase.

There are three main approaches to compiling or collecting an opinion lexicon:

a manual, a dictionary-based, and a corpus-based. The manual approach is time-consuming and labour-intensive, as it requires effort to create an opinion lexicon by assigning a corresponding opinion based on human judgement to each word in a dictionary. Alternatively, the other two approaches aim to automatically construct a lexicon by incremental induction from an initial seed set.

The dictionary-based approach depends on seeding opinion words and finding their synonyms and antonyms from a known dictionary in order to create a list of opinion words. For example, Hu and Liu [57] started with a seed sample of manually annotated words with positive/negative sentiment, and the induction process was then applied by searching in a known lexicon (i.e., WordNet) for their synonyms and antonyms to expand the opinion lexicon iteratively. The iterative process stops when no new opinion word can be found.

The third approach is corpus-based, which starts with a list of seed opinion words and extends the lexicon by learning the opinions associated with the words in a larger corpus in order to find domain opinion words that are related in a context-specific orientation. This approach can be done by utilizing semantic orientation or statistical methods. One example of a semantic orientation lexicon is proposed by Saif et al. [58], in which lexicon generation is based on word meaning as denoted by the SentiCircle lexicon. An example of a statistical method of lexicon induction is proposed by Turney [59], whose approach is based on patterns of co-occurrences. The main observation is that if two words appear frequently in the same context, they will have the same polarity; thus, we can calculate the polarity of target words by calculating their relative frequency of co-occurrence with another word using point-wise mutual information (PMI) score.

It is worth noting that the corpus-based approach can also be used for generating a general-purpose sentiment lexicon if it is generated from a very large and very diverse corpus. However, regardless of the length of the corpus, it is unlikely to achieve the completeness of a dictionary-based approach [5].

3.2 Applications of Social Media Opinions

With the expansion of the use of social media globally (e.g., reviews, forum discussions and microblogs), individuals, organizations and companies are increasingly utilizing its content as an influential source of knowledge that benefits various stakeholders. The field of sentiment analysis aims to reveal opportu-

nities for a broad range of domain applications. Researchers have addressed sentiment analysis across several domains, such as business, politics, social life and medicine, to enhance aspects of everyday life. In this section, selected examples of general applications of sentiment analysis are reviewed. A more comprehensive review of sentiment analysis as applied to the medical domain is also made, this being the domain of focus in this thesis.

3.2.1 General Applications in SA Using Social Media

The growth of user-generated data in social media can provide considerable value by fulfilling business and individual needs. For example, there is now less need for companies and organizations to conduct surveys, opinion polls or focus groups in order to determine public opinion regarding their products or services. Individual customers can also easily find and access previous and existing customers' opinions and reviews related to a product on the web. For example, individuals can use online review services to share and read members' experiences of hotels and travelling (e.g., TripAdvisor and Yelp), books and goods (e.g., GoodReads and Amazon), or film reviews (e.g., Internet Movie Database [IMDb]) to help in their own decision making. Automated sentiment analysis of socially produced data can be of significant value, such as by identifying a particular aspect of an information space, determining themes that predominate within a large dataset, and allowing people to summarize opinions within a big dataset. For example, Pang et al. [36] proposed sentiment analysis of IMDb data to classify film reviews regarding their overall sentiment (i.e., to identify a review as either positive or negative).

Conducting sentiment analysis in a business context can provide value to companies by enhancing customer satisfaction as a result of the monitoring and analysis of customer reviews. For example, Kang and Park [60] visualize and measure customer satisfaction with regard to mobile services. Their experiment assessed 1,487 reviews from eight different mobile application services. Another study developed satisfaction analysis for cosmetic product to improve customer decision making using twitter data [61] and review comments [62].

In the political domain, microblogging platforms are employed widely for political deliberation. Two common approaches have been taken to forecasting elections: focusing on the volume of online messages related to specific candidates or issues; and/or conducting sentiment analysis on these messages [63]. For example, Tumasjan et al. [64] attempted to predict election results

by conducting sentiment analysis on 100,000 messages posted on Twitter and concluded that the quantity of the messages related to each party reflected the subsequent election results. On the other hand, another study relied on a lexicon-based method of sentiment analysis on Twitter to calculate sentiment orientations [14]. A comprehensive study integrated both approaches by investigating the volume of messages and using a sentiment analyser [65] to assess Twitter data. The research demonstrated that the sentiment analyser outperformed the volume of messages approach for inferring the results of political elections. Contrary to these studies, [66][67][68] argue that predicting political elections from social media (especially Twitter) is not as good an indicator as traditional election polls.

3.2.2 Applications of Socially Generated Medical Data

With the spread and richness of online media in the last decade, analysing such data has contributed to learning more about public opinion on health-related matters [42][48]. Text mining techniques and sentiment analysis have been applied to implement automated approaches that can provide considerable value to society. Sentiment analysis in the medical domain can be approached from various directions [6]. Many of the applications are designed to study and analyse particular medical issues, such as a specific disease or adverse drug reactions; there are also a few applications dedicated to more general health issues. The following sections present some of the work conducted on the sentiment analysis of medical text, and summaries of those example studies can be found in Table 3.4.

3.2.2.1 Mining Personal Health Information Polarity and Opinion

Sentiment analysis techniques have been implemented to identify sentiment polarity, intent or affect related to a general or specific concern. Several pieces of research address the polarity of the sentiments expressed by participants or patients on social media. For example, Biyani et al. [69] performed polarity sentiment classification in a semi-supervised setting on user posts in an online cancer support community, Cancer Survivor Network, to aid understanding of the dynamics of that community in providing effective emotional support to participants. Ali et al. [24] studied opinions (positive, negative and neutral) towards hearing aids by analysing posts related to hearing loss from three different health forums (MedHelp, alldeaf and Hearing Aid Forums). Wang et al.

[70] studied the use of supervised learning strategies in the field of depression, the aim being to distinguish between depressed and non-depressed users of microblogs by utilizing features obtained from psychological research; the overall goal was to facilitate the reduction in suicide. Another application studied the sentiment and opinions in microblog posts related to traditional Chinese medicines (TCM) [71]. As TCM has been widely debated and is controversial, it was deemed interesting to try to understand public opinion by conducting binary classification to separate posts into supporting or opposing TCM.

3.2.2.2 Analysing Emotions and Studying Emotional Affects

In a wider context, implementing a multi-class classification model that can detect sentiment and affect has been undertaken across online medical communities. Some studies have conducted multi-class sentiment classification for medical text based on one of a number of well-known emotional categories. For example, Melzi et al. [26] analysed user-generated medical text user-generated data based on Ekman's 1992 research on emotion categories. Others did not tend to follow any standard emotion categories [43][48].

Lu [47] successfully classified online posts from communities of breast cancer survivors and activists, using three general topic categories - treatment, emotional support and survivorship - to increase the efficiency of patient search queries and engagement in an online health community. Miller et al. [43] performed a twofold classification: they first explored tweets about the Zika virus and then employed automated classification to sort them into four disease-specific categories. Their intention was to organize the data by categorizing them (into symptoms, transmission, prevention, and treatment) so that public health officials could observe the issues of most concern when the categories were assigned. Bobicev et al. [42] also conducted sentiment analysis in two stages: first, by identifying a relevant set of categories that could be used to characterize emotion expressed in forums on an in vitro fertilization (IVF) website dedicated to reproductive technologies, which generated encouragement, gratitude, confusion, facts and endorsement categories; second, they performed sentiment analysis based on a domain-dependent lexicon (HealthAffect developed by the authors themselves) to aid their classification task and to identify the most dominant sentiments in posts. Furthermore, Zhang et al. [48] used previous medical research to guide user intent in online health forums. After generating five classes (manage, cause, adverse, combo and story), they utilized a supervised

learning method to recognize users' needs in posts from the HealthBoards online health community.

From a broader perspective of emotion categorization in socially generated medical discourse, some studies adapted one of a set of known predefined categories. For example, Melzi et al. [26] proposed an analysis approach to online health forums to detect emotions based on six emotion categories (based on Ekman's 1992 work): anger, disgust, fear, joy, sadness and surprise. They argued that their approach to identifying the polarity of a text (positive, negative) with the associated emotions could help health professionals to get closer to and understand patients.

3.2.2.3 Measuring the Quality of Document Content or Health Interaction

Measuring the quality of medical data is another application of sentiment analysis in the medical domain. For example, Denecke and Nejdl [72] measured the quality and credibility of user-generated medical content. Based on the information content in a medical weblog, a classification method was used to sort posts into affective or informative. A post was classified as affective when describing daily actions or expressing emotions regarding treatment or a disease; it was informative when providing facts and information related to a disease or treatment. The study found that there was considerable difference in the content of health-related web resources. For example, weblogs can be an important source of material about disorders, physiology and medication, whereas Wiki pages and encyclopedias offer more information about anatomy and procedures.

3.2.2.4 Analysing Drug Medical Data

One recurring feature in online medical discourse is the assessment of treatment outcomes, both positive and negative. Analysing the sentiments expressed in such discussion can be valuable for both health professionals and patients [24][6][73]. Patients commonly express their experience regarding specific drugs or treatments – potentially having the benefit of identifying risks posed by drugs. For example, in a sentiment approach applied to online drug reviews with regard to polarity classification, the text was analysed using an incremental model to create an opinion lexicon [73]. Also, to better understand public opinion of drugs, Na et al. [74] proposed clause-level sentiment classification based on a linguistic rule-based approach for drugs reviews in online forum discourse. They successfully detected the polarity of each clause in relation to

six different aspects related to drugs: overall opinion, effectiveness, side effects, condition, cost, and dosage. Patient stories about drugs and treatments on discussion forums can be analysed for earlier detection of adverse drug effect [6].

3.2.2.5 Detecting Health Care Quality

From another perspective, understanding patients' health care experience is essential to the process of providing care – and one of the main pillars of health care quality [75]. Consequently, assessing health care quality is another application for online medical data. For example, identifying patients' opinion of health care from free online comments (i.e., the National Health Service [NHS] website) was investigated by Greaves et al. [75]; this research conducted sentiment analysis to predict the polarity of public opinion of hospitals in the UK based on three different aspects: does the patient recommend the hospital (overall recommendation)? is the hospital clean (cleanliness)? and were the hospital staff respectful and was the patient treated with dignity (dignity)? Similarly, tweets can contain considerable information related to care quality and can be used to enhance the standard of care [76]. This conclusion was drawn after analysis of tweets related to NHS hospitals, the results of which was used to detect the polarity of different themes regarding individual hospitals, such as effectiveness, safety and fundraising.

Table 3.4: Summary of examples of studies on the application of sentiment analysis work in the medical domain

Study	Task	Level	Textual source	Application
[69] Biyani et al. (2013)	Polarity	Document	Forum posts	Mining personal health information polarity and opinion
[24] Ali et al. (2013)	Polarity	Sentence	Forum posts	
[70] Wang et al. (2013) [71] Shen et al. (2015)	Polarity	Document	Sina Weibo[9]	
[47] Lu et al. (2013) [48] Zhang et al. (2014) [42] Bobicev et al. (2015) [26] Melzi et al. (2014)	Multi-class classification	Document	Forum posts	Analysing emotions and studying emotional effects
[43] Miller et al. (2017)	Multi-class classification	Document	Twitter	
[72] Denecke (2009)	Binary	Document	Blog posts	Measuring the quality of document content or health interaction
[73] Asghar et al. (2014)	Polarity	Document	Online reviews	Analysing drug medical data
[74] Na et al. (2012)	Aspect sentiment analysis	Clause-level	Online reviews	
[75] Greaves (2013)	Aspect sentiment analysis	Document	Forum posts	Detecting health care quality
[76] Greaves (2014)	Aspect sentiment analysis	Document	Twitter	

[9]Sina is a social network service similar to Twitter.

3.3 Summary

This chapter provided an overview of several approaches to sentiment analysis, as well as the different characteristics and requirements of each method. Textual data can be transformed into several features so as to generate a representation that improves classification performance. We presented existing works on sentiment analysis, highlighting valuable results as well as weaknesses. Selected examples of general applications of sentiment analysis were reviewed and a detailed examination of their application in the medical domain was undertaken.

The next chapter presents our proposed framework for sentiment analysis of short-form socially generated data (Twitter data).

Chapter 4

Short-form Sentiment Analysis Using Ensemble Learning

4.1 Introduction

This chapter examines techniques for analysing short-form text, in particular Twitter. The main contribution of this chapter is to investigate the effectiveness of using a combination of existing lexicon resources and sentiment analysis methods as meta-level features in ensemble learning for sentiment classification. Ensemble learning is a fruitful strategy for categorizing short form data that could offer advantages over using either a single lexicon resource or a single classifier.

The outline of the chapter is as follows: Section 4.2 provides a description of our model of sentiment analysis expressed as a binary classification problem, placing it in context with some related work. In Section 4.3, we present the experimental evaluation of the proposed model. Then we illustrate the experimental setup details and results in Section 4.4 and present our conclusions in Section 4.5.

4.2 Improving Sentiment Analysis Through Ensemble Learning of Meta-Level Features

With regard to sentiment in short-form text, we focus on polarity: in this case, a binary classification of positive or negative. Our proposed sentiment classification approach consists of two steps: first, we identify a combination

of sentiment analysis methods and lexicons, which we call meta-level features; and second, we implement an ensemble method that adopts multiple supervised classifiers, as shown in Figure 4.1.

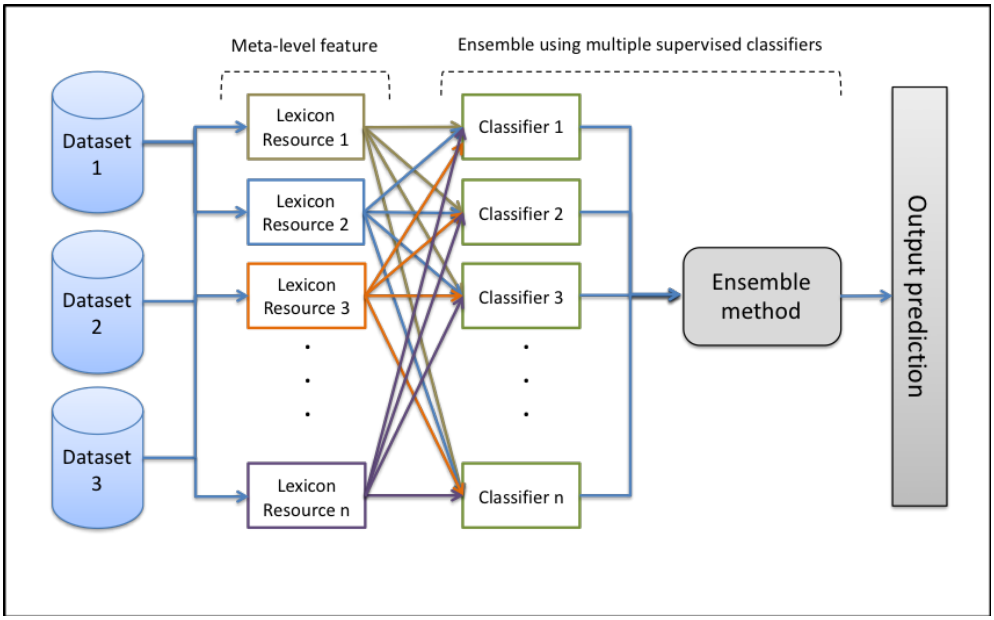


Figure 4.1: Schematic of the approach

4.2.1 Feature Engineering

4.2.1.1 Feature Hashing

In general text mining applications, a feature engineering task is needed to transform the input data – comprising various pieces of text content – into feature vectors that can include, for example, the frequencies of words/phrases.

The model used in feature representation is a feature hashing technique with n-grams. The feature hashing model converts streams of words into a set of integer features and vectors thereof, by creating a hashing dictionary that consists of n-gram features calculated using the terms present in a text. The advantages of using feature hashing is that it reduces the dimensional space and leads to faster lookup of feature weights for supervised learning, by representing text documents as equal-length numeric feature vectors. The hashing feature technique was developed using the Vowpal Wabbit library for 32-bit murmurhash v3 hashing[10] [78].

[10] https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-hashing#how-to-configure-feature-hashing

Feature hashing is set to a default hash table with bitsize=10, which can be adequate for many problems[11] , to generate a dictionary of unigrams and bigrams (i.e., a single word/phrase and a sequence of two words/phrases).

4.2.1.2 Meta-Level Features

Meta-level features are based on combining different existing methods and lexicon resources. Supervised, semi-supervised and unsupervised learning, as well as manual approaches, are used on these features. Furthermore, they represent different sentiment dimensions, polarity and strength. When the resources and methods ascertain the polarity of each tweet, the number of features in each lexicon resource can be calculated by finding those words that match between the text and the lexicon resources. If the method and lexicon resources present a strength value for each included word, then the feature is represented by the sum score of those values. The outcomes of all the aggregated lexicon resources are presented as a feature vector, as shown in Figure 4.2. Each row represents a tweet and the columns represent the features. As the model that was investigated was based on polarity sentiment classification, the meta-level feature includes polarity lexicons. This is in line with Bravo-Marquez et al. [21], who stated that the lexicon-based approach is the most effective for polarity detection. The polarity lexicon features are a combination of seven existing lexicon resources: SentiWordNet, the Bing Liu Lexicon, AFINN, NRC Hashtag, the Sentiment140 lexicon, the Sentiment140 method, and SentiStrength. The combination offers advantages over using a single lexicon resource. These features can help to address the nature of twitter data characteristic problems: grammatical shortcomings, abbreviations, etc. Table 4.1 shows the type and number of features extracted and the range of values for each. We explain each of the polarity lexicon features below. Algorithm 1 is an outline of how we use the lexicon features.

4.2.1.2.1 SentiWordNet

SentiWordNet 3.0 is a lexicon source for sentiment classification and opinion mining developed by Baccianella et al. [79] and is an improved version of SentiWordNet 1.0 proposed by Esuli and Sebastiani [80]. SentiWordNet 3.0 is based on WordNet 3.0, a well-known lexicon database for English, in which words are grouped into sets of synonyms, called synsets. SentiWordNet annotates

[11]https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-hashing

all synsets with a value between 1 and 0 to indicate the positivity, negativity or neutrality of each synset. This lexicon was developed by semi-supervised classification and a random walk process [79]. The lexicon is freely available to researchers. We extracted two features from the SentiWordNet lexicon: the positive value and the negative value (the pseudo-code is shown in Algorithm 4.3).

4.2.1.2.2 Bing Liu's Opinion Lexicon

This lexicon was generated and developed by Bing Liu and was applied in various papers of which he is the author or co-author [5]. The lexicon includes misspelled words, slang and some morphological variants. It has 2,006 positive and 4,783 negative words. The positive and negative features extracted for this research were the number of positive and negative words in each tweet that match Bing Liu's lexicon.

4.2.1.2.3 AFINN

AFINN-111 is an improved version of AFINN-96. The original version was called ANEW (Affective Norms for English Words) and was developed prior to the introduction of microblogging platforms [81]. It was generated using people's psychological reactions. A new version that would better suit the language used in the text of microblogging platforms was needed, so Nielsen created the AFINN lexicon [82]. Positive word scores are between 1 and 5 and negative word scores range from -1 to -5, as an indication of the strength of a word. The lexicon has 2,477 English words. We extracted two features, positivity and negativity, which are the rating value of all the words in tweets that match the AFINN-111 lexicon. As it was specifically designed and updated using microblogging sources (i.e., Twitter), it provides effective and informative features for evaluation datasets.

4.2.1.2.4 NRC Hashtag

The NRC Hashtag Sentiment lexicon resource was proposed by Mohammad et al. [44]. The lexicon was created automatically by adopting the use of 78 hashtags with positive and negative sentiments in words such as #angry, #joy and #sadness. The lexicon was generated from a set of about 775,000 tweets and a sentiment label was defined for those tweets according to their hashtags: positive or negative [44]. The NRC hashtag lexicon uses pointwise mutual information (PMI), which measures the degree of association between each word and the polarity label of a tweet in order to calculate the associated

sentiment score for each word found in the tweet dataset. A positive score points to a positive sentiment and a negative score to a negative one. Using this lexicon, we can extract positive and negative features by finding the matching words between the NRC Hashtag lexicon and the posts, then adding the values.

4.2.1.2.5 Sentiment140 Lexicon

The Sentiment140 lexicon was created by the same group that produced the NCR Hashtag [44] lexicon and was generated by following the same approach to calculating word sentiment. However, instead of using hashtags to indicate polarity, tweets are labelled according to their positive or negative emoticons. Sentiment140 was formed from a set of 1.6 million labelled tweets proposed by Go et al. [3]. We extracted the feature as we did for NRC Hashtag.

4.2.1.2.6 Sentiment140 Method[12]

The Sentiment140 Method is a web application for classifying a given text by its polarity. It provides an application program interface (API) for assigning a polarity to tweets. It was generated by using distance supervised learning techniques on 1.6 million tweets [3]. The method's founders considered emoticons in tweets as noisy labels for training data in sentiment analysis classification. Sentiment140 Method provides a single output (i.e., polarity value) for each post.

4.2.1.2.7 SentiStrength

SentiStrength is a web application for automatic sentiment analysis that evaluates the strength of short text [83]. It uses supervised and unsupervised learning methods [83]. From SentiStrength resources, we extracted a feature that includes polarity value.

[12] `http://www.sentiment140.com/`

AFINN positivity	AFINN negativity	Bing liu Positivity	Bing liu Negativity	NRC Positivity	NRC Negativity	Sen140 lexicon Positivity	Sen140 lexicon Negativity	Sentiment method Polarity
4	0	2	0	6.411	-13.578	16.493	-1.643	2
2	0	1	1	13.103	-4.918	7.784	-1.53	2

Figure 4.2: Sample of features vector space

Table 4.1: Features from lexicon and polarity resources

Field	Lexicon source	Extracted features	Description	Range of values
Polarity	Bing Liu	positive	Count the number of positive words that match lexicon	{0,. . . ,n}
		negative	Count the number of negative words that match lexicon	{0,. . . ,n}
	Sentiment140 method	polarity	Method polarity label	{0,2,4}
	SentiStrength	polarity	Method polarity label	{-1,1}
Strength	SentiWordNet	positive	Total of positive words values that match the lexicon	{0,...,∞}
		negative	Total of negative words values that match the lexicon	{0,...,∞}
	AFINN	positive	Total of positive words values that match the lexicon	{0,. . . ,n}
		negative	Total of negative words values that match the lexicon	{-n,. . . ,0}
	NRC Hashtag	positive	Total of positive words values that match the lexicon	{0,...,∞}
		negative	Total of negative words values that match the lexicon	{-∞,. . . ,0}
	Sentiment140 lexicon	positive	Total of positive words values that match the lexicon	{0,...,∞}
		negative	Total of negative words values that match the lexicon	{-∞,. . . ,0}

Algorithm 1: Meta-level features

```
Input :tweetStream, L
    For tweet ∈ tweetStream do
        Words ← tokenise(tweet)
        For word ∈ words do
            If hasWord (word, L_positive) then
                .
                .
                .
            If hasWord (word, L_negative) then
                .
                .
                .

    end
```

Figure 4.3: Meta-level feature Algorithm

Because of the characteristics of the language used on Twitter, some pre-processing steps are required in order to reduce the feature dimensional space. The first step involves removing links, punctuation, special characters and digits and replacing them with white space. Then, all capital letters are converted to lower case to unify the data format. Finally, letters that are repeated more than twice in a sequence are replaced with two occurrences, a technique used by Go et al. [3] (e.g., “greeeeeeat” or “greeeat” is converted to “greeat”).

4.3 Experimental Evaluation

4.3.1 Classifiers

Four classifiers were used as our base learners to fulfil the sentiment classification task: a two-class support vector machine, a two-class Bayes point machine, two-class logistic regression and a two-class decision forest. The reason for selecting those four classifiers in our study is that they have been widely used in previous sentiment analysis research and are known to perform well in text classification problems [42][21][24][10][44][26].

We considered various supervised learning algorithms in conducting our classification experiment, as each classifier deals with data using different knowledge and results in different patterns of errors. In our experiment, classifiers were built and trained to predict unseen test data. After the base learners were trained, our ensembles operated on a majority voting method, which is one of the most

common ensemble techniques used in classification tasks [84].

4.3.1.1 Two-Class Support Vector Machine

A support vector machine (SVM) [85] has been found to be one of the most successful classifiers in binary text classification, and specifically in sentiment analysis [21][11][44]. An SVM is a non-probabilistic binary linear classifier that can accommodate a large number of features in high-dimensional feature space. During the training process, the algorithm aims to identify the maximum-margin hyperplane, which is chosen from among several hyperplanes, that aims to increase the margin between the decision boundaries in the two classes on either side [1], as shown in Figure 4.4. To do this, the algorithm divides the input data of N number of instances into two classes, of positive and negative instances, representing them as points in this space. The algorithm attempts to assign each new instance to one of the predefined categories, mapping them to the same space. The algorithm can then map input data/features into high-dimensional space using a *kernel function*.

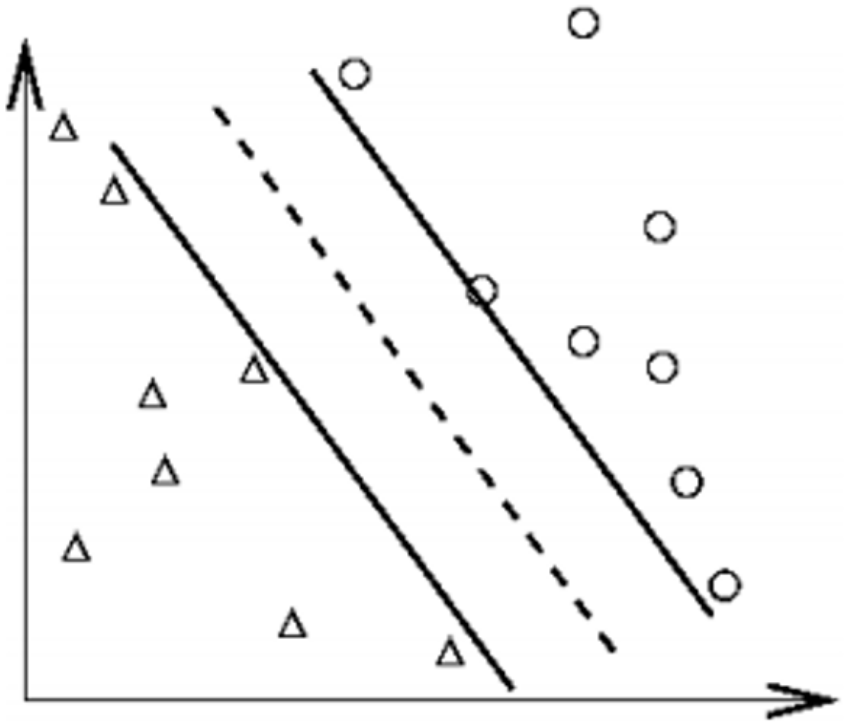


Figure 4.4: A general representation of SVM algorithm [1]

4.3.1.2 Two-Class Bayes Point Machine

Bayes point machines [86] are linear classification algorithms that use a Bayesian approach. The general idea is that, given a set of linear classifiers, we can determine one "average" classifier (the Bayes point), which can approximate the theoretically optimal Bayesian average in terms of generalization performance. In the training stage, it employs a uniform prior probability distribution, while training instances define a posterior distribution or weight vector. Given new test instances, the Bayes point assigns them real-valued output that represents the prediction measures confidently. The "hardest" pattern in the training process appears to have the most significant effect on the final expansion coefficient [86].

The implementation of a Bayes point machine in the Microsoft Azure Machine Learning Studio (MALMS) includes two improvements on the original algorithm:[13] (i) the use of the expectation propagation message-passing algorithm; and (ii) a parameter sweep and normalization process are no longer required in the implementation stage. As a result of those improvements, the Bayes point machine enhances the robustness of the model, reduces the need for parameter tuning and is much less prone to overfitting data.

4.3.1.3 Two-Class Logistic Regression

Logistic regression is a widely known statistical method in machine learning schemes for classification tasks [21][24]. The algorithm aims to predict the probability of the binary target of a given observed value by fitting a logistic function to the data. The predictions in logistic regression are calculated using the logistic function. Logistic regression fits an S-shaped curve that predicts the target value between 0 and 1 of any given input, as can be seen in Figure 4.5

The algorithm attempts to identify the optimal values of the coefficients using maximum-likelihood estimation for the input data. The algorithm uses a commonly employed model to optimize the parameters, known as L-BFGS limited-memory, which scales well with high-dimensional data. Further technical description can be found in Andrew and Gao [87].

[13] `https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-bayes-point-machine`

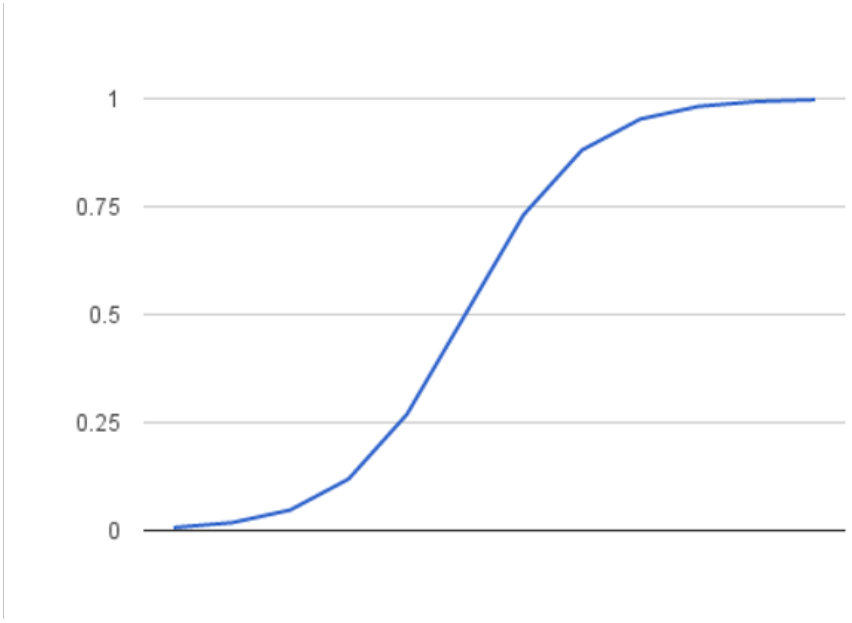


Figure 4.5: S-shaped curve in Logistic regression [2]

4.3.1.4 Two-Class Decision Forest

The two-class decision forest [88] model is based on decision forest algorithms and, in itself, constitutes a fast ensemble model. It can scale well to cope with noisy data and high-dimensional features and can deal with non-linear boundaries.

The model operates by creating individual models and combining them to classify new instances. There are several ways to construct an ensemble for a decision forest model. Most commonly, the decision forest conducts the ensemble by implementing multiple decision trees and combining their results in a voting scheme, whereby the most popular class is the one selected.

The model starts by creating multiple decision tree classifications based on the full datasets provided. However, each classifier will begin from a different and randomized starting point. Decision trees solve problems by constructing a tree representation, subdividing the feature space into an internal node or a root that corresponds to a class label and employing a statistical method that defines the order of the features in the decision tree as either a root or internal node.

After generating the output of a non-normalized frequency histogram of target labels from each decision tree classifier in the decision forest, the aggregation process scores the outputs and normalizes them in order to calculate the probability related to each target label. The trees with a higher prediction confidence

will be assigned a greater weight in the final prediction result in the decision forest.

All the supervised learning classifiers used in this experiment utilized Microsoft Azure Machine Learning Studio implementations and all parameters were set to their default values [14].

4.3.2 Classifier Ensemble for Tweet Sentiment Analysis

Ensemble learning is a technique in machine learning that trains multiple learners, referred to as base learners, to solve a problem [84]. According to Dietterich [89], there are three significant reasons for using ensemble learning: 1) statistical: when the result relies on combining classifiers, it can reduce the chance of selecting the wrong classifier; 2) computational: some of the learning algorithms are based on a local search, where it is possible to become stuck in local optima—by applying an ensemble with many different classifiers, we may achieve a better approximation than with any single classifier; and 3) representational: when the hypothesis space does not represent the true target function, an ensemble can expand that space to give a better approximate. While, employing an ensemble approach will not always guarantee a better result than the best base learner, it will decrease the error rate of selecting a poor classifier by outperforming random selection [90]. To achieve an effective ensemble, two elements should be considered: the diversity and accuracy of each classifier [84][91].

The idea of combining multiple supervised learners to achieve predictive classification has been explored by the research community – and is an approach that has gained attention within sentiment classification tasks. Su et al. [92] applied an ensemble learning algorithm approach, stacked generalization, to sentiment classification. They achieved good results by applying five different supervised learning techniques to three different domains. Clark and Wicentwoski [93] applied another ensemble learning approach – a combination of multiple Naïve Bayes classifiers, in which each is trained using a single type of feature – n-grams, sentiment lexicons, parts of speech or emoticons – and assigning weight to words that have repeated letters. The confidence-weighted voting scheme was then used to classify binary sentiment at the phrase level. Wang et al. [94] combined three different approaches, namely, bagging, boosting and random subspace; five supervised learning algorithms were used as base classifiers with a bag-of-words

[14] `https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-classification`

feature.

In our experiments, we investigate the potential benefits of combining multiple base supervised learners, whereby tweets and all the extracted features described previously are fed in as vectors of sentiment features. In order to evaluate the proposed approach, labelled data are needed to train the model and evaluate the performance. As previously stated, we concentrated on a polarity classification task: either positive or negative. A voting algorithm was used by combining four base learners into an ensemble learning system: SVM, Bayes point machine, logistic regression and decision forest. The meta algorithm vote then makes the final decision as in Catal and Nangir [95]. A voting algorithm can be implemented with various rules, such as majority voting, maximum probability, average probability and the multiplication of probability. In this study, we used majority voting as the voting algorithm, whereby the majority class results from each classifier are used to identify and classify the label for the data. The majority vote algorithm was used as it gave the best performance compared with other voting rules, in the work by Catal and Nangir [95].

4.3.3 Datasets

To evaluate the effectiveness of our approach, we consider three labelled datasets: Stanford Twitter Sentiment (STS), SemEval-2016 and Health Care Reform (HCR). These datasets are suitable for performance assessment because a positive, negative or neutral label has been assigned to each tweet. Example of positive and negative tweets from each dataset are given in Table 4.2. The datasets are further described below and the number of positive and negative tweets in each dataset is summarized in Table 4.3.

4.3.3.1 Stanford Twitter Sentiment (STS)

Stanford Twitter Sentiment[15] was proposed and constructed by Go et al. [3]. The dataset contains 1.6 million tweets that are automatically labelled as positive or negative, using their emoticons as noisy labels. For the training data, the scraper sends a query such as ":)" to the Twitter API (to find tweets with positive emoticons that express positive emotions) or ":(" (a negative emoticons, which express negative emotions). The full list of emoticons can be found in Go et al. [3]. The tweets were collected from 6 April 2009 to 25 June 2009. We randomly selected 12,000 of those tweets.

[15]The data source was a forum on the Sentiment140 website.

Table 4.2: Examples of annotated Tweets from the datasets

Database	Example	Labelled
Stanford (STS)	im back to the online world!!!yey!!!feel like a druggie getting another hit....pure pleasure!	Positive
	there is no way i can go to school today im way to sick	Negative
SemEval-2016	I Love TEEN WOLF, can't wait till January when it comes back on.	Positive
	I'm sat watching the worst Harry Potter film ugh	Negative
Health Care Reform (HCR)	I miss America...the American people are still strong, but half of our leadership has abandoned us...sad, but true #tcot #ocra #hcr #tlot	Negative
	The #hcr is one of the best things to ever happen in America. Free healthcare is the way forward. #thatisall	Positive

4.3.3.2 SemEval-2016

This dataset is provided by the Semantic Evaluation of Systems Challenge (SemEval-2016), which sets challenging tasks for researchers who are interested in semantic analysis problems [96]. The challenge provides a dataset, in which each tweet has been annotated manually as positive, negative or neutral. In this work, we use only tweets labelled positive or negative.

4.3.3.3 Health Care Reform (HCR)

The Health Care Reform (HCR) labelled dataset was created by Speriosu et al. [97]. It was collected by extracting tweets that had the hashtag "#hcr" (health care reform) in March 2010. The authors then manually annotated a subset of collected data for four classes: positive, negative, neutral and irrelevant. In this work, we focus on classifying tweets into positive and negative; thus, we exclude neutral tweets.

4.3.4 Experimental Setup

The model was developed using Microsoft Azure Machine Learning (MALMS), a cloud-based service that facilitates developers and data scientists in building

Table 4.3: The Twitter Dataset statistics

	Positive	**Negative**	**Total**
STS	5,999	6,001	12,000
SemEval	4,385	1,415	5,800
HCR	542	1,380	1,922

and developing machine learning models. This software platform is hosted on a fully managed cloud that can deploy the model into production as a web service – which can then be accessed by any device and can use different data sources.

MALMS suits our purposes for several reasons:

- It has flexibility in sourcing and pre-processing data. Program code in Python or R can simply be embedded into it.
- It includes linear and non-linear algorithms, and models can be developed using classification, regression, anomaly detection, and clustering. Furthermore, it can deal with different types of data types: Boolean, categorical, numeric, string, datetime and timespan. It includes visualization to represent the developed model.
- Azure Machine Learning has a very effective cloud-based predictive analytics service, which is fast and creates simple solutions from predictive models.
- It is also easy to learn because the interface is based on dragging and dropping the nodes needed to create the experiment.
- Furthermore, any experiment can be shared with others, who can then easily extend your work.

The MAMLS platform, requires the following as input data:

$$X_{[N*d]}, Y_{[N*1]} \tag{4.1}$$

Where X refers to N data samples in a d-dimensions data space and Y refers to the actual class label for each data sample [98]. We used randomized data splitting, which is commonly employed as a rule of thumb, in which 70% of the data sample is used to train the model, and the remaining 30% of the data is used as test data to evaluate the performance.

To assess performance, we utilized a matrix for overall accuracy, precision, recall and F1-score. We defined an evaluation binary matrix B (tp,fp,tn,fn), where tp is the number of true positives, fp is the number of false positives, tn is the number of true negatives, and fn is the number of false negatives, and N is the total number of examples in dataset. Overall accuracy is computed using formula 4.2:

$$\textit{Overall accuracy} = \frac{(tp + tn)}{N} \tag{4.2}$$

Precision, which is the fraction of correctly classified positive examples over all the examples that are classified as positive, is calculated using formula 4.3:

$$\textit{Precision} = \frac{tp}{(tp + fp)} \tag{4.3}$$

Recall (also called sensitivity, or true positive rate), the fraction of correctly classified positive examples over all the positive examples, is calculated using formula 4.4:

$$\textit{Recall} = \frac{tp}{(tp + fn)} \tag{4.4}$$

F1-score, also called the F-measure, is used to combine the precision and recall into a harmonic mean, as shown in formula 4.5:

$$\textit{F1-score} = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{4.5}$$

In some of our data, as shown in Table 4.3, the class distributions were skewed (i.e., the number of positive and negative tweets was not balanced); thus, we performed random oversampling to assign a higher weight to the minority class and to avoid biasing the classifiers towards one specific class [99]. An oversampling technique randomly duplicates a selected sample of instances from the minority class to increase the population and uses all instances in the majority class.

4.4 Results

The analysis was carried out in three phases: 1) construct the ensemble; 2) apply meta-level features on each classifier; and 3) combine them into the ensemble using meta-level features on each base learner. We compared the results from the three approaches with stand-alone classifiers, which were set as our baseline. We evaluated the model that combines a meta-level approach with an ensemble approach to address the potential for improving performance.

The results of the polarity classification tasks are shown in Table 4.4 for each of the three datasets. The best performance baseline classifier for the SemEval-2016 and HCR datasets is the two-class decision forest, whereas, for the Stanford dataset, it is two-class logistic regression. The meta-level approach outperformed the baseline by just over 5% in terms of accuracy, F-measure and the average for both the Stanford and SemEval datasets. The meta-level improvements were less than 5% in the HCR dataset, which might indicate that the HCR dataset is easier to classify than either Stanford or SemEval.

According to the results, the ensemble with meta-level features shows better outcomes when compared with the original ensemble. The ensembles in both cases scored better for accuracy than the average of their base learners. Thus, applying our approach could avoid the common classification task problem of a poor selection of classifier. By considering the average of the polarity tasks, we observed that there is no significant difference between the best classifier, the two-class decision forest, and the proposed ensemble approach, which scored 82.4% and 81.0%, respectively.

Table 4.4: Polarity classification performance

		Stanford dataset					SemEval dataset					HCR				
		Acc	**Prec**	**Rec**	**F**	**Avg**	**Acc**	**Prec**	**Rec**	**F**	**Avg**	**Acc**	**Prec**	**Rec**	**F**	**Avg**
Baseline	**Two-Class Support Vector Machine**	0.643	0.629	0.668	0.648	0.647	0.729	0.737	0.73	0.733	0.732	0.721	0.654	0.705	0.678	0.69
	Two-Class Bayes Point Machine	0.659	0.653	0.656	0.654	0.656	0.759	0.767	0.757	0.762	0.761	0.788	0.724	0.792	0.757	0.765
	Two-Class Logistic Regression	0.667	0.654	0.682	0.668	0.668	0.758	0.762	0.765	0.763	0.762	0.759	0.695	0.753	0.723	0.733
	Two-Class Decision forest	0.63	0.617	0.652	0.634	0.633	0.852	0.891	0.808	0.847	0.85	0.82	0.748	0.857	0.799	0.806
	Average	0.65	0.638	0.665	0.651	0.651	0.775	0.789	0.765	0.776	0.776	0.772	0.705	0.777	0.739	0.749
Ensemble	**Ensemble classifiers (majority voting)**	0.66	0.648	0.676	0.662	0.662	0.802	0.818	0.786	0.802	0.802	0.812	0.754	0.815	0.783	0.791
Meta-level	**Two-Class Support Vector Machine**	0.766	0.758	0.769	0.763	0.764	0.816	0.821	0.817	0.819	0.818	0.758	0.697	0.74	0.718	0.728
	Two-Class Bayes Point Machine	0.762	0.75	0.775	0.762	0.762	0.828	0.837	0.824	0.831	0.83	0.811	0.75	0.818	0.783	0.791
	Two-Class Logistic Regression	0.791	0.784	0.794	0.789	0.79	0.843	0.848	0.845	0.846	0.846	0.79	0.733	0.782	0.757	0.766
	Two-Class Decision Forest	0.771	0.766	0.771	0.768	0.769	0.913	0.951	0.873	0.911	0.912	0.834	0.755	0.89	0.817	0.824
	Average	0.773	0.765	0.777	0.771	0.771	0.85	0.864	0.84	0.852	0.852	0.798	0.734	0.808	0.769	0.777
Combining	**Ensemble classifiers (majority voting)**	0.783	0.777	0.784	0.78	0.781	0.873	0.889	0.858	0.873	0.873	0.832	0.788	0.818	0.803	0.81

4.5 Summary

In this chapter, we outlined a series of experiments conducted on the sentiment classification of social media text using ensemble learning methods. Each base learner in the ensemble used meta-level features. Those features covered a combination of several existing lexicon and method resources for sentiment analysis. Feature hashing was also used in the representation of tweets. The experiments investigated three datasets to verify the effectiveness of the approach across different data. Our experiment results showed that these ensemble classifiers can minimize the error rate by avoiding poor selection of stand-alone classifiers, which is an effective way to ensure stability. In addition, using meta-level features mitigates problems associated with the sparsity of data. In that context, the meta-level ensemble approach can achieve promising results.

In the following chapter, meta-level features are investigated using long-form medical socially generated datasets. These features are used in order to enhance performance in multi-class sentiment analysis.

The work reported in this chapter was published in Alnashwan et al. [77].

Chapter 5

Long-form Sentiment Analysis

5.1 Introduction

In this chapter, we identify sentiments or affects expressed in online medical forums that discuss Lyme Disease. There are two goals in our research: first, to identify a complete and relevant set of categories that can be used to characterize Lyme Disease discourse; and second, to test and investigate strategies, both individually and collectively, for automating the classification of medical forum posts into those categories. We present a feature-based model that consists of three different feature sets: content-free, content-specific and meta-level features. Employing inductive learning algorithms to build a feature-based classification model, we assess the feasibility and accuracy of the automated classification proposed.

We first outline the motivation in Section 5.2. Section 5.3 discusses related work. In Section 5.4, we describe the details of our model of automated sentiment classification recognition. We present the experimental evaluation of the proposed model in Section 5.5. Then we illustrate the experimental result in Section 5.6 and present our conclusions in Section 5.7.

5.2 Motivation

In recent years, there has been a growth in the number of online health communities (e.g., MedHelp and Dailystrength) that include forums dedicated to general health issues (e.g., diet and fitness) or specific diseases (e.g., Lyme Disease). Within these online communities, members can discuss health-related

issues with others across large geographical distances. Members of a community participate in order to receive information and support through discussing aspects of their health-related issues, to ask questions about symptoms, treatments and side effects, and to share their health experiences [6].

This rapid increase in online health communities has seen a rise in the volume, velocity, and variety of data contributed to the forums concerned. For example, the MedHelp[16] community currently hosts around 11 million health discussions created by millions of users. These posts are submitted by users with varying levels of education and knowledge and from different social backgrounds, leading to a considerable range of submissions: long and short posts, information-providing and information-seeking, general and more specialized content, and structured and unstructured text [6]. This presents a significant technical challenge when attempting to analyse online medical discussions. Analysing such data can contribute to learning more about public knowledge or opinions on health-related matters [6][42], or to understanding and extracting information about the dynamics of a social network, such as by recognizing the dominant health issues in a community. A recent survey reported that about 72% of web users search for health information, including issues related to a specific disease, treatment or general health information [102].

With the spread and richness of online media in the last decade, the need for text classification has increased, which has led to this becoming an important application and research topic. There are very well-established techniques and algorithms for performing classification. However, it is hard to choose one universally efficient and effective technique that is applicable to diverse real-time data.

5.2.1 Why Lyme Disease?

Lyme Disease (LD), and the forums concerned with it, is a major focus of this study. It is a disease that is topical, has varied symptoms (which are not solely confined to this disease), can be complicated to diagnose and is the subject of controversy as to how it might be treated. Much of the difficulty is caused by its multi-faceted appearance and non-specific symptoms, which can be misdiagnosed as autoimmune diseases; due to this difficulty, Lyme Disease is known as "the great imitator". Members interact in online LD forums to gain the kind of social support that is important for people that may be suffering from the

[16] https://www.medhelp.org/about

condition. Often, members post messages to support others, to clarify confusion regarding an infection, or simply to share their experiences, so that sufferers (or potential sufferers) can better deal with the disease. In short, posts can come from a diverse population, can address diverse issues and can have diverse aims. We believe that sentiment classification of LD posts can help to understand the dynamics of such networks and can extract order out of the confusion created by the diversity of posts; in effect, we believe that we can produce an elementary mapping of a complex information space, one which users can employ to assist navigation within that space. We propose a multiclass classification model to help individuals browse and search for specific information that meets their needs, by categorizing diverse information into different classes. Classifying a large dataset provides a clearer view of where individuals should look and where they are currently positioned in the information space, allowing them to focus on particular aspects of the set. This information can also be useful for health professionals and policy makers. We return to these applications later.

5.3 Previous Studies

Various approaches, techniques and methods have been applied across different tasks to address the sentiment analysis problem. According to Wang et al. [28], the two popular methods for sentiment analysis are: natural language processing (NLP) and machine learning (ML) approaches. An NLP task involves converting text into a set of feature tokens that characterize the content of the text but are more amenable to computation. For ML techniques, the emphasis has been on supervised learning methods. An important research question is how to integrate different feature sets in supervised classification schemes. Feature sets can be categorized into content-free features and content-specific features. Content-free features include lexical and structural features [40]. Content-specific features are composed of words that have significance for a specific domain, e.g., the words 'treatment' and 'rash' have specific meanings within the context of Lyme Disease. Studies have shown higher text-classification performance after including content-specific features [40][38].

One of the key concepts in text classification is detecting semantic information represented in text; towards this end, a number of different techniques have been developed for lexicon adoption. Sentiment lexicon resources, which are referred to as meta-level features by Bravo-Marquez et al. [21], have been widely used as features in supervised classification schemes [42][21][40]. For

example, Bravo-Marquez et al. [21] combined 10 existing sentiment analysis methods and resources as a feature set in a supervised classifier that focuses on different aspects, such as polarity, strength and emotion. Sentiment analysis was performed using three different ML algorithms. The results suggest that a lexicon-based approach is the most suitable for polarity classification, whereas part-of-speech features are more appropriate for subjectivity classification. Most of the previous studies were conducted using these types of features individually. However, work done by Zheng et al. [37] and Abbasi and Chen [38] used a combination of the feature sets in an attempt at authorship identification of online messages. In contrast, our work proposes an ensemble feature set that includes content-free, content-specific, and meta-level features for sentiment or affect analysis. In addition, we adopt and evaluate the effectiveness of adding an existing domain lexicon as a new feature. As first demonstrated by Bobicev et al. [42], we show that general lexicons are not fully representative of health-related text data; however, it is no small task to generate a specific lexicon for each health domain.

Here we mention three recent papers with work overlapping ours. Zhang et al. [48] performed hierarchal classification with two cascading SVM algorithms using pattern and word features to identify user intents in online medical forums. Lu [47] successfully classified online posts for communities of breast cancer survivors and activists. His approach used three general categories: treatment, emotional support and survivorship; and three classification methods: C4.5, Naive Bayes and SVM with n-gram-based features, domain-specific features and sentiment features. In contrast to Zhang et al. [48] and Lu [47], we construct a feature-based model that consists of 10 different lexicon resources and lexical features and a key extraction and n-gram feature that depends on the most relevant n-gram as determined by the weighting scheme. Furthermore, we analyse sentiments that appeared in user posts in online health communities and generate a set of sentiment labels that are more sophisticated and more comprehensive for health-related discussions. The third related work is by Bobicev et al. [42], who performed sentiment analysis of online forums related to reproductive technologies based on one domain-dependent lexicon (HealthAffect), which the authors developed to aid their classification task. Their results show that the HealthAffect lexicon outperformed other existing lexicons in a multiclass classification task conducted on forums from an in vitro fertilization (IVF) website dedicated to reproductive technologies. For this reason, it was our intention to investigate use of this lexicon in our model, with the caveat that it

is health-focused but not specific to a particular disease.

5.4 Design and Implementation of Automated Sentiment Classification Recognition

A characteristic of online medical forums discussion, including those for Lyme Disease, is the presence of affect or sentiment. Hypothesizing a binary sentiment classification (positive or negative polarity) would be too limited and could not adequately address the range of sentiments expressed by a participant due to the complex nature of disease-related posts. Consequently, our intention was firstly to build and develop a set of sentiment categories for posts from disease-related forums and, secondly, to perform automated sentiment classification recognition with respect to these categories that would be meaningful, feasible and reliable.

The approach to implementation consists of two main stages: (1) domain-dependent categories identification, and (2) feature engineering.

5.4.1 Domain-Dependent Categories Identification

5.4.1.1 Dataset

As with other medical issues, Lyme Disease is discussed on several different forums, which themselves can be disparate in content. We collected anonymised data from several user forums, reflecting a broad cross-section of community networks populated by past, existing and potential patients. The forums/websites facilitate the forming of vibrant online health communities, in which thousands of communications on various conditions take place daily.

5.4.1.2 Identifying Categories

Forum posts are usually long enough to convey sentiment [42]. This study adopted a bottom-up approach to develop and identify sentiment categories that are specific to Lyme Disease forum discussion, the intention being to render these categories more suitable for a multi-class classification task. The sentiment analysis was carried out in two stages. First, we identified a set of categories that could capture the full range of sentiments or affect expressed in the discussions. Second, we investigated and evaluated whether those categories were sufficient for all or most of the posts in Lyme Disease medical forum discourse.

As shown in Figure 5.1 below, the process began with the selection of random posts crawled from Lyme Disease medical forums. These posts vary in length, the minimum word count being eight and the maximum 730, with an average of 165 words. In consultation with a number of medical professionals who have interest in Lyme Disease, each post was manually read and then categorized based on their observation. This facilitated the identification of a set of seed categories that emerged from the data. The seed categories started to be repeated after post number 67. Consequently, it was decided to terminate this step after manually classifying 93 posts, as no new categories were required. The seed categories were then compared both for similarities and differences in order for them to be summarized and potentially merged. This initially resulted in 22 categories. Some of the posts could represent more than one sentiment when considering this as a sentence- or paragraph-level sentiment classification task. However, in this study, we intended to place each of the original posts into one meaningful seed category, which led us to a document-level classification task of identifying the dominant categories for each post. Some of the posts were uncomplicated and easy to code. For example: *"Hi everyone! Please help me Lyme test result. This is my sons test. Is he has borrelia or not? **** Thanks!"* This post is very clear (if ungrammatical), so was categorized under the seed category "Confused about having Lyme (if they have Lyme or not)".

After development of the seed categories, the categories were compared for similarities in order for them to be abstracted to a higher categorization level, which we refer to as core categories. This process concluded after identifying six core categories. Table 5.1 presents the six core categories associated with the 22 seed categories; for clarity, an example of each core category is given.

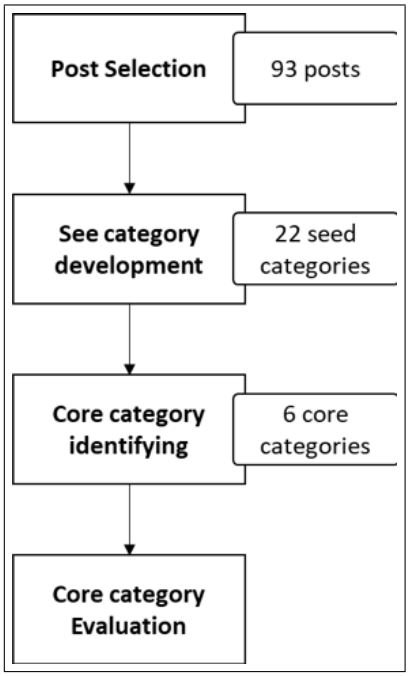


Figure 5.1: Process for identifying sentiment categories

5.4.1.3 Validating Category Selection

We constructed annotated posts using a crowdsourcing service: Mechanical Turk (MTurk), Amazon's online web service. MTurk is crowdsourced and enables individuals and companies (known as requesters) to take advantage of the power of a number of human workers (MTurkers) in completing a set of Human Intelligence Tasks (HITs). Tasks should be generated carefully to obtain annotations of high quality. Thus, a requester is able to access human intelligence in a simple, scalable and cost-effective way by generating tasks (HITs) which are distributed globally to thousands of high-quality and on-demand MTurkers for completion. In our case, the posts were distinguished according to their length so that the reward per assignment was fair. A flow diagram for this process can be seen in Figure 5.2.

We formulated the classification task using a short introduction about Lyme Disease, including a description of the dataset as well as guidelines for the MTurkers as to how to perform the task. We provided the text in Table 5.1 for further guidance.

Although most of the posts were long enough to represent several sentiment categories, as explained earlier, we requested that the MTurkers select the single dominant sentiment for each post. For category evaluation purposes, we added

Table 5.1: Description of categories and their subcategories

Categories	Seed categories
Asking about treatment	Asking about a specific treatment
	Asking about medication
	Medication is not helping and asking for an alternative
Lyme infection confusion	Being confused about having Lyme Disease (if they have Lyme or not)
	Is it worthwhile pursuing a particular doctor?
Lyme symptoms confusion	Confusion if a specific disease (not Lyme Disease) has these symptoms
	A patient is diagnosed with Lyme, but is confused if the symptoms relate to Lyme
	A patient who does not have Lyme, but is confused about the symptoms
Depressed and frustrating	Desperate and depressed
	Disappointed with the community
	Loneliness
	Worried and confused about having new symptoms
	Disagreement
Awareness and encouragement	Awareness and support
	Encouragement and support
	Providing general information
	Gratitude to his/her doctor
Seeking general information	Asking about advice
	Asking for information (from a doctor or specialist)
	Asking for information related to products (such as a Rife machine)
	Seeking test information
	Seeking a job
None of the above	If the post cannot be annotated with one of the above categories
	Kindly write your suggestion about a new class or category that can fit this post

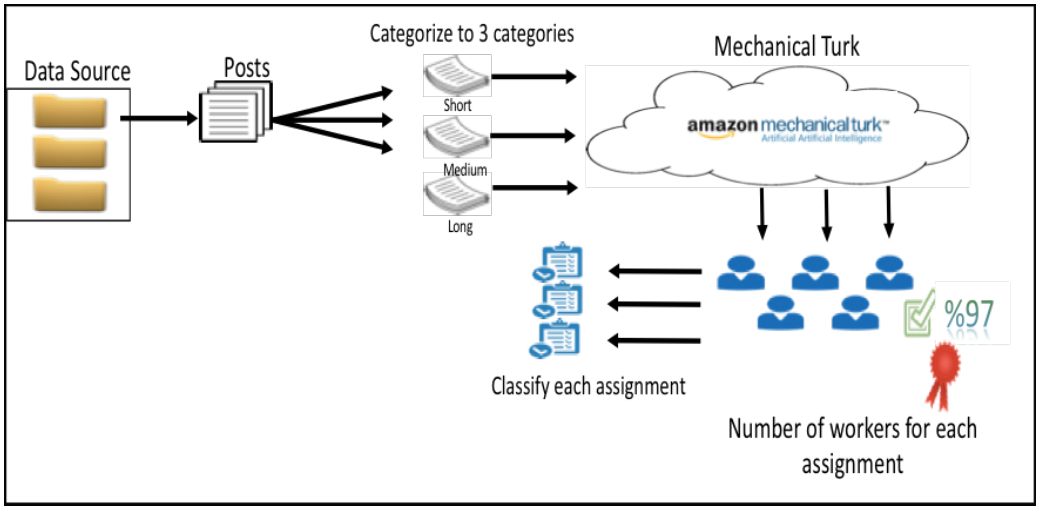


Figure 5.2: Flow diagram for annotation prosses

a final category, "None of the Above" and asked the MTurkers to suggest and justify a new category based on their viewpoint. By adding this step, we could evaluate if the proposed categories were sufficiently complete to match all or most Lyme Disease medical posts.

The aim of the evaluation of the 93 posts was to ensure that we received high-quality annotations from well-regarded MTurkers. The final category chosen, and its content, was determined by majority voting. Posts were omitted where the majority of the annotators were unsure of the most suitable category. The resulting annotation from the MTurkers was checked and compared with a domain expert annotation to verify if they matched. Figure 5.3 shows the distribution of the categories in the collected dataset, whereby every category has multiple assignments.

Of the 372 classification decisions, only 17 (4.6%) were "None of the Above". Ten of these decisions suggested "Test result confusion", three decisions suggested "Co-infections" and four did not offer any suggestion for a new category. We took this as confirmation that our proposed approach to identifying a comprehensive, yet distinct, set of sentiment categories was suitable.

5.4.2 Feature Engineering

The first step, as outlined above, was to identify a set of categories that could represent a comprehensive overview of sentiment or affect expressed in the discussions. We then needed to generate appropriate feature sets that would enhance the automated classification task. Feature choice is one of the most

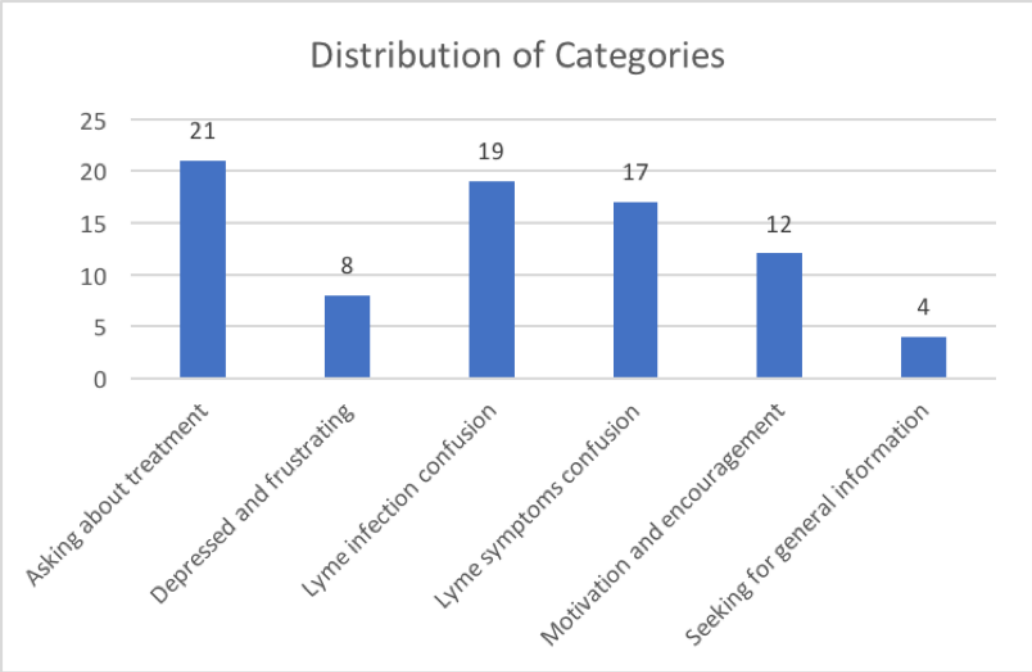


Figure 5.3: Statistics for the distribution of the categories

important and challenging aspects of supervised learning, as discussed in Section 3.1.1.1.

In a wider context, the aim was to identify a reliable multi-class classification model that could detect sentiments and affect across a large range of forum text. The approach should be capable of accommodating the diversity of language expressions in individual posts related to Lyme Disease. Our proposed approach was based on a feature set model, whereby high-dimensional text is reduced to a more manageable — and computable — set of features. In particular, we set out to develop an automated multi-class classification task based on three types of feature set: content-free features, meta-level features and content-specific features.

In the following sub sections, we first describe a baseline for the experiments, then we present the three different sets of features used to facilitate the classification and the rationale for using them.

5.4.2.1 The Baseline

Our baseline features simply embody feature hashing with n-grams. The feature-hashing model converts streams of words into vectors of integer components, by creating a hashing dictionary that consists of n-gram features calculated using the terms repeated in text. The advantages of using feature hashing is that it

reduces the dimensionality and leads to faster lookup of feature weights for the supervised learning machine, by representing text documents as equal-length numeric feature vectors. The hashing feature technique was developed using the Vowpal Wabbit library for 32-bit murmurhash v3 hashing[17] [78]. Feature hashing is set to a default setting with a bitsize of 10 in hashing each unigram (i.e., a single word/phrase).

5.4.2.2 Content-Free Features

Content-free features, that include lexical features, can be further divided into a word base and a character base. In our research, we included the character base and word base used in Zheng et al. [37]. In total, we adopted 59 lexical features for each post, collectively labelled F1. Table 5.2 includes the extracted features and their description.

5.4.2.3 Meta-Level Features

Meta-level features are based on several existing sentiment analysis lexicons and resources, which rely on a range of approaches to extract the sentiments from a piece of text: supervised, unsupervised and concept-based approaches. The outcome of these approaches is set in three dimensions: polarity, strength and emotion. From each lexicon resource, we generated a feature by calculating the number of words that matched between a post and the lexicon resource.

The results of adding these values are represented as dimensions in the feature vectors. The meta-level features include polarity lexicons (POL), emotion lexicons (EMO) and a domain-specific lexicon (DOM). Table 5.3 shows the type and number of features extracted and the range of values for each.

5.4.2.3.1 Polarity Lexicon Features

According to Bravo-Marques et al. [21], a lexicon-based approach is the most appropriate for polarity detection. We implemented this in our research, as outlined in Chapter 4, for analysing tweets for binary classification. However, simple polarity and strength detection is not sufficiently rich for analysing long-form text, as found in online Lyme Disease forums; we needed something more sophisticated. Applying meta-level features, as in the previous chapter, we combined seven existing sentiment analysis lexicons and resources in a feature

[17] https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-hashing#how-to-configure-feature-hashing

Table 5.2: Adopted lexical feature in our model

	Feature ID	Lexical Feature	Name of feature	Description
Character-based features	1	Total number of characters (C)	NumOfChar	-
	2	Total number of alphabetic characters/C	alphabetic_ characters/C	-
	3	Total number of upper-case characters/C	upper-case_ characters/C	-
	4	Total number of digit characters/C	digit characters/C	-
	5	Total number of white-space characters/C	white-space characters/C	-
	6	Total number of new lines	new lines	-
	7-32	Frequency of letters (26 features)	alphabet	A-Z
	33-54	Frequency of special characters (22 features)	specialchar	∼, @, #, $, %, ^, &, *, -, _, =,+, >, <, [,], {, }, \, |, /
Word-based features	55	Total number of words (M)	NumOfWord	-
	56	Total number of short words (less than four characters)/M	ShortWord	e.g., for, eat
	57	Average word length	AverageWord	-
	58	Average sentence length in terms of word	AverageSentence lengthWord	-
	59	Average sentence length in terms of character	AverageSentence lengthCharacter	-

set, then added emotion lexicons and a domain-dependent lexicon. This offers advantages over using a single lexicon resource.

The seven sentiment analysis lexicons and resources were: SentiWordNet, Bing Liu Lexicon, AFINN, NRC-hashtag, Sentiment140 lexicon, Sentiment140 method and SentiStrength; these are described in the previous chapter (see Sections 4.2.1). The additional two feature sets, emotion lexicons and a domain-dependent lexicon, are described below.

5.4.2.3.2 Emotion Lexicon Features

SenticNet: "SenticNet 2 is built by means of sentic computing, a new paradigm

that exploits both AI and Semantic web techniques to better recognize, interpret, and process natural language opinions" [103]. SenticNet defines the semantics and sentiments of over 14,000 common sense knowledge concepts. It is generated by employing an ensemble of graph mining and dimensionality-reduction techniques. SenticNet outcomes are a sentiment value related to each of the concepts found in a text. Each concept is correlated to one of five categories: Pleasantness, Attention, Sensitivity, Aptitude, and Polarity. We extracted the five emotion-oriented features by finding the words that matched between the SenticNet lexicon and the posts.

DepecheMood: [104] DepecheMood is an emotion lexicon with high coverage that includes approximately 37,500 terms. It was developed using an automated method with crowdsourced news articles from rappler.com. The word emotion matrix was generated by an affective annotation implicitly provided by readers. The emotions are: Afraid, Amused, Angry, Annoyed, Don't Care, Happy, Inspired, and Sad. We extracted these eight features as we did with the features from SenticNet.

5.4.2.3.3 Domain-Specific Lexicons

HealthAffect: The HealthAffect lexicon was created from a collection of about 700 posts to in vitro fertilization (IVF) forums and is thus medical domain-dependent [42]. It was created using a pointwise mutual information measure between each extracted n-gram and each proposed class. Detecting sentiments and analysing the data resulted in five classes: Encouragement, Gratitude, Confusion, Endorsement, and Factual; this was achieved by calculating the semantic orientation (SO) of each n-gram and for each class. After indicating all potential SOs, each n-gram is classified to the class that represents its maximum SO [42].

We followed Bobicev et al. [42] in representing a feature vector from the HealthAffect lexicon. We extracted emotion-oriented features, which are the sum of the score for each of the lexicon emotions for all the terms found in the posts. We then classified posts in the category for which the maximum number of terms was found. To overcome imbalance bias in the number of terms in the lexicon dimensions, we normalized the number of terms in the post by the total number of terms for each category. Finally, we represented each category with an integer value.

Table 5.3: Description of the extracted lexicon feature

Field	Lexicon source	Extracted features	Description	Range of values
Polarity	Bing Liu	Positive	Count the number of positive words that match the lexicon	{0,..,n}
		Negative	Count the number of negative words that match the lexicon	{0,..,n}
	Sentiment140 method	Polarity	Method polarity label	{0,2,4}
	SentiStrength	Polarity	Method polarity label	{-1,1}
Strength	SentiWordNet	Positive	Total of positive word values that match the lexicon	{0,..,∞}
		Negative	Total of negative word values that match the lexicon	{0,..,∞}
	AFINN	Positive	Total of positive word values that match the lexicon	{0,..,n}
		Negative	Total of negative word values that match the lexicon	{-n,..,0}
	NRC Hashtag	Positive	Total of positive word values that match the lexicon	{0,..,∞}
		Negative	Total of negative word values that match the lexicon	{-∞,..,0}
	Sentiment140 lexicon	Positive	Total of positive word values that match the lexicon	{0,..,∞}
		Negative	Total of negative word values that match the lexicon	{-∞,..,0}
Emotion	SenticNet	5 features (pleasantness, attention, sensitivity, aptitude, and polarity)	Total of word values that match each of emotions in the lexicon	{-∞,..,∞}
	DepecheMood	8 features (afraid, amused, angry, annoyed, do not care, happy, inspired, and sad)	Total of word values that match each of emotions in the lexicon	{0,..,∞}
	HealthAffect	(encouragement, gratitude, confusion, endorsement, and factual)	Sum of the score for each of the emotions for all terms found in posts. Then classified posts for which the maximal number of terms was found.	{1, 2, 3, 4, 5}

5.4.2.4 Content-Specific Features

According to Zheng et al. [37], content-specific features are important distinguishing factors for online messages. In this study, a content-specific feature can constitute a key phrase extraction feature (F5) or an n-gram extraction text representation feature (F6).

5.4.2.4.1 Key Phrase Feature

The purpose of using the automated key phrase extraction feature is to extract important topical words and phrases from a text. This feature has the potential for improving many NLP techniques by identifying the terms that best describe the subject, so that sentiments can be captured from phrases that have combinations of modifiers and nouns.

5.4.2.4.2 N-gram Feature

We set out to transform long text data into n-gram features, which could then be extracted and used as a surrogate for text; to do this, we transformed frequently occurring text segments into an equivalent numeric vector representation. This feature generates a content-specific dictionary from all n-grams in the training text dataset. The dictionary is the result of measuring the term frequency/inverse document frequency (TF-IDF) score for each n-gram [105]. We extracted unigrams and bigrams from the text and, due to the high dimensionality of such a text feature, we excluded any n-gram that appeared in fewer than five posts or more than 80% of the posts. To select the most highly correlated n-grams in the dictionary, we used Chi-squared feature selection to choose 2,000 desired features. This vocabulary and weighting were subsequently used when extracting the features from an unknown (test) dataset.

5.4.3 Classification Approach

Our proposed approach relies on different feature sets (see Figure 5.4). The features extracted can be categorized into two types: domain independent (DI) and domain dependent (DD). According to Biyani et al. [69], DI feature sets represent features that are used to extract expressed sentiments on a social media text in general, whereas DD feature sets are specific to a particular community. Of the feature sets, domain-specific lexicon features (DOM), key phrase features (KEY), and N-gram features (NG) are all extracted from Lyme Disease posts, and thus are specific to the Lyme Disease community and can be categorized

as DD features. The remaining sets are DI features, as they contain extracted lexical features (LEX), polarity and strength features (POL) and emotion lexicon features (EMO) from text that express sentiment and affect on online social media in general (see Table 5.4).

In order to evaluate the effectiveness of the features generated by our dataset, we started from a baseline and added features individually. We chose this incremental method in order to assess the impact of each feature, both individually and collectively.

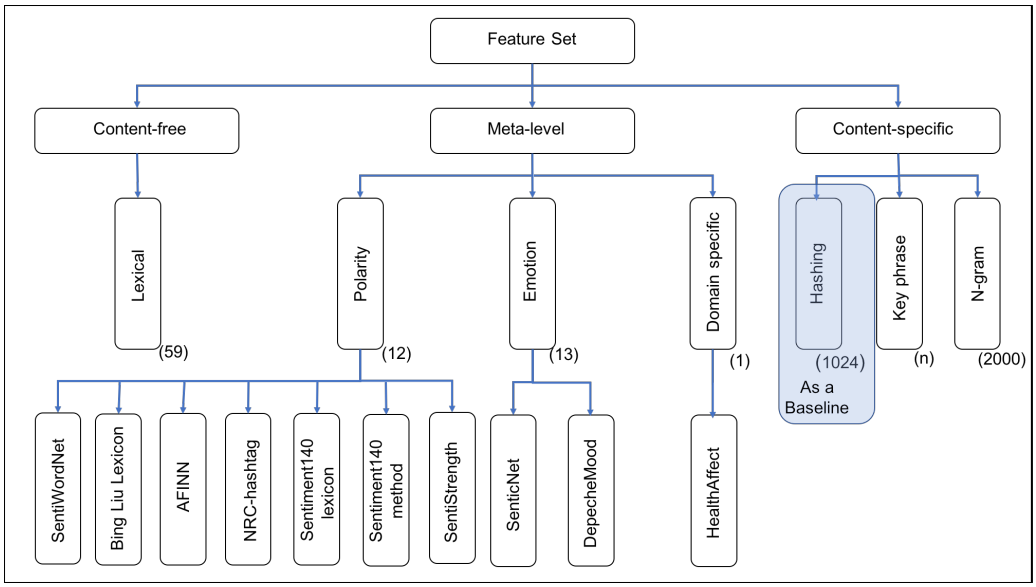


Figure 5.4: Feature set

5.4.3.1 Automatic Feature Selection/Reduction

As this approach could potentially result in a large number of features, it was beneficial to have strategies that could limit the number and/or determine the best features to apply; this was achieved using a technique called feature selection [106]. Feature selection is a well-known technique in machine learning that precedes training of the learner classifiers. The aim of this technique is to reduce the dimensionality, which can lead to enhanced performance and improved speed by discarding the features that are irrelevant. Research shows that classification can be improved by applying feature selection [107]. One of the best-known feature selection methods for estimating the importance of individual features with respect to the class label is the Chi-squared (χ^2) statistic. The χ^2 statistic measures how close the expected values are to the actual results (see equation 5.1):

Table 5.4: Categorizing different feature sets used for multiclass classification

Categorization of the extracted feature sets			
Content-free features	Lexical features (LEX)	Include a word base and a character base	Domain-independent feature sets
Meta-level features	Polarity lexicon features (POL)	A combination of seven existing lexicon resources to extract text polarity or strength	
	Emotion lexicon features (EMO)	Two existing lexicon resources that concentrate on extracting emotion indicated from text	
	Domain-specific lexicon features (DOM)	A lexicon resource generated from health-related posts that classifies text according to five different classes	Domain-dependent feature sets
Content-specific features	Key phrase features (KEY)	Important topical word and phrase extraction from a text	
	N-gram features (NG)	Dictionary generated by extracting unigrams and bigrams	

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5.1}$$

where O_i are the observed values, and E_i are the expected values under the assumption that the variables are drawn from a convenience sample of an independent variable. In this study, we used the training set and selected the 500 features with the highest χ^2 values.

5.5 Experimental Evaluation

5.5.1 Dataset

In order to construct an experimental dataset, we utilized a web crawler to collect approximately 3,000 posts across Lyme-related forums. Then, we constructed annotated posts using the services of Amazon's Mechanical Turk. Each post generated an annotation from five different high-quality MTurkers (with a guarantee that each annotator had a > 97% approval rate and had been granted a "Mechanical Turk Master" qualification). The flow diagram for the annotation process can be seen in Figure 5.5. As the posts were usually long enough to represent several sentiment categories – the posts had 161 words on average

– we asked the MTurkers to choose the most dominant sentiment category for each post. After getting an initial five annotations, each post was assigned to its final category in accordance with a simple majority. The posts were omitted when two or more categories that were assigned equal votes by the annotators (9.5% omitted posts). Figure 5.6 shows the vote distribution of the data (after omitting the posts labeled with an uncertain sentiment category or when "None of the Above" category had been selected).

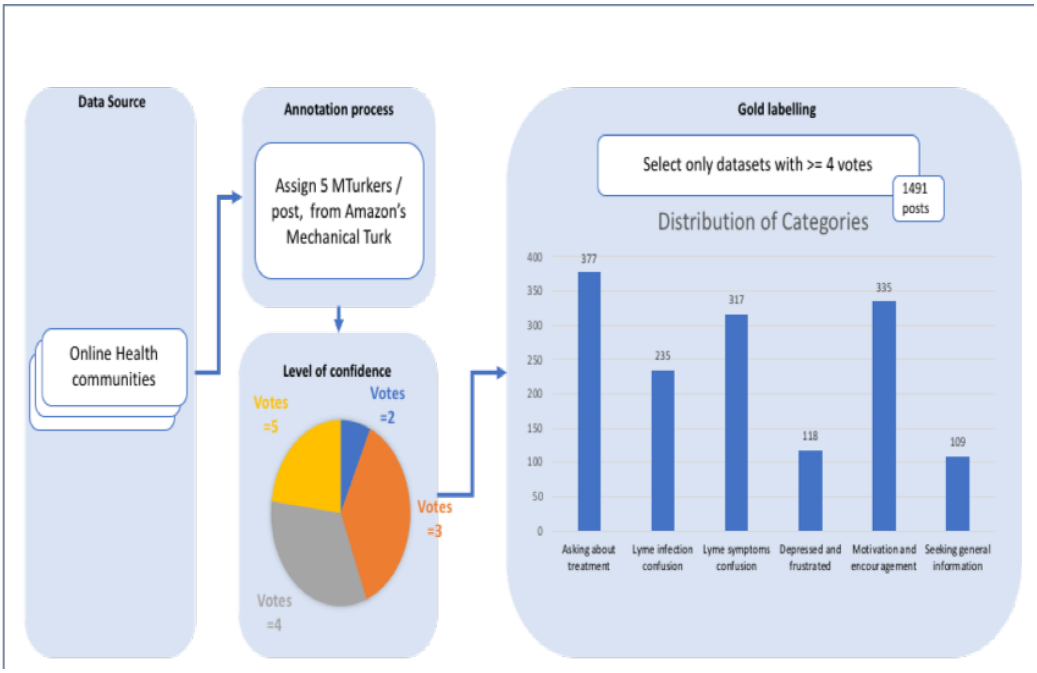


Figure 5.5: Flow diagram of Lyme data collected for multi-class sentiment classification

The quality of the annotation was assessed using *Fleiss' kappa* [108]. When manually labelling multi-class classification tasks, the measure evaluates the degree of inter-rater agreement between multiple annotators. Fleiss' kappa is defined in equation 5.2:

$$\textit{Fleiss' kappa} = \frac{(P - P_e)}{(1 - P_e)} \tag{5.2}$$

where the denominator (i.e., $(1 - P_e)$) is the degree of agreement that is obtainable above chance, and the numerator (i.e., $(P - P_e)$) gives the degree of agreement actually achieved above chance [108]. A Fleiss' kappa of 1 indicates complete agreement between raters. Landis and Koch [4] proposed an interpretation of Fleiss' kappa values that can be seen in Table 5.5. In spite of the diverse data, we acquired moderate agreement between the annotations, with

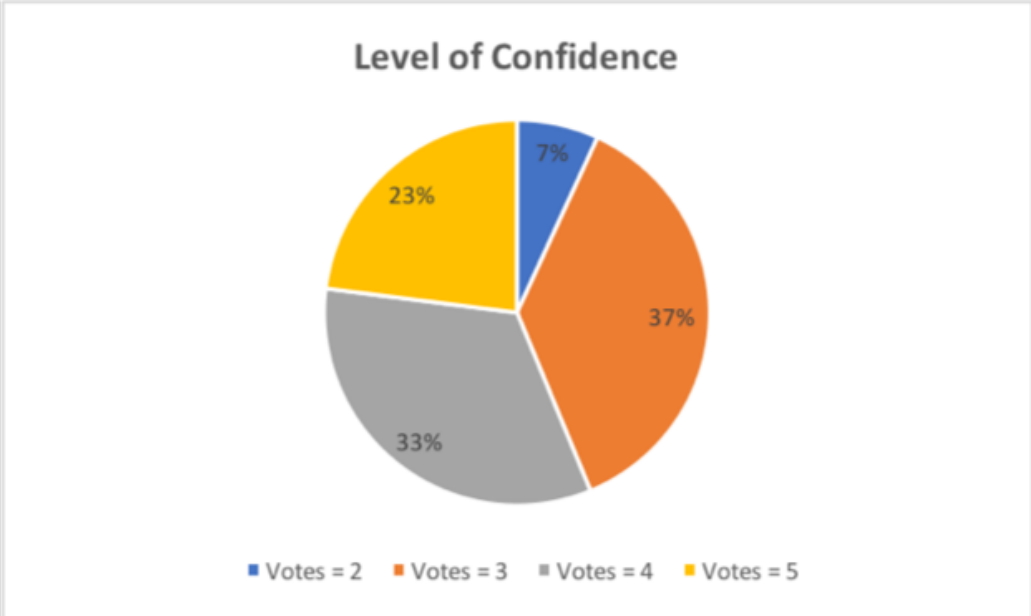


Figure 5.6: Vote distribution

Fleiss' kappa = 0.57.

Table 5.5: Fleiss' Kappa interpretation [4]

	Interpretation
<0	Poor agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

5.5.2 Gold Labelling

For our experiment, a gold standard subset was extracted from the larger collection in order to achieve a high-quality annotated dataset with substantial agreement. The data selected was that assigned to the same class by at least four annotators. The dataset selected resulted from 1,491 posts. As per our assumption, the gold standard dataset produced a higher Fleiss' kappa of 0.76, which indicates substantial agreement between the annotators (see Table 5.5). Figure 5.5 shows the final distribution of the gold standard labels among the sentiment classes and Figures 5.7 and 5.8 present examples of posts related to

Lyme Disease along with their associated class.

5.5.3 Data Pre-processing

We required pre-processing techniques to address the informality and lessen the noise confronted in the social media text. Noisy data can include HTML tags, hyperlinks, advertising in different sources, and any content or text that has no effect on the sentiment orientation but can mislead the performance of the classifiers. At the word level, for example, each word represents one dimension – so failing to remove irrelevant words can cause an increase in the dimensionality of the problem and lead to a more difficult and computationally complex classification problem. There are several data pre-processing techniques that can clean and simplify text to make it easier to identify features. We applied only the most basic and common operations in text processing that have been used previously and shown to be beneficial for sentiment analysis [29][3][47][90]:

- **Stop-words elimination:** this elimination was applied to some common words that have less potential of affecting class prediction (e.g., prepositions). It is performed using a predefined list of the most common stop-words. The stop-word list used was composed of about 312 words (e.g., “a”, “about”, “above”).
- **Removing certain classes of characters:** such as digits, special characters, email addresses and URLs.
- **Reducing word lengthening:** this includes normalizing word-lengthening. For example, when a word includes a sequence of repeated characters, such as “saaaaad”, it is reduced to a sequence of two characters: “saad”.
- **Normalizing character case:** all capital letters are converted to lower case to unify the data format.
- **Lemmatization:** this is applied to convert multiple related words to a single canonical form. Azure Machine Learning utilizes natural language processing libraries that involve multiple linguistic operations to implement lemmatization. First, sentence separation is done using natural language tools that depend on lexical analysis. Then comes the tokenization step, which determines the boundaries of words. The tokenization rules are established by a text analysis library devised by Microsoft Research. As it is hard to identify the part of speech for each word in any sequence of

words, Azure Machine Learning uses a disambiguation model to decide the single most likely part of speech according to the input context. Finally, the dictionary form is generated. A single word can have more than one dictionary form or lemma. For example, the word “increase” can be a verb or a noun, but only the single most likely dictionary form is created in Azure Machine Learning.

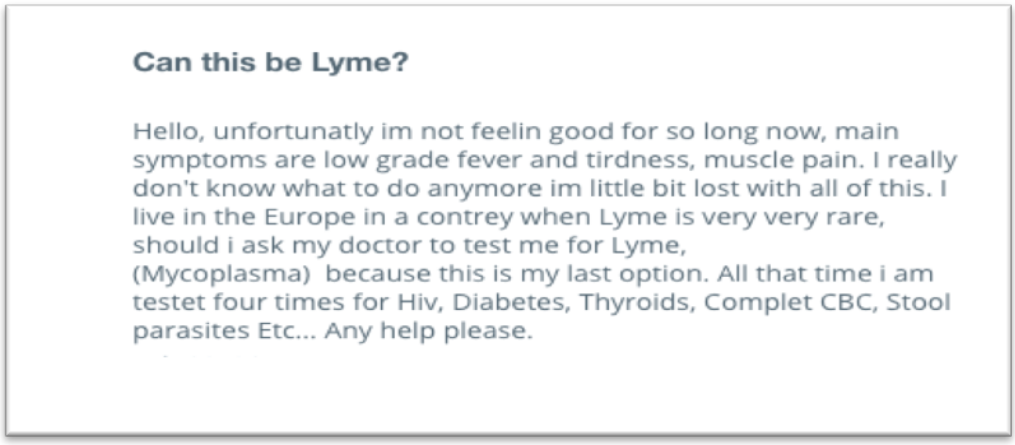


Figure 5.7: Example of a post related to Lyme infection confusion

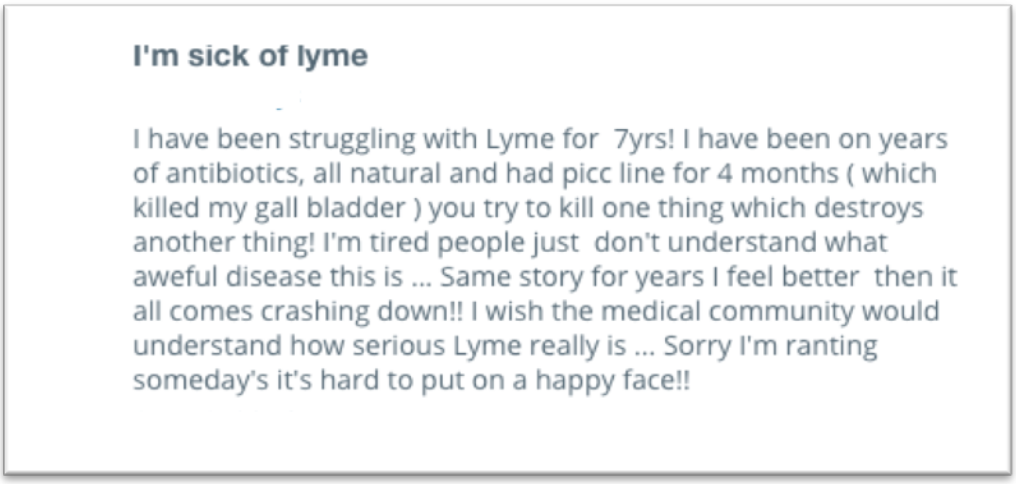


Figure 5.8: Example of a post related to the Depressed and frustrated class related to Lyme Disease

5.5.4 Machine Learning Techniques

As before, we used the Microsoft Azure Machine Learning Studio (MAMLS) platform. We considered various supervised learning algorithms to conduct our classification experiment (Multiclass Decision Forest, Multiclass Logistic Regression, Multiclass Neural Network, etc.) with all the parameters set to their

default values in Azure Machine Learning. Early experiments suggested that Multiclass Logistic Regression and Multiclass Neural Network produced a higher result than Multiclass Decision Forest, as shown in Table 5.6. Consequently, we discarded the use of Multiclass Decision Forest and employ only Multiclass Logistic Regression and Multiclass Neural Network in the experiments.

Table 5.6: Performance in terms of accuracy for three supervised learning algorithms

Supervised learning algorithms	**Baseline**	**Best model**
Multiclass Decision Forest	0.503	0.517
Multiclass Logistic Regression	0.607	0.738
Multiclass Neural Network	0.55	0.752

We used an 80/20 data split, which is one of the splits commonly used as a rule of thumb with randomized and stratified splitting: 80% of the data sample was used to train the model and the remaining 20% was used to test the performance evaluation.

5.5.5 Evaluation Metrics

To assess performance, we utilized a matrix for overall accuracy, precision and recall. We computed precision and recall matrixes using two different approaches: micro-averaging and macro-averaging. Micro-averaging tends to be effective with the most frequent classes, while macro-averaging considers each class equally [105].

We defined the evaluation binary matrix B (tp_i, fp_i, tn_i, fn_i), where tp_i is the number of true positives, fp_i is the number of false positives, tn_i is the number of true negatives, and fn_i is the number of false negatives, for binary evaluation of the i^{th} class label, and N as the number of classes. The overall accuracy is computed using formula 5.3. The micro-average is computed using **??** and 5.5 and the macro-average is computed using 5.6 and 5.7, for both precision and recall.

$$\textit{Overall accuracy} = \frac{\textit{(total number of correctly predicted)}}{\textit{(total number of prediction made)}} \tag{5.3}$$

$$\textit{Micro-average precision} = \frac{\sum_{(i=1)}^{N} tp_i}{\sum_{(i=1)}^{N} tp_i + fp_i} \tag{5.4}$$

$$\textit{Micro-average recall} = \frac{\sum_{(i=1)}^{N} tp_i}{\sum_{(i=1)}^{N} tp_i + fn_i} \tag{5.5}$$

$$\textit{Macro-average precision} = \frac{\sum_{(i=1)}^{N} \frac{tp_i}{tp_i + fp_i}}{N} \tag{5.6}$$

$$\textit{Macro-average recall } = \frac{\sum_{(i=1)}^{N} \frac{tp_i}{tp_i + fn_i}}{N} \tag{5.7}$$

5.6 Experiment Results

The analysis was carried out in stepwise increments, whereby we progressively added features and compared the results with the baseline features. The order in which we added feature categories reflected our intention of incremental development. It progresses from a simple approach (DI) to a more sophisticated one (DD). For instance, LEX features are unlikely to perform a classification task effectively on their own. Therefore, we intended to evaluate the feature set by adding new features that may be able to supplement the existing ones. Moreover, previous works generally begin their classification model with DI, this acting as a foundation for a DD feature set in classification tasks, such as [37] [40].

Table 5.7 documents the performance of the proposed classification model for different types of feature and technique. As can be seen from Table 5.7, the overall accuracies were very high, particularly when considering the challenging nature of the task. In most cases, the value of all five measures increased as the various features were added. The feature set (LEX+POL+EMO) outperformed both the feature set (LEX+POL) and LEX, each of which in turn outperformed the baseline in most cases. There was generally an increase in accuracy after adding more features, e.g., in extending from (LEX+POL+EMO) to (LEX+POL+EMO+DOM+KEY). This was especially evident after adding a domain-dependent feature.

From the results, it can be summarized that the whole feature set (LEX+POL+EMO+DOM+KEY+NG) gave the highest classification accuracy thus far. Our classification model achieved an accuracy of 73% and 75% for the two classifiers by applying feature selection and reducing dimensionality to the most relevant features. However, we did observe some anomalies: in most of the performance measures, features DOM and KEY performed well individually; but combining

them could reduce performance. We believe there is interaction between these two features. To assess the impact of DOM (HealthAffect lexicon) on the classification model, we built a model using all the feature space but withholding DOM. The model achieved higher accuracy, of 74% and 75%, indicating that this feature is not beneficial for what we are doing. This might well be due to the type of data used to build the HealthAffect lexicon, which is specifically related to IVF treatment and has no correlation with the disease that is our particular domain. It would suggest that, to be of practical use, domain lexicons must be very specific to the task at hand.

We compared the performance of the two machine learning algorithms and their reliability for the represented data. Table 5.7 reveals that Neural Network outperformed Logistic Regression. The greatest accuracy was achieved when using Neural Network and selected features from (LEX+POL+EMO+KEY) and NG. The Multiclass Neural Network scored higher by $< 3\%$ when all the feature sets were combined. The greater performance of the approach indicates the considerable contribution of the richness of feature space represented in the multi-class model. Feature selection also assisted in boosting performance. The results show consistency for overall accuracy and for both micro- and macro-averaging. The combination of all features, excluding DOM, had the best score for the two supervised learning algorithms.

Table 5.7: Experimental result for different feature set on Lyme Disease

			Confidence $\geq$ 4 (Baseline)	**LEX**	**LEX + POL**	**LEX + POL + EMO**	**LEX + POL +EMO +DOM**	**LEX + POL +EMO +DOM + KEY**	**LEX + POL +EMO +DOM + KEY +NG**	**(LEX + POL +EMO + KEY) selected 500 +NG**	**(LEX + POL +EMO +DOM + KEY) selected 500 +NG**
Multiclass Logistic Regression	**Train, Test (80/20)**	Overall Accuracy	0.607	0.638	0.638	0.638	0.651	0.648	0.715	**0.738**	0.732
		Micro-average precision	0.607	0.638	0.638	0.638	0.651	0.648	0.715	**0.738**	0.732
		Macro-average precision	0.628	0.627	0.682	0.676	0.686	0.681	0.705	**0.752**	0.74
		Micro-average recall	0.607	0.638	0.638	0.638	0.651	0.648	0.715	**0.738**	0.732
		Macro-average recall	0.539	0.559	0.567	0.567	0.582	0.575	0.66	**0.683**	0.678
Multiclass Neural Network	**Train, Test (80/20)**	Overall Accuracy	0.55	0.597	0.624	0.631	0.624	0.668	0.721	**0.752**	0.745
		Micro-average precision	0.55	0.597	0.624	0.631	0.624	0.668	0.721	**0.752**	0.745
		Macro-average precision	0.516	0.554	0.593	0.594	0.594	0.645	0.7	**0.738**	0.73
		Micro-average recall	0.55	0.597	0.624	0.631	0.624	0.668	0.721	**0.752**	0.745
		Macro-average recall	0.516	0.556	0.587	0.595	0.588	0.626	0.671	**0.711**	0.705

5.7 Summary

In this chapter, we proposed a model for the sentiment recognition of user posts in online health communities. Analysing sentiment in medical discussion is quite different from analysing polarity. While polarity can be appropriate in many cases – such as political blogs and consumer reviews of a product – it is not adequate for medical forums where, even within a single post or thread, there can be several competing themes. We identified comprehensive and appropriate categories that are adequate for applying to user posts in medical forum discussion. Our experiments investigated posts related to Lyme Disease within three different medical forums. We formulated our medical sentiment analysis as a multi-class classification problem to classify each user-generated post into one of the following classes: Depressed and frustrated, Lyme infection confusion, Lyme symptom confusion, Asking about treatment, Awareness and encouragement, and Seeking general information.

Our medical sentiment analysis was conducted using an incremental feature-based multi-class classification model. Content-free features, content-specific features and meta-level features with feature selection showed varying capabilities for sentiment recognition in users' online posts. Our experiment results show that multiple features used with a feature selection technique can maximize the performance of the classifiers.

The research described in this chapter presents several practical benefits that can have both direct and indirect impacts on health communities, as well as on health organizations and practitioners, due to the potential for enhancing health outcomes. In essence, our proposed model helps interested parties to understand where they are currently positioned within the information space. This would, for example, help health organizations to direct investment to those areas that would most benefit the community, thereby optimizing investment in generating awareness in a cost-effective manner. Moreover, this could help to identify misinformation regarding health issues and disseminate pertinent health information to target communities.

Overall, our experiments showed the proposed approach is highly effective and a promising start for future investigation. This approach can be adapted to a variety of applications in today's information-driven healthcare industry. For instance, this model may be adaptable to other disease-focused forums, as knowledge about social posts about one type of disease may be transferrable to

another type of disease. The model can also be aggregated with professional data from experts, such as doctors, to enhance the patient-centric delivery of healthcare. We believe that this model can also be helpful in improving the healthcare response to a specific disease within a pre-identified community and can support the creation of a new digital epidemiology [109] that would provide timely information about a disease and the health dynamics within a certain community.

In the following chapter, we evaluate our model further by assessing its ability to adapt to an online medical forum discussing a different disease.

The material in this chapter was published in Alnashwan et al. [100] and as part of [101].

Chapter 6

Application to an Alternative Disease Dataset

6.1 Introduction

In this chapter, we present further evaluation of the feature-based model (discussed in the previous chapter) by assessing its ability to adapt to data from another online medical forum (i.e., discussing Lupus). We identify sentiment or affect expressed in online medical forums that discuss Lupus. Specifically, we investigate the effectiveness of combining the proposed feature set to extract useful information for classifying texts accurately into generated domain-dependent classes that derive directly from this domain. While this works for Lyme Disease discourse, we postulate that it is not unique: we believe that the research can also be applied to other health and medical topics, such as Lupus.

There are two goals in this chapter (following similar goals of the previous chapter): first, to evaluate the set of categories identified that can be used to characterize Lupus; and second, to test and investigate strategies, both individually and collectively, for automating the classification of medical forum posts into those categories. Again, we investigate a feature-based model involving content-free, content-specific and meta-level features. Employing inductive learning algorithms to build a feature-based classification model, we assess the feasibility and accuracy of the proposed automated classification technique.

6.1.1 Why Analyse Lupus Disease Posts?

The main motivation for using Lupus is that it shares several of the characteristics of Lyme Disease, in that it gives rise to a certain amount of confusion and can be misdiagnosed due to the symptoms presenting differently between patients.

The outline of the chapter is as follows. Section 6.2 describes the details of the adaptation of the model. We present the experimental evaluation of the proposed model in Section 6.3. We then illustrate the experimental results in Section 6.4 and present our conclusions in Section 6.5.

6.2 Design and Implementation of Automated Sentiment Classification Recognition

The approach implemented consists of two main stages: (1) the identification of domain-dependent categories within another disease-related discourse and their evaluation; and (2) adaptation of the feature engineering.

6.2.1 Domain-Dependent Categories Identification

6.2.1.1 Dataset

As with Lyme Disease, Lupus is discussed on a number of forums. In order to identify appropriate categories, we collected random posts related to Lupus disease from across several forums; we subsequently used crowdsourcing to validate these categories.

6.2.1.2 Identifying Categories

After reading each post and consulting with people who are working in the field, seed categories were generated and assigned to each relevant post. After examining a number of randomly selected posts, the seed categories started to be repeated and the process was drawn to a close after post number 81. The seed categories were combined according to their similarities and differences, resulting in the creation of core categories. As a result, six core categories emerged from the Lupus online community posts.

Perhaps surprisingly, the categories developed for Lyme Disease also proved to be an adequate fit for the Lupus dataset. The two datasets (i.e., Lyme Disease

and Lupus posts) shared the same core categories and were assigned to similar subcategories. Table 6.1 presents the six core categories associated with 25 seed categories related to posts about Lupus.

Table 6.1: Description of Lupus discussions categories and their subcategories

Category	Includes
Asking about treatment	Is a specific medication causing a specific symptom or side effect?
	Asking for experience of changing the medication doses
	Asking about if a medication helped?
	Asking about a good medication for Lupus
	Asking for a treatment that can deal with a particular pain
	Asking if the patient feels normal after a specific medication
Lupus infection confusion	Being confused about having Lupus disease (if they have Lupus or not)
	Being confused about the interpretation the result of the test and does it means diagnosed with Lupus.
Lupus symptoms confusion	A patient is diagnosed with Lupus, but is confused if the symptoms relate to Lupus
	A patient who does not have Lupus, but is confused about the symptoms
	Asking how long does symptoms will last
Depressed and frustrating	Desperate and/ or depressed and/ or scared and/ or nervous
	Disappointed with the community
	Loneliness
	Disagreement
Awareness and encouragement	Awareness and support
	Encouragement and support
	Gratitude to his/her doctor
Seeking general Information	Asking about advice
	Asking to understand Lupus
	Asking for information (good doctor or specialist or hospital)
	Can a patient do (something) if they have Lupus
	Asking about advice when you cannot see a doctor
	How patients with similar disease live their life?
	Seeking for information about tests (what to test on themselves
None of the above	If the post cannot be annotated with one of the above categories.

6.2.1.3 Evaluating Categories

We constructed the annotated posts (the annotation steps were as presented in Section 5.3.1) using Amazon's MTurk crowdsourcing service. The classification task was formulated by including a brief introduction about Lupus, a description of the nature of the datasets and guidelines for the MTurkers as to how to perform the task. Furthermore, Table 6.1 was included in the task as supplementary material.

As discussed previously, although there are some posts that can represent multiple sentiment categories, we asked the MTurkers to select the predominant category for each post. A "None of the above" category was added to investigate the adequacy of the proposed categories, and to gain insight into new categories, if there were any, suggested by the MTurkers.

In the evaluation process for the 81 posts relating to Lupus, five different high-quality MTurkers (i.e., well-regarded MTurkers who met similar conditions and qualifications to those described in Chapter 5) were assigned to classify each post. Figure 6.1 shows the distribution of the categories in the data collected after omitting posts with equal votes (as this led to difficulties in determining their associated categories). Note, however, that the categories are more imbalanced than the corresponding Lyme Disease Categories.

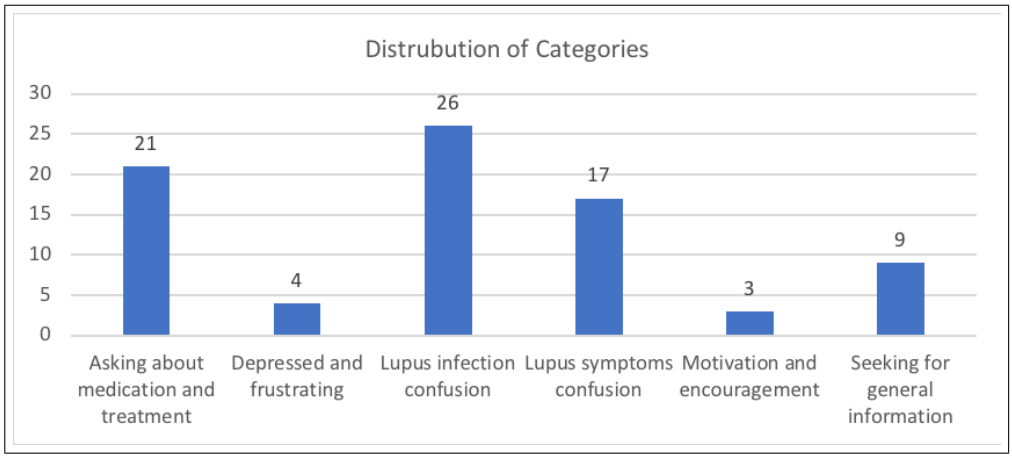


Figure 6.1: Statistics for the distribution of the categories related to Lupus posts

Of the 405 classification decisions, only 14 (3.5 %) were assignment to "None of the above". No MTurkers offered any suggestions for new categories. We inferred from this that the proposed categories (i.e., identical categories to those relating to Lyme Disease posts) present suitable and comprehensive sentiment groupings.

6.2.2 Feature Engineering

For the purpose of investigating whether the feature-based model – with content-free, content-specific and meta-level features – is adequate in this instance, we carried out the classification task on Lupus posts. To begin with, we established a baseline based on feature hashing using n-gram, following the same approach referred to in Chapter 5.

6.2.3 Classification Approach

As presented in the previous chapter, the features were examined in a stepwise manner, starting from the baseline and adding features individually in order to assess the impact of each feature, both individually and collectively. The experiment also applied the same technique as before for feature selection.

6.3 Experimental Evaluation

6.3.1 Dataset

To conduct our experiment, we considered posts dedicated to Lupus taken from different online health communities. We collected approximately 2,800 posts and followed the same annotation process we undertook previously. We then requested Master MTurkers to classify these. The vote distribution for the Lupus dataset for the collected dataset (i.e., 2,800 posts) is shown in Figure 6.2.

6.3.2 Gold Labelling

A gold-standard dataset was obtained from the annotation process and consisted of 1,085 posts. Figure 6.3 shows the final distribution of the gold-standard labels among the sentiment classes. Figure 6.4 presents an example of a post related to Lupus, with its associated class.

As can be observed in Figure 6.3, one of the classes has only about 4% of the overall data, which is quite small; this signifies the unbalanced distribution of the data. We did consider applying techniques to handle the unbalanced data, such as oversampling (using the Synthetic Minority Oversampling Technique, or SMOTE). However, this could lead to misinterpretation when comparing results across the two medical forum discourses.

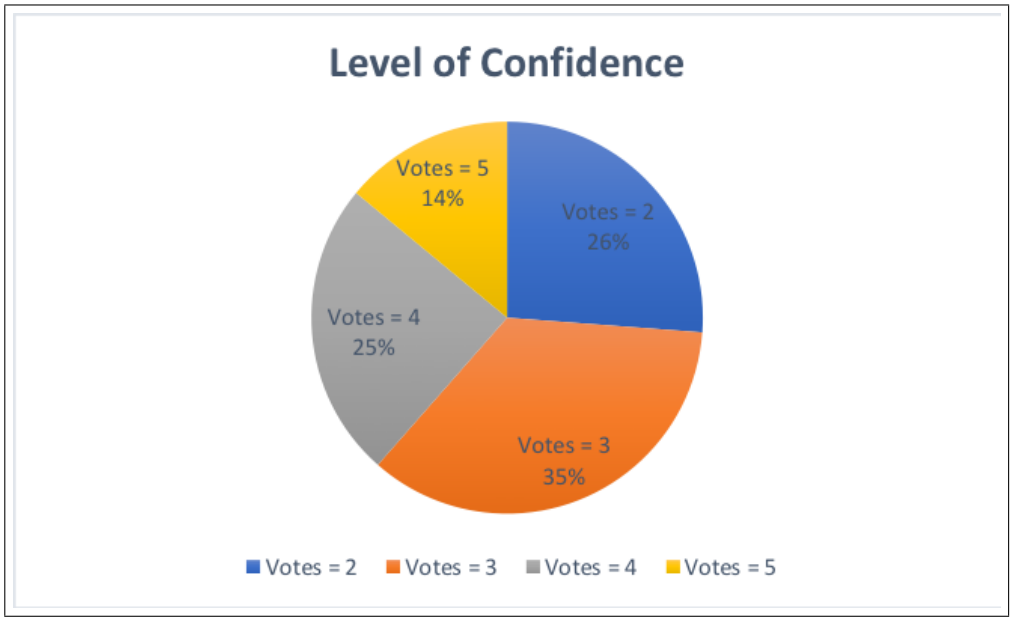


Figure 6.2: Vote distribution for Lupus posts data

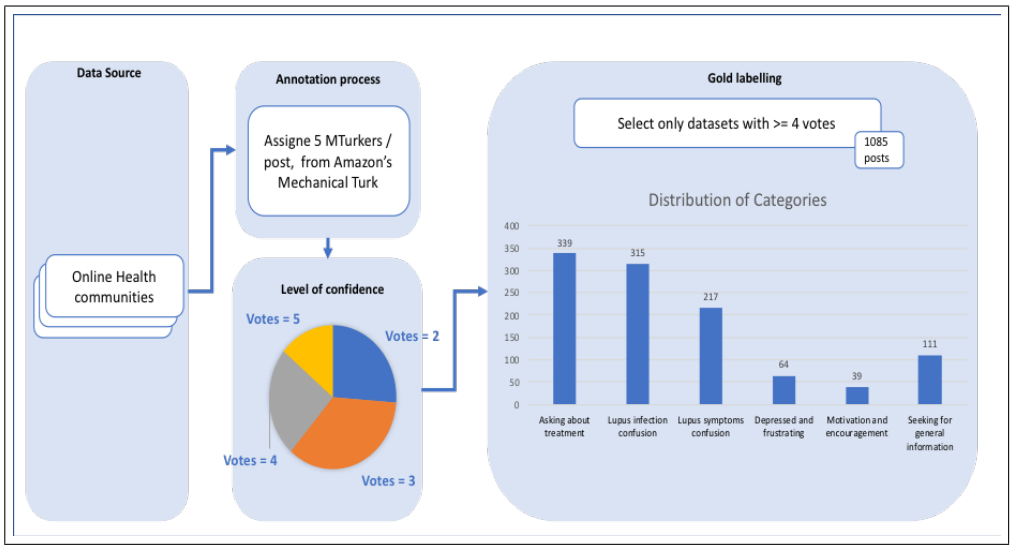


Figure 6.3: Flow diagram of Lupus data collected for multi-class sentiment classification

6.3.3 Machine Learning Techniques

We analysed the multi-class classification model using the same supervised learning algorithms and evaluation metrics as before in order to test the effectiveness of the same model. We considered two supervised learning algorithms, multi-class logistic regression and multiclass neural network, as these had produced the most reliable results in relation to the Lyme Disease medical forum discourse.

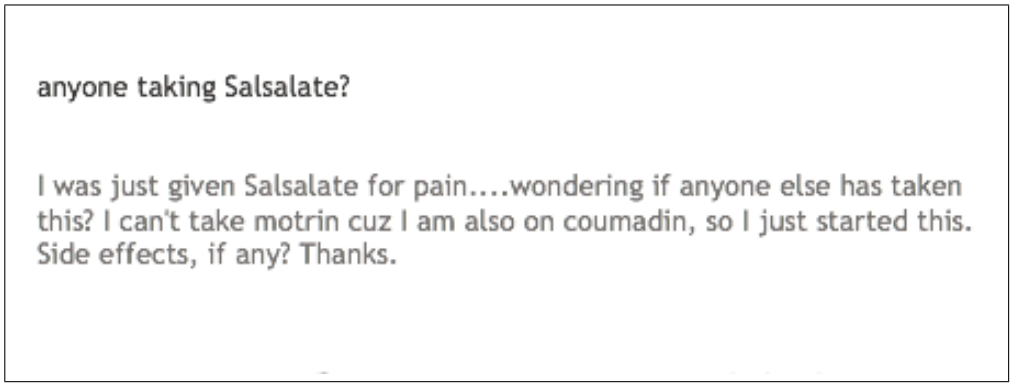


Figure 6.4: Example of a post related to Asking about treatment class related to Lupus disease

6.4 Experiment Results

The analysis was carried out in stepwise increments, whereby we added features and compared the results with the baseline features. Table 6.2 documents the results of applying the multi-class classification model to Lupus medical forum discourse. Generally, the evaluation measures continued to improve through the incremental addition of features, although adding F4 (HealthAffect) diminished performance. As with the Lyme Disease dataset, this confirms that F4 does not make an effective contribution to our model.

The performance also corresponds across the machine learning algorithms, whereby the multiclass neural network scored 74%, outperforming multiclass logistic regression by 2%, to achieve the best result. It can be observed that this result is consistent across two different diseases in the medical forum discourse. However, the Lyme Disease dataset scored higher than that of Lupus in the best performance measures. We suggest this is attributable to the amount of data used to train the learners, an imbalance in class distribution or a combination of both. The investigation does imply that our model can be applicable to different types of disease in medical forum discourse.

Table 6.2: Experimental result for different feature set on Lupus disease

			Confidence ≥ 4 (Baseline)	LEX	LEX + POL	LEX + POL + EMO	LEX + POL +EMO +DOM	LEX + POL +EMO +DOM + KEY	LEX + POL +EMO +DOM + KEY +NG	(LEX + POL +EMO + KEY) selected 500 +NG	(LEX + POL +EMO +DOM + KEY) selected 500 +NG
Multiclass Logistic Regression	**Train, Test (80/20)**	Overall Accuracy	0.636	0.641	0.673	0.668	0.668	0.664	0.691	**0.7**	**0.7**
		Micro-average precision	0.636	0.641	0.673	0.668	0.668	0.664	0.691	**0.7**	**0.7**
		Macro-average precision	NaN	NaN	0.633	0.63	0.63	0.628	0.71	**0.676**	**0.676**
		Micro-average recall	0.636	0.641	0.673	0.668	0.668	0.664	0.691	**0.7**	**0.7**
		Macro-average recall	0.394	0.396	0.484	0.482	0.482	0.479	0.519	**0.524**	**0.524**
Multiclass Neural Network	**Train, Test (80/20)**	Overall Accuracy	0.51	0.645	0.627	0.627	0.627	0.677	0.714	**0.724**	0.714
		Micro-average precision	0.51	0.645	0.627	0.627	0.627	0.677	0.714	**0.724**	0.714
		Macro-average precision	0.51	0.579	0.564	0.567	0.582	0.645	0.711	**0.725**	0.714
		Micro-average recall	0.613	0.645	0.627	0.627	0.627	0.677	0.714	**0.724**	0.714
		Macro-average recall	0.618	0.505	0.508	0.51	0.535	0.563	0.596	**0.602**	0.587

*NaN indicating missing value or cannot be handled due to predicting with no value for a certain class.

6.5 Summary

In this chapter, we presented the results of an investigation of a model for the recognition of sentiment in user posts in online health communities, this time using Lupus posts. The work described in the chapter was twofold. First, we examined and evaluated the proposed categories. Having identified and assigned appropriate forum posts, we concluded that the generated classes were reliable and a good fit for applying to user posts. Second, we investigated the medical sentiment multi-class classification problem. We classified each user-generated post into one of the following classes: Depressed and frustrated, Lupus infection confusion, Lupus symptom confusion, Asking about treatment, Awareness and encouragement, and Seeking general information (i.e., the same classes proposed in Chapter 5).

Our medical sentiment analysis was conducted using the proposed feature-based multi-class classification model that employed content-free, content-specific and meta-level features, together with feature selection for dimensionality reduction. Our experiment results showed that the proposed feature-based model has the ability to adapt to an online medical forum discussing Lupus. The experimental results demonstrate the effectiveness of our approach.

In the next chapter, we assess the performance of the supervised feature-based model in Chapters 5 and 6 using formal statistical evaluation analysis. We then discuss and summarize the findings of the empirical investigations.

The work described in this chapter was published in Alnashwan et al. [101].

Chapter 7

Evaluation and Discussion

7.1 Introduction

This chapter evaluates the supervised feature-based model of medical discussions related to Lyme Disease and to Lupus discussed in Chapters 5 and 6. Performance is assessed by adding features in a stepwise manner and using appropriate evaluation metrics and techniques. The chapter then critically evaluates the impact of combining the three types of feature set – Content-free, Meta-level and Content specific – using statistical techniques.

The outline of the chapter is as follows: we present experimental evaluation of the purposed models in Section 7.2. In Section 7.3, we investigated the significant effect of considering different group of feature sets. The chapter ends with a summary of the findings of our empirical investigations.

7.2 Performance Evaluation

To assess the success rate of a machine learning classifier, the dataset was split into training and testing data and the evaluation was carried out on a test set, which contains unseen instances that were not involved in building the classifier.

In Chapters 5 and 6, we evaluated performance using the 'hold-out' approach, whereby the dataset was split into two independent sets: training and testing data. The training dataset was used to train the classifiers and then we then classify instances of an unseen testing set to evaluate performance. Although this is a valid evaluation approach, it depends on only one set for training and

one set for testing – not all instances in the testing dataset are involved in the training process. It is more effective to use the whole dataset thoughout the evaluation process. However, due to the cost of obtaining labelled data, this may not always be viable. Instead, we used another evaluation approach: k-fold cross validation [110]. The k-fold cross-validation approach we employ randomly partitions a dataset into k-folds, each fold representing a dataset portion of equal size and maintaining approximately the same distribution of classes as the original dataset – a strategy known as stratified cross validation. For each k-fold, k$-$1 folds are used to train the classifier and evaluation is then carried out on the remaining fold. The training process is repeated k number of times on different partitions of the training dataset and each fold is used for testing in turn. The evaluation results are averaged from all the folds to score the overall performance. Cross validation can result in more robust and reliable performance compared with the hold-out approach, as it is trained on multiple training and testing splits [110]. In this chapter, we compare the performance of different multi-class classification models using 5-fold cross validation for each dataset. The value k=5 is a commonly used number of folds [16]; also this value has been considered to ensure there are sufficient assignments to every category in the evaluated dataset.

To assess performance, we utilize metrics for overall accuracy, and micro-averaging and macro-averaging precision and recall, as described in Chapter 5. We also use an additional metric for evaluating the classifiers: the F1 score. The F1 score is the harmonic mean of the precision and recall. We report the F1 score for micro-averaging and macro-averaging: Micro F1-score is used as a global metric, calculating the total true positive, false negative and false positive; whereas macro F1 score considers the average F1-score for each class.

Feature selection techniques are used to reduce the high dimensionality of the feature space in the classification model. As discussed in Chapter 5, we assessed the importance of features using the Chi-squared (χ^2) test. In addition, we evaluate the performance with another feature selection technique that is widely used, the Fisher score [111], which can deal with continuous data values without discretization and normalization of the data. The Fisher Score computes the score for each feature independently based on the distance and variance of each class.

The analysis was carried out in stepwise increments, as done in Chapter 5, whereby we added features and compared the results with the baseline features.

Table 7.1 documents the performance of the classification model for the Lyme Disease dataset for different types of feature and technique. We can observe that, in most cases, the value of all seven measures increased as the various features were added. For example, the feature set (LEX+POL) outperformed the LEX feature set, and both sets outperformed the baseline. The performance was stable, or showed a slight increase in accuracy in F1 scores, after extending from (LEX+POL+EMO) to (LEX+POL+EMO+DOM+KEY). Combining the whole feature set with feature selection gave the highest classification accuracy and F1 scores. This suggests that the two feature selection techniques improved the performance of the multi-class classification model. The χ^2 and Fisher scores showed similar performance within this dataset.

In a similar vein, we studied the performance of the feature-based model on posts related to Lupus. The performance after combining the proposed feature sets in a stepwise manner revealed consistency with the data relating to Lyme Disease. The results are shown in Table 7.2.

Figure 7.1 summarises the effect on accuracy of adding the features in a stepwise fashion, both for Neural Network (NN) and Logistic Regression (LR) learning algorithms. From that figure, we can observe that there is a decrease in accuracy after adding DOM to the feature set, and thus it was interesting to investigate performance when omitting DOM. Similar to the results in Chapters 5 and 6, the best outcome occurred when combining all the feature sets with χ^2 feature selection but omitting DOM. The performance increase slightly when omitting the DOM feature in the Lyme Disease dataset whereas, with Lupus, there was no change in the performance when adding or removing the DOM feature in the model. This can indicate ineffectiveness of using this feature.

We compared the performance of the two machine learning algorithms and assessed their reliability for the two datasets. As shown in Figure 7.1, the Neural Network and Logistic Regression classifiers performed similarly for the baseline and then the score increased consistently with the addition of features in a stepwise approach. The largest increase for both machine learning algorithms was when adding the F6 (n-gram text extraction) feature to the other features. The Neural Network outperformed Logistic Regression in both experiments.

Table 7.1: Feature-based model performance on a Lyme Disease dataset using cross validation

		Confidence ≥ 4 (Baseline)	**LEX**	**LEX + POL**	**LEX + POL + EMO**	**LEX + POL +EMO +DOM**	**LEX + POL +EMO +DOM + KEY**	**LEX + POL +EMO +DOM + KEY +NG**	**(LEX + POL +EMO +DOM + KEY) (Chi) selected 500 +NG**	**(LEX + POL +EMO +DOM + KEY)(Fisher) selected 500 +NG**	**(LEX + POL +EMO + KEY) (Chi) selected 500 +NG**	**(LEX + POL +EMO + KEY) (Fisher) selected 500 +NG**
Multi-class Logistic Regression	Overall accuracy	0.598	0.641	0.651	0.647	0.646	0.647	0.699	0.706	0.708	**0.711**	0.71
	Micro-average precision	0.598	0.641	0.651	0.647	0.646	0.647	0.699	0.706	0.708	**0.711**	0.71
	Macro-average precision	0.6	0.624	0.672	0.663	0.656	0.656	0.688	0.7	0.7	**0.706**	0.704
	Micro-average recall	0.598	0.641	0.651	0.647	0.646	0.647	0.699	0.706	0.708	**0.711**	0.71
	Macro-average recall	0.506	0.539	0.571	0.566	0.562	0.562	0.629	0.633	0.635	**0.641**	0.638
	F1 score - micro	0.598	0.641	0.651	0.647	0.646	0.647	0.699	0.706	0.708	**0.711**	0.71
	F1 score - macro	0.513	0.546	0.586	0.581	0.575	0.575	0.646	0.651	0.652	**0.658**	0.656
Multi-class Neural Network	Overall accuracy	0.596	0.638	0.643	0.649	0.646	0.665	0.709	0.72	0.715	**0.724**	0.72
	Micro-average precision	0.596	0.638	0.643	0.649	0.646	0.665	0.709	0.72	0.715	**0.724**	0.72
	Macro-average precision	0.571	0.619	0.619	0.619	0.629	0.644	0.684	0.702	0.694	**0.704**	0.695
	Micro-average recall	0.596	0.638	0.643	0.649	0.646	0.665	0.709	0.72	0.715	**0.724**	0.72
	Macro -average recall	0.55	0.584	0.587	0.592	0.591	0.598	0.646	0.667	0.662	**0.671**	0.669
	F1 score - micro	0.596	0.638	0.643	0.649	0.646	0.665	0.709	0.72	0.715	**0.724**	0.72
	F1 score - macro	0.558	0.596	0.596	0.601	0.646	0.609	0.66	0.68	0.674	**0.684**	0.68

Table 7.2: Feature-based model performance on a Lupus disease dataset using cross validation

		Confidence ≥ 4 (Baseline)	**LEX**	**LEX + POL**	**LEX + POL + EMO**	**LEX + POL +EMO +DOM**	**LEX + POL +EMO +DOM + KEY**	**LEX + POL +EMO +DOM + KEY +NG**	**(LEX + POL +EMO +DOM + KEY) (Chi) selected 500 +NG**	**(LEX + POL +EMO +DOM + KEY)(Fisher) selected 500 +NG**	**(LEX + POL +EMO + KEY) (Chi) selected 500 +NG**	**(LEX + POL +EMO + KEY) (Fisher) selected 500 +NG**
	Overall accuracy	0.652	0.663	0.676	0.672	0.676	0.676	0.723	0.724	**0.727**	0.724	**0.727**
Multi-class Logistic Regression	Micro-average precision	0.652	0.663	0.676	0.672	0.676	0.676	0.723	0.724	**0.727**	0.724	**0.727**
	Macro-average precision	0.656	0.669	0.66	0.645	0.667	0.669	0.736	0.733	**0.739**	0.729	**0.739**
	Micro-average recall	0.652	0.663	0.676	0.672	0.676	0.676	0.723	0.724	**0.727**	0.724	**0.727**
	Macro-average recall	0.449	0.46	0.488	0.485	0.494	0.498	0.585	0.588	**0.592**	0.589	**0.592**
	F1 score-micro	0.652	0.663	0.676	0.672	0.676	0.676	0.723	0.724	**0.727**	0.724	**0.727**
	F1 score-macro	0.461	0.475	0.515	0.511	0.523	0.528	0.627	0.629	**0.633**	0.629	**0.633**
	Overall accuracy	0.636	0.653	0.66	0.661	0.653	0.699	0.733	0.75	0.739	**0.751**	0.735
Multi-class Neural Network	Micro-average precision	0.636	0.653	0.66	0.661	0.653	0.699	0.733	0.75	0.739	**0.751**	0.735
	Macro-average precision	0.53	0.589	0.602	0.602	0.589	0.652	0.705	0.728	0.718	**0.728**	0.708
	Micro-average recall	0.636	0.653	0.66	0.661	0.653	0.699	0.733	0.75	0.739	**0.751**	0.735
	Macro-average recall	0.491	0.54	0.56	0.564	0.548	0.583	0.598	0.628	0.628	**0.633**	0.617
	F1 score-micro	0.636	0.653	0.66	0.661	0.653	0.699	0.733	0.75	0.739	**0.751**	0.735
	F1 score-macro	0.504	0.557	0.574	0.577	0.563	0.607	0.632	0.662	0.66	**0.666**	0.648

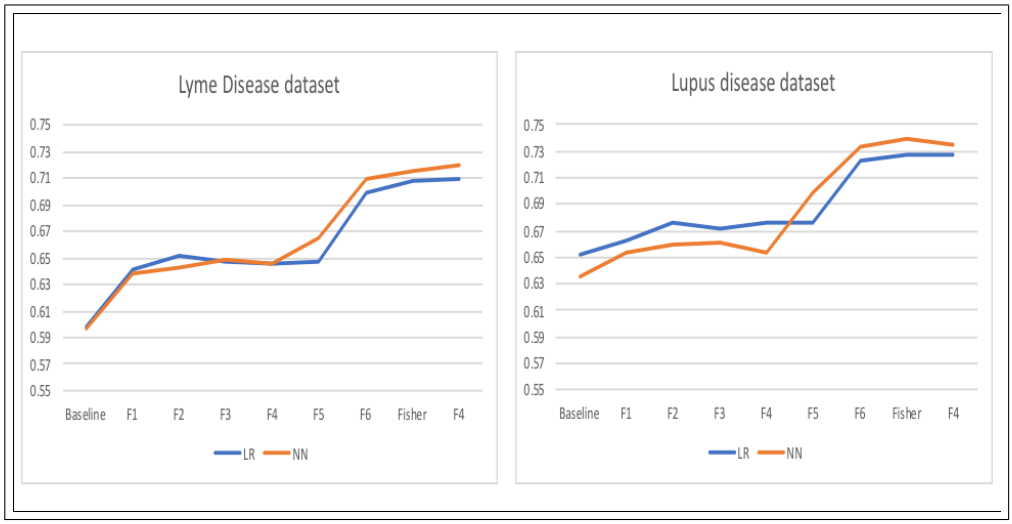


Figure 7.1: The learning curve for the feature-based model

7.3 Evaluation of Content-free, Meta-level and Content-specific Feature Sets

7.3.1 Statistical Tests

Statistical tests are needed for the comparison of performance between machine learning methods and to confirm that any differences revealed did not occur by chance [110]. In our case, we compared the performance of the trained model with different extracted feature sets using statistical tests, choosing the *corrected resampling paired t-test* to evaluate any statistically significant difference in performance. The comparison was carried out between the different groups of feature sets on the classification model with the cross-validation using the corrected resampling paired t-test with 5% significant level [112]. The null hypothesis is that there is no difference in the mean performance between the two classification models. This statistical test is preferred to the simple paired t-test as it corrects the variance estimator between samples. In this section, the test is conducted on accuracy and F1 scores.

We carried out the experiments by evaluating the performance of three different feature groups (Content-free, Meta-level and Content-specific) extracted incrementally using a stepwise approach, as categorized and discussed in Chapter 5. The statistical significance of the results for the two classifiers of the medical posts related to Lyme Disease are shown in Table 7.3. We observed a significant improvement when combining the Baseline with each of the Content-free,

Meta-level and Content-specific feature sets. Furthermore, combining all three feature sets led to a statistically significant improvement when compared to any combination of Baseline, Content-free and Meta-level – and to the Baseline alone.

Table 7.3: Accuracy of the classification model with three different types of feature sets on the Lyme Disease dataset. The best performance per column is given in bold. Underlining refers to statistically significant differences with respect to the Baseline ($p<0.05$); * refers to statistical significance with respect to the previous model in each column ($p<0.05$)

	Multi-class Logistic Regression			**Multi-class Neural Network**		
	Accuracy	**F-score -micro**	**F-score - macro**	**Accuracy**	**F-score - micro**	**F-score - macro**
Baseline	0.598	0.598	0.513	0.596	0.596	0.558
Baseline + Content-free	0.641*	0.641*	0.546*	0.638*	0.638*	0.596
Baseline + Content-free + Meta-level	0.646	0.646	0.575	0.646	0.646	0.646
Baseline + Content-free + Meta-level + Content-specific	**0.699***	**0.699***	**0.646***	**0.709***	**0.709***	**0.660***

We also investigated the statistical significance of the performance of different types of feature sets on medical posts related to Lupus, as can be seen in Table 7.4. The results indicate that for Logistic Regression, after combining the Meta-level feature with the Baseline and Content-free feature, the results improve significantly in comparison with the Baseline only and with the Baseline plus the Content-free feature; however, there is no significant difference between their performance using the Neural Network. Overall, combining all the feature sets improved the accuracy and F1 score significantly over the Baseline in all experiments.

Table 7.4: Accuracy of the classification model with three different types of feature set on the Lupus dataset. The best performance per column is given in bold. Underlining refers to statistically significant differences with respect to the Baseline ($p<0.05$); * refers to statistical significance with respect to the previous model in each column ($p<0.05$)

	Multi-class Logistic Regression			**Multi-class Neural Network**		
	Accuracy	**F-score -micro**	**F-score -macro**	**Accuracy**	**F-score - micro**	**F-score - macro**
Baseline	0.652	0.652	0.461	0.636	0.636	0.504
Baseline + Content-free	0.663	0.663	0.475	0.653	0.653	0.557
Baseline + Content-free + Meta-level	0.676*	0.676*	0.523*	0.653	0.653	0.563
Baseline + Content-free + Meta-level + Content-specific	**0.723**	**0.723**	**0.627***	**0.733***	**0.733***	**0.632**

7.3.2 Goodness of Fit

In classification problems, the class distribution can be unbalanced and may be skewed towards a majority class; this may well result in a good estimation of accuracy that has occurred by chance. The kappa statistic [113] addresses this issue by normalizing the classifier accuracies using the different frequencies of the class distribution. The kappa statistic is utilized to calculate the agreement between predicted and observed classes, while taking into account the possibility of the agreement occurring by chance [110]. The maximum value of kappa is 1 when a classifier performs perfectly and 0 indicates the classifier prediction was no better than what would be expected by a random classifier. The kappa statistic is calculated as shown in 7.1:

$$Kappa = \frac{Accuracy - p_{exp}}{1 - p_{exp}} \tag{7.1}$$

$$p_{exp} = \sum_{(i=1)}^{m} (\sum_{(j=1)}^{m} \frac{X_{ij}}{N} * \sum_{(j=1)}^{m} \frac{X_{ji}}{N}) \tag{7.2}$$

where p_{exp} is the probability of arriving at agreement by chance and calculated as in 7.2; m refer to the number of classes and N is the total number of instance in dataset to create the confusion matrix X in multi-class classification problem. The confusion matrix includes m*m cells, where X_{ij} cell indicates the total

number of instance predicted with class i and the actual class is *j*.

In this section, we report the results of the experiments discussed in Section 7.3.1 using the kappa statistic measure. In Tables 7.5 and 7.6, the kappa statistic increases as we combine the extracted feature types using a stepwise approach for both datasets. In the dataset related to Lyme Disease, performance improved significantly after utilizing the Content-free feature when compared with the Baseline. Whereas, after extracting the Meta-level feature set and adding to the model, performance as measured by kappa is significantly higher for the dataset related to Lupus disease. As a general observation from the results, the Neural Network classifier outperformed the Logistic Regression classifier in most cases. The final feature-based model in both datasets is the best performing and has a statistically significantly improvement over the Baseline and (Baseline + Content-free + Meta-level).

In further investigation, we applied another statistical significance test to verify whether our feature-based model is reliable and that there had been no overfit. We used target shuffling, "which is a process that reveals how likely it is for results to have occurred by chance" (John Elder, Founder of Elder Research)[18] . Target shuffling is performed by randomly shuffling the class labels, applying a classification task and then repeating the classification of the randomly shuffled class labels several times. Running target shuffling several times resulted in a mean accuracy of 0.228%, which indicates a significance level of < 0.0001. Therefore, we are confident that our model is reliable and that there is no overfit.

[18]www.elderresearch.com/target-shuffling.

Table 7.5: Kappa performance of the classification model with three different types of feature set on the Lyme Disease dataset. The best performance per column is given in bold. Underlining refers to statistically significant difference with respect to the Baseline ($p<0.05$); * refers to statistically significance with respect to the previous model in each column ($p<0.05$)

	Multi-class Logistic Regression	**Multi-class Neural Network**
	Kappa	**Kappa**
Baseline	0.488	0.492
Baseline +Content-free	0.544*	0.547*
Baseline + Content-free + Meta-level	0.552	0.557
Baseline + Content-free + Meta-level + Content-specific	**0.621***	**0.635***

Table 7.6: Kappa performance of the classification model with three different types of feature set on the Lupus dataset. The best performance per column is given in bold. Underlining refers to statistically significant difference with respect to the Baseline ($p<0.05$); * refers to statistically significance with respect to the previous model in each column ($p<0.05$)

	Multi-class Logistic Regression	**Multi-class Neural Network**
	Kappa	**Kappa**
Baseline	0.521	0.513
Baseline + Content-free	0.537	0.536
Baseline + Content-free + Meta-level	0.559*	0.541
Baseline + Content-free + Meta-level + Content-specific	**0.626***	**0.641***

7.4 Error Analysis

In this section, we conduct error analysis to investigate the underlying cause of the misclassification of instances in the feature-based model in both datasets: posts related to Lyme Disease and to lupus. For simplicity, we confine our analysis to the multi-class logistic regression classifier; recalculating for the neural network classifier is straightforward.

Figure 7.2 presents the confusion matrices for the Lyme Disease and lupus datasets and delivers a general overview of the error analysis. The diagonal cells show the percentage of correctly classified instances; the other cells are instances of misclassification and error.

From our observation, one of the main causes of error coming from the posts was mixed classes (posts covering more than one class). A post may be quite long, and can express concepts identified with more than a single class. For example, a participant can start a post by writing about the confusion of having the disease and then express how that causes a high level of depression and frustration. Furthermore, as can be seen from Figure 7.2, the infection confusion and symptom confusion classes were found to be overlapping in both datasets. [As those two classes are the most closely related, noticing onset of symptoms may indicate having an infection; a specific infection could present known symptoms but could also have many others also]. In addition, the highest error from both datasets came from "Seeking general information". Interestingly, a considerable number of instances related to this class can be misclassified with the "asking about treatment" class. One of the main reasons is that the post can include general questions about the disease and also expressing some confusion about a treatment or the best treatment for a given situation.

In addition to mixed classes, there are a number of potential causes of confusion in classifying posts. As a consequence of the nature of social media, abbreviations, acronyms and misspellings of words/terms are frequently used. Some of those terms are non-standard and are unique to the posts in which they appear, which leads to errors in classification. Another cause of misclassification is short posts that include little content, from which it would be difficult to identify any features generated.

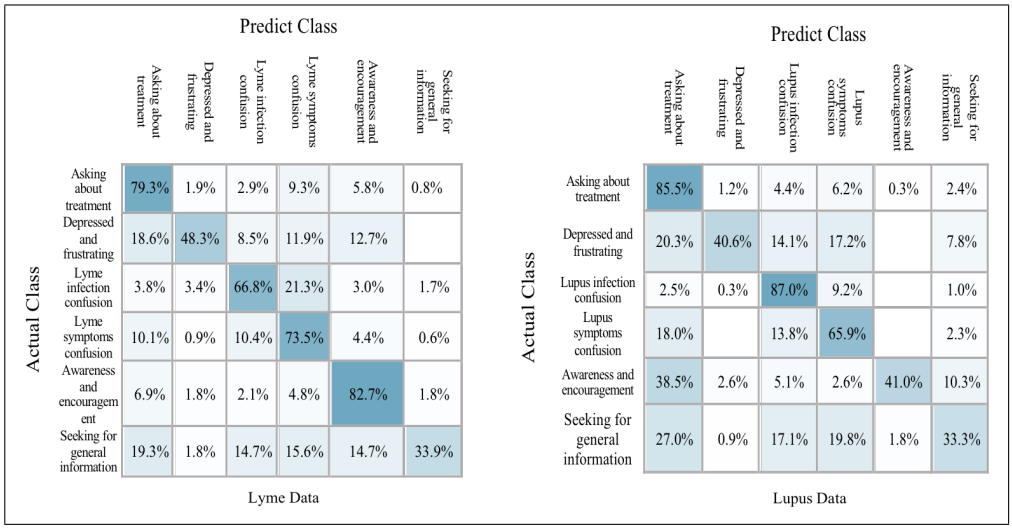


Figure 7.2: Confusion matrix for Lyme and Lupus data

7.5 Main Findings

In the experiments reported in this chapter, we investigated our supervised feature-based model using two datasets of medical online discourse and two different machine learning algorithms. Below we summarize the results and discuss the main findings from the evaluation analysis.

- Some features can overlap and interact due to the complexity of having various features. However, from the analysis of the results, we were able to isolate the effects of adding each feature type. For example, combining LEX and POL enhanced the performance of the multi-class classification model. In contrast, adding DOM reduced the performance in most cases; this was not a surprise, as intuitively, DOM represents the HealthAffect lexicon, which was designed for a different medical field than the diseases discussed in our data.
- The results also indicate that NG is the most beneficial when combined with others in all the experiment results. This finding is attributable to the appearance of the words and phrases in the compiled list of the most informative features that were expected to provide clues for the classifiers.
- We evaluated the performance of combining incrementally the three different groups of feature sets: Content-free, Meta-level and Content-specific. The results illustrate that combining all three groups of feature sets significantly boosted the multi-class classification performance and significantly outperformed any other feature set combinations.

- Applying machine learning, it is almost impossible to achieve 100% accuracy in performance due to the challenges of the sentiment analysis of socially generated data, which exhibit a high level of noise, as discussed in Chapter 3. We argue that the feature-based sentiment analysis model described herein achieves a good level of performance within the chosen medical domains.

7.6 Summary

In this chapter, we outlined a series of experiments we conducted with our feature-based classification model. The evaluation process was presented in two sections. In the first section, we evaluated the effect of adding each feature set to the others using cross-validation evaluation techniques and reported the results using different evaluation metrics. In the second section, we conducted an evaluation of the experiments conducted with three grouped feature sets: Content-free, Meta-level and Content-specific, to assess the performance of combining them in a stepwise manner using statistical significance evaluation analysis. The experimental results demonstrate the effectiveness of our approach.

In many sentiment analysis applications, the labelled training datasets are fairly expensive to obtain, whereas unlabelled datasets are readily available. In the next chapter, we address this issue through the adaptation of a well-known semi-supervised learning technique, in which *co-training* is implemented by combining labelled and unlabelled data.

Chapter 8

Semi-supervised Approach - Modified Co-training

8.1 Introduction

Most of the current sentiment and text classification approaches were developed using the supervised learning approach, which requires labelled training data. But labelled data is hard to come by – certainly outside of the standard datasets available to machine learning researchers. We believe that the limited availability of labelled data has been challenging and limiting for most of the researchers who wish to build and evaluate sentiment analysis and classification systems for applied domains. Furthermore, while the process of manually labelling data is expensive, tedious, time-consuming – and in some cases unfeasible – unlabelled user-generated data are often readily available and cost-free. Therefore, semi-supervised learning as a strategy for machine learning has drawn considerable attention, as it aims to leverage the benefits of using unlabelled data in order to enhance the performance of learning approaches with limited labelled data. One question arises, however: can we achieve a high degree of accuracy in the classification of medical discourse using unlabelled data? This chapter addresses this question.

In this chapter, we focus on the co-training learning method, one of the better-known approaches in semi-supervised learning, originally proposed by Blum and Mitchell [51]. We previously demonstrated the effectiveness of our feature-based model in extracting useful information for classifying texts into domain-dependent classes (i.e., domains relating to Lyme Disease and Lupus) in a fully

supervised learning setting (see Chapters 5 and 6). We now investigate the effectiveness of combining those features with co-training learning in automated sentiment classification. To the best of our knowledge, we are among the first to apply co-training to multi-class classification within a medical domain. We believe that the research described here can also be applied to other socially generated health and medical data, especially for diseases that share the same characteristics as those targeted in this research, and perhaps to similar types of online discourse.

The outline of the chapter is as follows. Section 8.2 presents some of the previous related work. In Section 8.3 we describe our method and experiment design. We present the experimental setup of the proposed model in Section 8.4. We then illustrate the experimental results in Section 8.5 and present our conclusions in Section 8.6.

8.2 Previous Studies

Most sentiment and opinion analysis has been based on supervised learning methods, performed on user-generated text in forums and blogs,[42][20] product reviews [115], and political discussion [31] (see Chapter 3). Semi-supervised learning is an alternative approach to machine learning, which learns from both labelled and unlabelled data in the training stage. Most existing semi-supervised learning approaches are derived from the following techniques: (1) expectation maximization (EM), (2) graph-based algorithms, and (3) co-training (as described in Chapter 3). However, according to Yu [23], EM in a semi-supervised setting is limited by the mixed model assumption, and graph-based semi-supervised learning is not ideal when processing large-scale data. Consequently, in this chapter, the focus is on co-training for a more generalizable model.

The co-training learning strategy was originally developed by Blum and Mitchell [51]. The model requires two redundant and independent views of a space, in which each view includes sufficient information to classify the target correctly. For example, Blum and Mitchell [51] investigated a co-training algorithm for a web classification problem. Each web page was represented by two distinct views: the content appearing on a specific web page and the content appearing in the hyperlinks that refer to that web page. Algorithms then define two separate classifiers: a page-based classifier and a hyperlink-based classifier, each to be

trained on a small set of labelled data. The co-training algorithm then learned from the enriched labelled data as generated iteratively by the two classifiers. It is asserted that co-training in a semi-supervised setting can reduce classification error by leveraging the use of separate classifiers over different views to combine their predictions.

Several extensions of the original co-training algorithm have been developed, specifically in natural language processing applications. Some of the extensions include the following:

- Based on view variations, Li et al. [116] proposed an effective two-view model based on personal/impersonal views in co-training algorithms. The two views were extracted using the following heuristic rules: "identify personal sentence" (when the first word in the sentence includes a personal pronoun) and "identify impersonal sentence" (when the first word in the sentence includes an impersonal pronoun). The authors found that applying the two proposed views in a co-training model was effective for binary sentiment classification in eight different domains.
- Biyani et al. [69] applied semi-supervised co-training algorithms to detect the polarity (positive/ negative) of user posts in online health communities. Domain-independent (DI) and domain-dependent (DD) features were extracted for each post as the two views in the co-training settings.

For much real-world data, a natural split is not always apparent; so the absence of natural feature splits that might generate two views deters some researchers from applying a co-training technique. Fortunately, studies have shown the effectiveness of co-training by developing separate learners, which can derive either from different feature splits or different learning algorithms. For example, Zhou and Goldman [117] successfully applied single-view co-training with two different classifiers.

The research in this chapter is motivated by the challenge of multi-class sentiment classification in medical and health-related discourse. The obstacles faced include machine learning when there is a scarcity of labelled data and the lack of previous work on co-training with multiple classes. In a wider context, this study aims to identify a reliable multi-class classification model that can identify sentiment and affect across a wide range of medical forum texts in a semi-supervised setting.

As explained above, co-training can be applied when a dataset has a natural

division of features. In our case, we do not have this division and need to adapt it. In contrast to the work by Biyani et al. [69], the current study aims to develop a co-training model based on three different feature sets: content-free, content-specific and meta-level features, which we categorize into two views – domain-independent and domain-dependent – to recognize sentiment automatically in a multi-class classification problem.

8.3 Methods and Experiment Design

In the design outlined below, we adapted and evaluated a semi-supervised model using the pre-processed gold-standard datasets previously derived from forums relating to Lyme Disease (see Chapter 5) and Lupus (see Chapter 6).

8.3.1 Co-training Model

As stated by Blum and Mitchell [51], the success of a co-training model relies on the following three assumptions: first, data are split into at least two different sets, called views; second, the views are compatible and, specifically, each view is sufficient for classification; and third, these views are conditionally independent for a given class. The authors argue that co-training in a semi-supervised setting can reduce classification error by leveraging the use of separate classifiers over different feature views and combining their predictions. Subsequent work by Balcan et al. [118] shows that the condition of independence can be relaxed, with little affect on performance.

To run a co-training model, classifiers, say C_1 and C_2, should be assigned, generating a small set of labelling data, $L = (X_1, Y_1), (X_2, Y_2), \ldots\ldots (X_i, Y_i)$ and a large set of unlabelled data, $U = X_1, X_2, X_3, \ldots\ldots X_n$. Instead of training C_1 and C_2 with the whole set of unlabelled U data, a small subset U' is more beneficial [51]. The rationale behind selecting a smaller pool, U', of unlabelled data is that it can indirectly force C_1 and C_2 to select the examples that are more representative of the underlying distribution that generated U [51].

A diagram of the co-training is shown in Figure 8.1 and the pseudocode is presented in Figure 8.2. The co-training model proceeds as follows:

- After predefining the small set of labelled data, L, unlabelled data, U, and a small pool of unlabelled data, U', for each iteration, the classifiers C_1 and C_2 are trained on L, on the assumption that the performance of the initial

classifiers will be low but better than random. Then each of the classifiers is allowed to label the unlabelled set, U'.

- Next, the examples with the highest confidence are selected from amongst the U' data generated from C_1 and C_2 and added to L with their assigned labels.
- The U' dataset is replenished with a predefined number of examples selected randomly from each class.
- The process will continue for K iterations, K being set to a value that is usually predefined. At the end of the co-training model, C_1 and C_2 are trained using the L dataset augmented with all the selected auto-labelled examples.

In the test stage, the final prediction is assigned by the product of using the two classifiers, C_1 and C_2, which are the output of the co-training model.

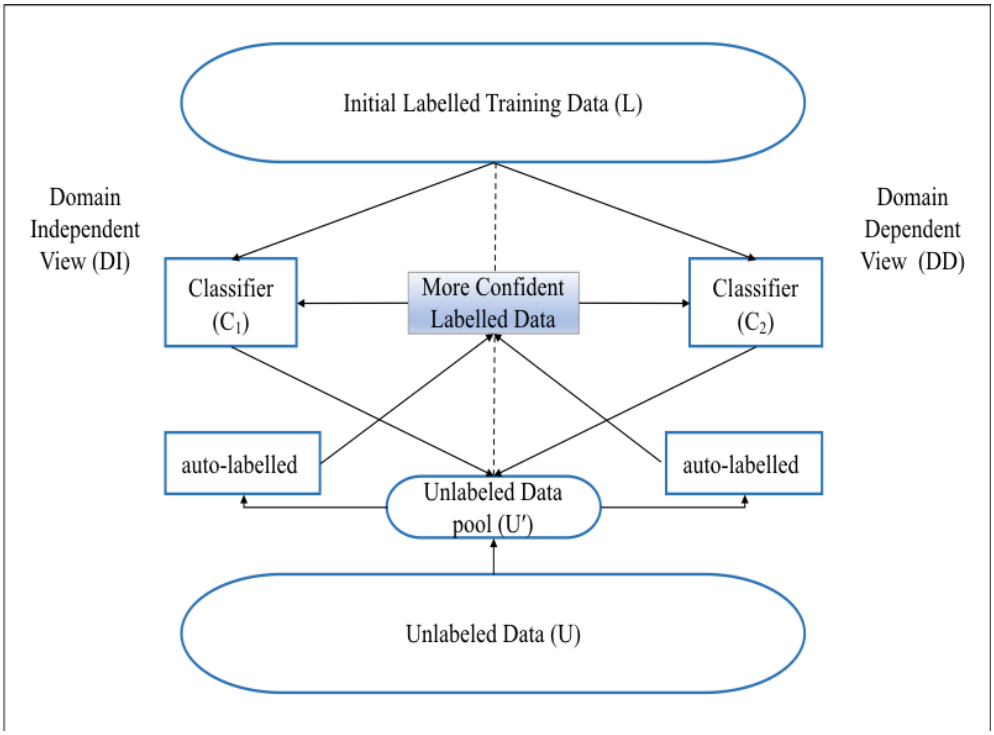


Figure 8.1: Procedure followed by the modified co-training model

In our research, we apply a modified version of the co-training model in an attempt to verify its applicability to a domain-dependent multi-classification model. This modified version differs from the original co-training model by less restrictive assumptions. This affects two aspects of co-training: 1) building the classifiers, and 2) selecting auto-labelled data.

Algorithm The enhanced Co-training algorithm

Given:

Set of L labelled training instance.
Set of U unlabeled instance.
Classifiers C1 and classifier C2

Parameters:

Set an initial pool of U' that is created by randomly sampling u instances from U.
K the number of iteration.

Loop for K iteration:

Use L to train classifier C1 with considering only DI views
Use L to train classifier C2 with considering only DD views
Apply C1 to U' to label and assign confidence scores to each example
Apply C2 to U' to label and assign confidence scores to each example
Reduce disagreement between F(C1) and F(C2)
Select examples with highest confident that is not less that a threshold to add to L
Randomly choose u examples from U to replenish U'

End loop

Output: Classifiers C1, C2 and L.

Figure 8.2: Pseudocode for enhanced co-training

In our case, the classifiers C_1 and C_2 are built using an adapted version of co-training involving two views: a domain-independent (DI) feature set and a domain-dependent (DD) feature set, based on the feature set previously used in Chapters 5 and 6.

C_1 and C_2 are generated in the co-training by DI and DD as the two different views of each post from online health communities. The DI feature set consists of the following: content-free features, which include lexical features that can be further divided into a word base and a character base (LEX); and meta-level features, which are based on several existing sentiment analysis lexicon resources; polarity lexicon features (POL) and emotion lexicon features (EMO).

The second view in the co-training is represented by the DD feature set, which consists of the following: extracting emotion-oriented features from a medical-domain-dependent lexicon (DOM), a key phrase extraction feature to extract important topic words and phrases from a text (KEY), and an n-gram text extraction feature (NG).

The feature engineering details, a description of the features used and the rationale for using them can be found in Section 5.3.2. However, the last feature in the DD feature set needed to be changed due to the use of chi-squared feature selection, which relies on a labelled dataset to score the correlation between the text and the label corresponding to it, as the co-training model is being

proposed to demonstrate the effectiveness of using unlabelled data. The n-gram feature was thus extracted without using chi-squared. Although we set the n-gram feature to the same setting as the previous work, unsupervised feature selection was applied because of the high dimensionality of a text feature. We used count-based feature selection, with seven as the minimum number of total instances required for inclusion.

8.3.2 Determining Auto-Labelled Data

The original co-training model is conducted in an iterative manner, in which classifiers assigned labels to a random set of unlabelled data with, for instance, the highest-confidence values being added to labelled sets. The model learned from the enriched labelled samples as generated iteratively by the two classifiers.

The independence assumption condition described by Blum and Mitchell [51] cannot be met in many practical situations due to the occurrence of noisy data or classifiers not being sufficiently effective on their own. Consequently, it is beneficial to relax the assumption; as suggested by those authors, an alternative is to minimize the number of examples with conflict prediction.

In practice, classifiers can be loosely dependent and may not be fully sufficient individually. For example, Biyani et al. [69], Yu [23] and Collins and Singer [119] conducted their own variations of co-training with text classification problems by minimizing the conflict prediction. The rationale for minimizing the disagreement between the classifiers is to reduce the error rate in a co-training model.

To reduce noise, we applied co-training with a weak assumption by minimizing the auto-labelled data and selecting only examples on which both classifiers agreed. We then selected the top-N highest confidence examples from the comparable high confidence view classifiers according to the class distribution shown in Table 8.1, subject to a threshold θ.

In each co-training run, and after training each classifier on a DI and a DD view on the current L dataset, N candidates were selected from the U' training sample that showed the highest levels of confidence from the comparable selection of the two learners for each example. Then, the L dataset was updated with the selected candidates if the confidence score was higher than the threshold θ. In this study, N was set to values that represented the underlying distribution of the original dataset, as shown in Table 8.1.

To select the threshold, we conducted experiments using pre-set thresholds θ of different values (0.65, 0.75, 0.85, 0.95 and 0.99) and without a threshold (see Figure 8.3) on one of the datasets (i.e., the Lupus dataset). The co-training model run without the addition of a threshold had an average accuracy of 59.9% over the two proportions of labelled data (20% and 25%). The average accuracy also generally increased when raising the value of the threshold. There was about a 3-percentage improvement in performance when using a threshold, starting from 0.65 and increasing until reaching 0.95. However, performance decreased slightly when increasing the value to 0.99. This might be due to narrowing the number of examples that met the condition, which resulted in fewer labelled data, reflecting lower model performance. Therefore, the threshold θ was set to 0.95 for the co-training experiments.

Table 8.1: Ratios of the class distribution for both datasets

	Asking about treatment: Lyme/Lupus infection confusion: Lyme/Lupus symptoms confusion: Depressed and frustrating: Motivation and encouragement: Seeking for general information
Lyme Dataset	3:2:3:1:3:1
Lupus dataset*	6:6:4:1:1:2

* Distribution is based on the average of the least two classes.

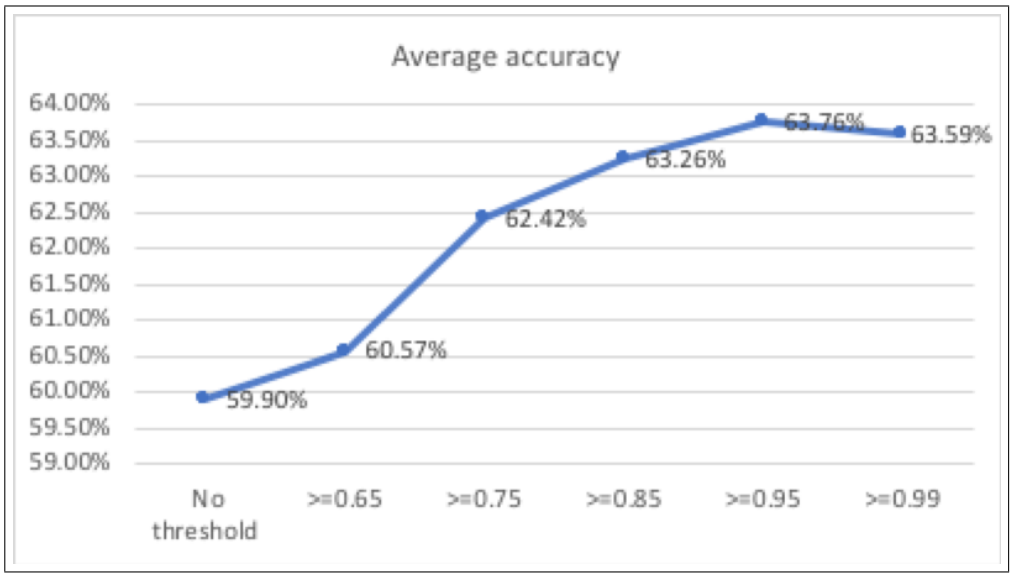


Figure 8.3: Average accuracy of different threshold values. When the confidence score is higher than the threshold θ, the labelled dataset (U') is selected as the training data (L).

8.3.3 Baseline Runs

As illustrated in Figure 8.4, each dataset was randomly divided into three portions: a test set for evaluation process; then, as a co-training dataset, a labelled (L) dataset and an unlabelled (U) dataset. For each experiment, 20% of the data were reserved as an evaluation dataset and 80% were set as a co-training training dataset, which consisted of L and U. Datasets were partitioned using different proportions of L and U, with 20% or 25% of the training data for each class allotted to L and the remainder of the dataset treated as U. We followed a similar class distribution ratio for the proposed dataset for L, U and the test set.

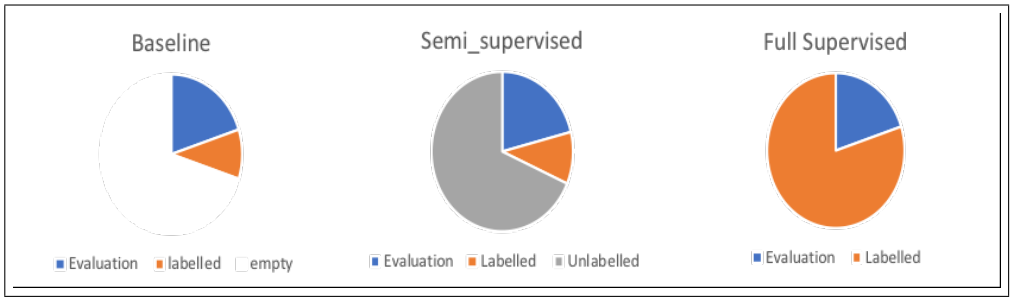


Figure 8.4: Data split for semi-supervised learning runs, baseline supervised learning runs, and fully supervised learning runs.

The performance of each of the co-training models was compared with the performance of the corresponding baseline, which included only the feature hashing (i.e., following the baseline characteristics outlined in all previous chapters) and the fully supervised algorithms (i.e., that included all the features previously described).

8.4 Experiment Setup

8.4.1 Number of Iterations

To achieve full co-training potential, the number of iterations (K) was not strictly defined. Instead, the co-training process stopped when an unlabelled dataset (U) could not supply enough data to replenish the smaller pool of the unlabelled dataset (U').

8.4.2 Unlabelled Data Available at Each Iteration

First, we selected the labelled data from each class according to the class distribution. Then, in each iteration, N unlabelled candidates were selected from U to replenish U' in each iteration of the classification process, classifying all unlabelled data U, or generating a new set of U' from U to be classified in each iteration. Experiments conducted by Yu [23] suggest that replenishing U' significantly outperforms building U' from a new set. Furthermore, when Blum and Mitchell [51] and Yu [23] implemented experiments that allowed classification of examples directly from the larger unlabelled dataset U, they found that co-training achieved higher results when using a smaller pool U'. The rationale behind this could be that replenishing U' could force the classifiers to reclassify the examples with a low prediction score as a result of the previous iteration. On the other hand, generating a new subset U' in each iteration leaves the classifier to classify simple examples [23]. Therefore, we replenished U' with N candidates in each iteration for all the experiments in this study.

To determine the number (N) of unlabelled pool (U') entries to be learned from the two classifiers in each iteration process, experiments were conducted using different percentage values: 6%, 8%, 16%, and 48%. In this study, the percentage attribute was used, rather than a fixed number of examples, due to the dissimilarity of the data volumes, so that a fair and consistent comparison could be made. From the experiments conducted on one of the datasets (i.e., the Lupus dataset) and by calculating the average improvement with 25% and 20% of labelled data, it was found that when running the classifiers on 48% of the U dataset on each iteration the performance of the co-training scored an average accuracy of 58%, as shown in Figure 8.5. Reducing the proportion of unlabelled data used to run classifiers to 16% for each iteration increased performance by 2.5%. Co-training runs classifying 8% of the unlabelled dataset at each iteration increased performance by 1.4%, and co-training runs classifying only 6% of the unlabelled data on each iteration improved accuracy by about 1%. Consequently, U' was set to an optimal value of 6% for the ongoing experiments on unlabelled dataset U.

The choice of the number of examples that should replenish U' within each iteration reflects the co-training learning rate. To achieve stability in the co-training model, the number of examples should be low in volume compared with the original L dataset and represent the underlying distribution of each class, and within each dataset. For this reason, we set the value of examples

that should be added from U to U', as shown in Table 8.1, reflecting the data distribution for Lyme Disease and Lupus discourse.

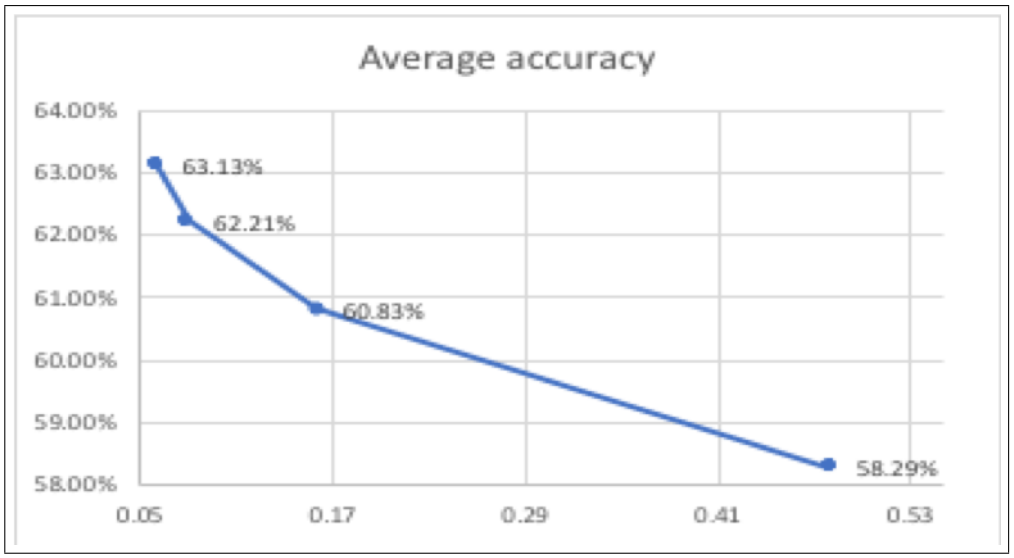


Figure 8.5: Average accuracy of different proportions of N to create an unlabelled pool (U') for the two classifiers in each iteration process. Experiments were conducted using the following percentage values: 6%, 8%, 16%, and 48%.

8.4.3 Platform

When we tried to develop the model using the Microsoft Azure Machine Learning Studio, we faced difficulties regarding the platform, as the algorithm needs to be run iteratively. For this reason, we implemented the co-training model using the Waikato Environment for Knowledge Analysis (WEKA). WEKA was developed at the University of Waikato and is a popular open-source resource that consists of the implementation of different machine learning algorithms, written in Java. We accessed WEKA using the Python wrapper for the Java machine learning workbench Weka using the JavaBridge library.

All the experiments included in this section were conducted using SimpleLogistic supervised learning algorithms (i.e., from WEKA[19]), with all parameters set to their default values.

According to Landwehr et al. [120], SimpleLogistic uses LogitBoost and a simple regression in each iteration to select the most relevant attributes in the data when

[19]WEKA is Waikato Environment for Knowledge Analysis. The wrapper is available for download at: `https://pypi.org/project/python-weka-wrapper/`

performing logistic regression and stops before convergence to the maximum likelihood solution.

8.5 Experiment Results

The co-training model was trained using two classifiers with two different views (domain independent and domain dependent) and across two topical disease datasets (Lyme Disease and Lupus). The results are summarized in Table 8.2. Baseline supervised learning classifiers were trained using only 20% and 25% of each class of labelled data from the co-training training dataset, whereas the fully supervised learning classifiers were trained using 95% and 100% (i.e., the whole co-training training dataset) of the labelled data from the co-training training dataset.

To conduct a deeper investigation of our co-training model, we implemented the initial co-training with parameter settings comparable to our modified co-training. In the initial co-training model, after we relaxed the assumption of consistency in each iteration and we followed the default setting in selecting examples from the auto-labelling data, as followed by Yu [23], Li et al. [116] and Biyani et al. [69]. The highest confidence examples of each class from C_1 and C_2 were chosen to expand the labelled data. The results of the co-training model are shown with regard to coverage performance, as reflected by the last iteration of the co-training model.

In the Lyme Disease medical forums dataset, modified co-training runs using two classifiers (DI and DD views) improved the performance of the corresponding initial co-training, each of which outperformed the baseline. For example, the first co-training run using two classifiers, each trained with 239 initial labelled posts, showed a classification accuracy of 59.4%; the modified co-training run using the same two classifiers achieved an accuracy of 62.1%; and the baseline for the same classifier algorithm performed with an accuracy of 51%. Thus, the modified co-training run outperformed the baseline supervised run by more than 11% in the evaluation measure. The fully supervised model, which trained its classifier initially with 1,133 labelled data, scored 66.8% accuracy, a gain of only 4.7% over the modified co-training runs, but required an additional 894 labelled data items.

For the Lupus medical forums dataset, the baselines had an accuracy of 54.4% when training supervised learning with 173 labelled posts and 59.5% with 218

labelled posts. The modified co-training model improved performance by 7.8% and 4.6% over the corresponding baselines. The modified co-training accuracies for both labelled data portions were outperformed by the fully supervised learning run by only 3.3% and 2.8% when training their classifiers with a full set of training data.

From a general perspective, the modified co-training model improves the accuracy of classifying medical sentiment as a multi-class classification problem. For both disease datasets, co-training runs generally outperformed their corresponding baseline supervised and initial co-training runs and almost matched the performance of fully supervised classification learning.

For Lyme Disease data, the evaluation measure for the modified co-training reached midpoint between the initial co-training and fully supervised learning. However, in the Lupus dataset, the initial co-training did not improve on the baseline, which we suggest is attributable to the amount of data used to train the learners, an imbalance in class distribution, or a combination of the two.

Table 8.2: Classification accuracy (%) of co-training with DI and DD views

	Lyme Disease dataset		**Lupus dataset**	
Run type	**Labelled data**		**Labelled data**	
	20%	**25%**	**20%**	**25%**
Baseline	51.01%	58.05%	54.38%	59.45%
Initial co-training	59.40%	61.75%	54.38%	58.53%
Modified co-training	62.08%	65.44%	62.21%	64.06%
Fully supervised*	66.78%	69.80%	65.44%	66.82%

* Fully supervised runs using an additional 55% of labelled data.

8.6 Summary

In this chapter, we aimed to address the shortage of sentiment labelled data that can prevent researchers developing accurate and productive sentiment classification approaches by examining semi-supervised learning. We formulated our multi-class sentiment analysis as a co-training problem, a widely known and effective semi-supervised learning algorithm. A co-training approach was developed using a feature-based model for the sentiment recognition of user

posts in online health communities. We studied less restrictive adaptations of co-training models based on views that naturally partitioned into two sets: DI and DD views.

From the results of this study, we can conclude the following: (1) our experiments showed that modified co-training performs well when combining the feature information in medical sentiment classification; (2) even with limited labelled data, the performance of modified co-training is able to outperform the corresponding baseline in both datasets; and (3) the achievements of the proposed co-training model suggest that it is capable of approaching the performance of corresponding fully supervised learning trained on a large amount of labelled data.

All the sections of this chapter were published in Alnashwan et al. [114].

Chapter 9

Long-form Content and Topic Analysis

9.1 Introduction

In this chapter, we present a framework to analyse socially-generated medical content using two approaches complementary to sentiment analysis — *content analysis* and *topic analysis* — to ascertain their usefulness to various interested stakeholders, such as patients, medical professionals and policy makers. This work falls within the remit of an emerging field of study: what Denecke calls *Health Web Science* [6].

Each of the three approaches to data analysis has the potential to reveal different aspects of the data. As we discussed previously, sentiment analysis is the computational analysis of people's opinions, sentiments, emotions and attitudes expressed towards a concept or entity [5]. On the other hand, automatic content analysis of text documents relies on assessing high-frequency terms in order to deduce concepts or themes from text [122], while topic analysis is a method for identifying topics in a text collection and is based on a probabilistic model [123].

This chapter is organized as follows. Section 9.2 discusses the impact of medical social media from different stakeholder perspectives. We then introduce a framework that utilizes two additional techniques to analyse the information: content analysis is presented in Section 9.3 and topic analysis in Section 9.4. Section 9.5 contains discussion of the analyses and Section 9.6 presents a summary of the chapter.

9.2 Stakeholder Perspectives

Health-related social media is increasingly being generated, shared, and analysed [6]. The content is, primarily, narrative text about a medical topic that has been produced by individuals, doctors and other healthcare professionals. One of the most important benefits that social media offers is that health and medical information can be disseminated more widely and accessed more conveniently across a broad range of the population, despite differentiation in age, education level, gender and locality; this contrasts with traditional methods of communication and information dissemination. Although social media provides several benefits to the health community, there are issues regarding the quality and trustworthiness of the information and patient privacy [124]. There are many interested parties, including patients, medical professionals and social data analytics researchers. We separate the impact of medical social media into three different stakeholder perspectives: the individual or patient, the medical professional, and the organizational or policy viewpoint. In the same vein, Denecke gives the perspectives of three stakeholders who are invested in medical social media: patients or individuals, health professionals, and biomedical researchers, which can include organizations and researchers [6].

9.2.1 Individual Perspective

Individuals include patients who may be suffering from a health condition, and their families and carers, who participate on social media platforms. Social media has had a considerable impact on the speed and nature of the interaction between individuals and between individuals and health professionals [124]. It has also become an increasingly important medium for those seeking medical information. A recent survey [102] reported that about 35% US adults go online to identify a specific medical condition they, or one of their relatives, might have; these individuals are known as “online diagnosers”. In the report, 41% of the diagnosers asserted that their diagnosis, condition or disease was later confirmed by a medical professional. In addition, 26% of online users tend to search for others’ health experience in social media.

Online forums provide opportunities for individual participants to extend their reach beyond local general practitioners and hospitals [125]. Patients have the potential to use a variety of online health forums to seek answers to general questions, share their health knowledge (which is sometimes lacking in accu-

racy), provide and gain emotional support, learn from others, and express their concerns [126]. They also share their emotions and details about how they are feeling in order to have a better quality of life during their treatment.

9.2.2 Professional Perspective

Healthcare professionals include physicians, nurses, therapists, and public health officials. Healthcare professionals are becoming increasingly aware of the important role of patient involvement and patient preference in their healthcare, which can now be seen in social media discourse and play a part in medical decision making [6][127]. For example, in contrast with traditional surveys, online forums are a convenient way of obtaining a large amount of information from users and their families in order to evaluate patients' opinions and moods [6].

Physicians engage in social media for a number of reasons: to network and to consult other professionals about a complicated case, to gain insight into patient sentiment, to share information, and to give of their time by answering patient questions online. Denecke [6] reported on physicians using social media to enhance patient care. Medical professionals can also gain valuable insights into how patients are able to support and influence each other. One controlled study [128] demonstrated that written patient testimonials can have a significant impact on treatment choices.

9.2.3 Organizational Perspective

Health organizations include hospitals, professional societies, clinics, health systems and pharmaceutical companies. Organizations in general can benefit from social media to improve their image and visibility, market their services or products, seek funding, and create an online space in which to publish news about their activities [6]. Online technologies have also become a fundamental part of public health surveillance [127]. Web information is now a valuable resource for public health agencies that depend on different sources for daily surveillance activities. Moreover, analysis of these data could enhance the potential for biomedical researchers to predict disease outbreak more accurately [6].

Policy makers in health organizations and government health departments are also increasingly mining user-generated content on medical social media

platforms to inform decision-making or assess service quality [129].

9.2.4 Related work in stakeholder identification of medical social data

Lu et al. carried out a study where they also linked sentiment and content analysis of medical social media to stakeholder perspectives [130]. Whereas we focus primarily on Lyme Disease they collected data related to lung cancer, diabetes and breast cancer from an online forum (MedHelp). They identified three significant stakeholders: patients, caregivers, and specialists. The SentiWordNet lexicon was used to identify sentiment. Their content analysis consisted of feature extraction, probabilistic clustering, keyword extraction, and topic identification. Their feature extraction used word n-grams and UMLS terms. They recognized five significantly different health-related topics: symptom, examination, drug, procedure, and complication. They found that the topics of interest and sentiment expression differed significantly among different stakeholders across the different disease forums.

9.3 Content Analysis

In our work, content analysis was performed to obtain a general view of the medical information available in online discussions related to Lyme Disease. Automatic content analysis of documents relies on statistical and natural language processing methods to extract concepts or themes from text. Swartz and Ungur provide an overview of the various methods used to analyse social media data [122]. The rule of thumb for content analysis in qualitative research is to consider high-frequency words in order to identify vocabulary of strong potential interest [131]. For this reason, we extracted the most frequent concepts from medical posts using text mining techniques. To achieve a more semantically meaningful analysis, we mapped natural language to standard medical terminologies. The process is illustrated in Figure 9.1.

The gold standard dataset presented in Chapter 5, which was used in automated sentiment classification recognition from extracted posts related to Lyme Disease discussions, was used for further content analysis to draw additional insights from the same dataset. For the content analysis, we extracted key phrases[20]

[20]We used the Key Phrase Extraction model from Azure ML Studio implementation.

to derive meaningful concepts. Key phrases are those that contain single noun words, compound nouns or modifiers and nouns. The extracted key phrases were represented as n-gram features in a vector space representation. We used the widely known Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme, a numerical score calculated to reflect how important a term is to a post across a dataset. The TF part of TF-IDF is a statistical value that increases in proportion to the number of occurrences of a term in a document, while IDF relates to the frequency of the term in the data collected as a whole. Both word unigrams and bigrams are included in the term weighting. We applied count-based feature selection to exclude terms that appeared fewer than five times in the feature vector space. This process resulted in identifying 762 concepts from the dataset.

The concepts identified were further reduced using domain-specific terminologies by mapping the terms to the Unified Medical Language System (UMLS) Metathesaurus. The UMLS is the largest repository of biomedical terminology in the world. It consists of knowledge that represents around 1.7 million biomedical concepts, classified under 133 semantic types. We used the US National Library of Medicine's MetaMap[21] [132], a highly configurable program, to map text to concepts in the UMLS Metathesaurus. MetaMap provides a Java application interface to analyse text and extract related domain-specific concepts. Using MetaMap, we mapped concepts to a representative semantic type; otherwise, we considered the concepts as they were. This resulted in a reduced total of 144 concepts or semantic types.

Excluding concepts that did not represent any semantic type further reduced the number of concepts, resulting in 98 concepts or semantic types. Further reduction was then achieved by considering only those semantic types that appeared more than 10 times. This process resulted in identifying the 33 most frequent semantic types.

To illustrate these concepts and represent them in a meaningful way, we measured the similarity between the concepts using the Word2Vec[22] library. We adopted Word2Vec because it is an efficient implementation of a combination of models that are used in word embedding. Word2Vec is based on continuous bag-of-words and skip-gram architectures to configure the vector representation of concepts. In Word2Vec, the input is a large text dataset that is used to recon-

[21] `https://metamap.nlm.nih.gov/`
[22] https://code.google.com/archive/p/word2vec/

struct a vector space, in which words with similar meaning are proximate to one another in the space. To measure the similarity between each concept, we used a package called Gensim, which is a Python implementation of Word2Vec. We applied the pretrained model that used Google News dataset of about 100 billion words to generate a large word vector space. A distance tool is used to measure the similarity between words. After the similarity matrix had been constructed, and for representation purpose, a simple technique was needed in order to group similar concepts together. We used a hierarchical clustering function that performs agglomerative clustering with complete linkage method as default parameter, to perform a cluster analysis. A chart was then used to illustrate the family relationships in a conventional tree structure.

Figure 9.2 illustrates the concepts of most concern among the members of online health communities related to Lyme Disease, as represented by the 33 domain-specific semantic types. It is possible to draw several observations from Figure 9.2. For example, we can observe that forum participants mention general medical concepts that are related to Sign.Symptom and Disease.Syndrome semantic types. At the same time, the figure includes concepts related to Body.Location.Region and Body.Part.Organ.Component sematic types, which appear with a high degree of frequency. In the forum data, various concepts, such as area name, city or county, which can be defined as the Geographic.Area semantic type, were repeatedly mentioned in Lyme posts; this arises from the fact that there are a number of concerns about locations that have high risk of infection, a common issue with Lyme Disease.

Investigating this content analysis in greater depth could assist in developing the knowledge of practitioners, doctors and organizations so that they can understand the concerns aired by online members of Lyme Disease discussions. Mapping the concepts of most concern to users to domain-specific semantic types could help address the terminology gap that exists between health professionals and health consumers.

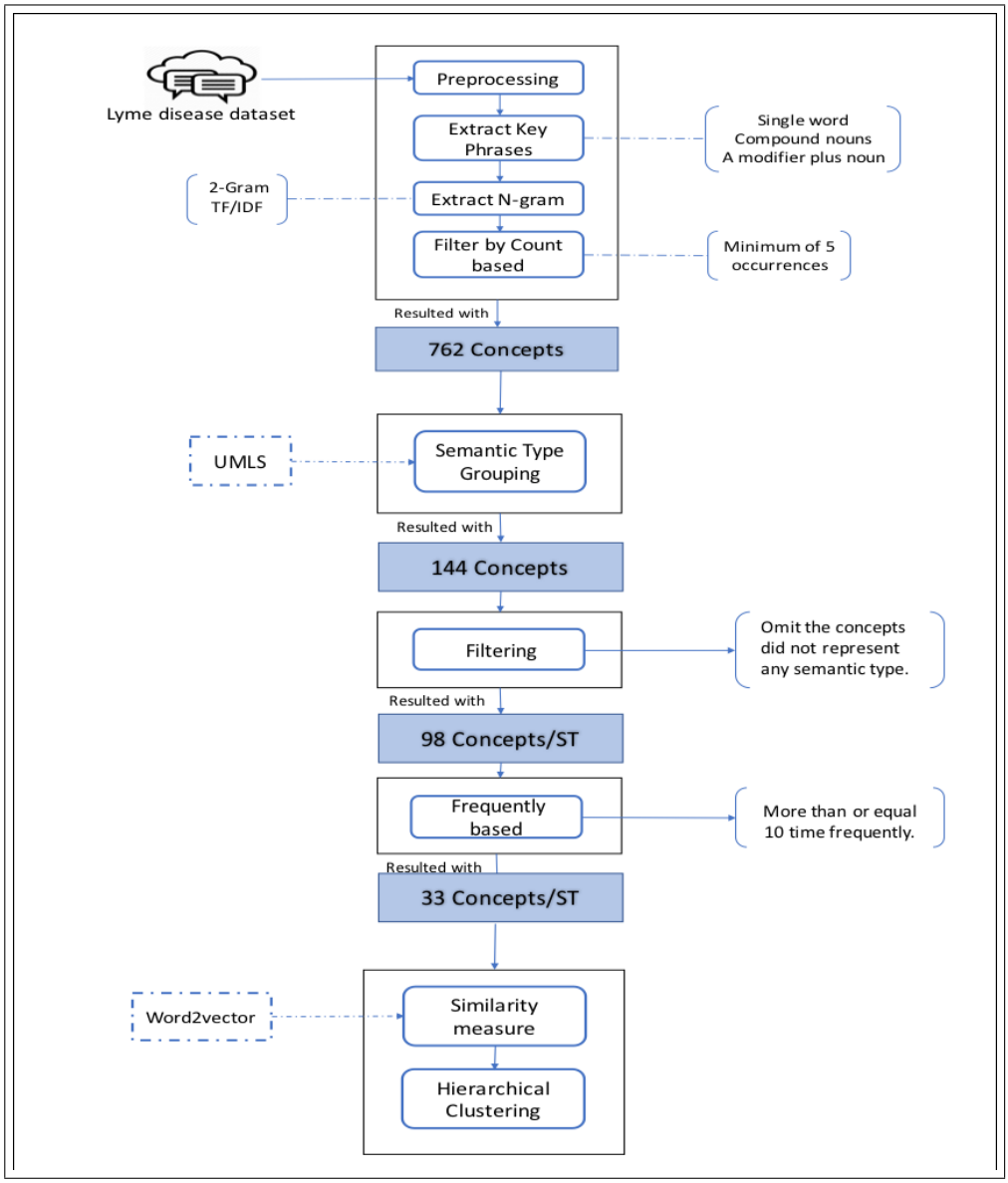


Figure 9.1: Content analysis process

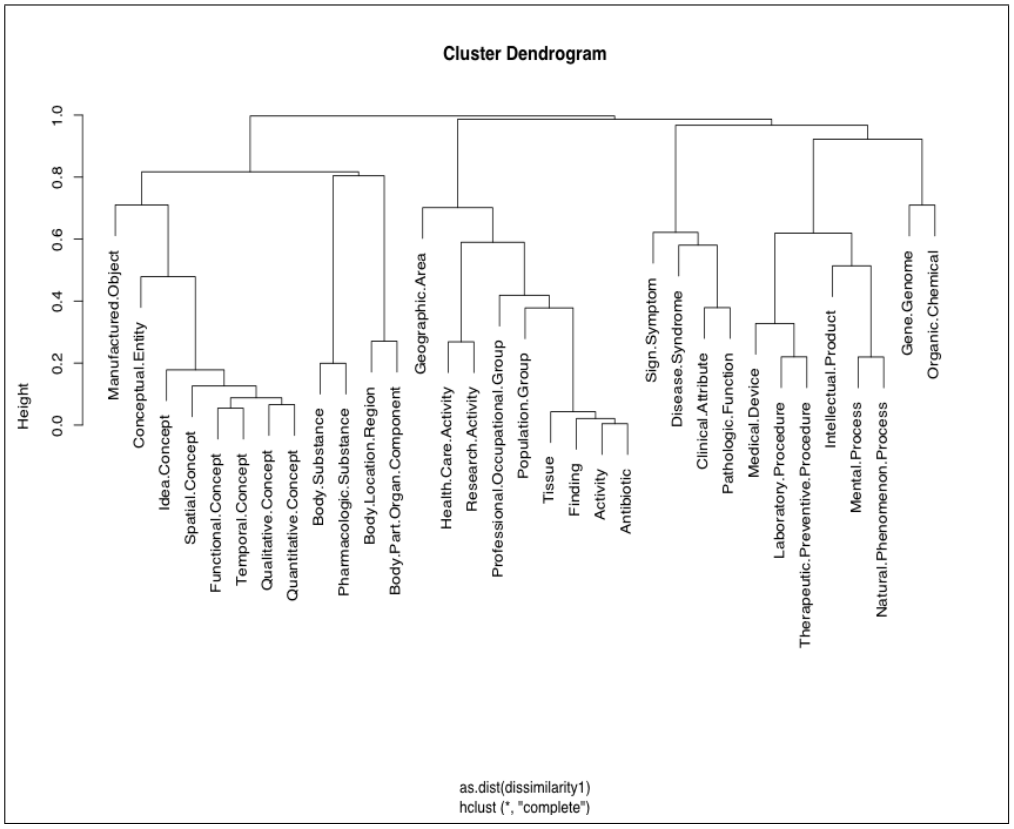


Figure 9.2: Presentation of extracted process

9.4 Topic Analysis

The other analysis we carried out was topic analysis. The field of topic analysis in social media has attracted the interest of many researchers working on the mining of discussions and the detection of hot topics of concern [43][133]. We aimed to detect different topics represented by participants or patients in the health community related to Lyme Disease forums.

Topic modelling approaches, such as the latent Dirichlet allocation (LDA) model [123], are commonly used to automatically identify hidden topics from a corpus [43][133]. LDA is based on a probabilistic model that depends on a hierarchy of components, the basic intuition being that each document can be assigned to multiple latent topics and each latent topic is distributed over a range of words [123]. We employed an unsupervised LDA generative model in order to detect topics in posts.

In this study, we first used pre-processing, such as Lemmatization, stop word removal and tokenizing, which then enabled us to generate topics. From a

technical viewpoint, we employed the Mallet[23] implementation to increase efficiency in the implementation of the LDA model.

Various topic coherence matrices have been proposed in order to evaluate topic quality. For example, following extensive study, Röder et al. [134] proposed a topic coherence measure derived from a combination of some known components, which resulted in a higher correlation to human ranking, where the best performance compared well with other existing measures.

We computed the coherence measure proposed by Röder et al. [134] in order to identify the optimal number of topics from the data by testing the different values that represent the number of topics, from 2 to 30. After running the model multiple times, we found that the value of the average coherence increased until it reached seven topics but decreased after that, indicating that this value optimises the performance of the model. Therefore, we decided to set the number of topics to seven in the LDA topic modelling.

The LDA model results were interpreted and each distribution forms a collection of words representing seven different topics. Table 9.1 illustrates the 20 keywords most associated with each topic.

Using topic analysis, we have thus identified seven significant "hot" health topics related to Lyme Disease (see Table 9.1). Topic labels were then interpreted by two medical experts based on the keywords extracted. According to these experts, the emergent topics found in the topic analysis of the Lyme Disease online discussion were seen to cover the general characteristics of the condition.

Topics #1 and #6 were closely related to symptoms, which is not surprising as the symptoms of Lyme Disease can cause a lot of confusion. Furthermore, people tend to discuss symptoms a lot, as these become one of their greatest concerns. Although Lyme Disease is known as the great imitator, there are some initial symptoms that could indicate infection after being bitten by a tick, such as starting to notice a rash on different parts of the body. The duration and body area are important factors in identifying the initial symptoms, as found in topic

[23] `http://mallet.cs.umass.edu/topics.php`

[24] Doxy stands for Doxycycline, an antibacterial medication.

[25] LLMD stands for Lyme Literate Medical Doctor.

[26] A die-off reaction, also called a Herxheimer reaction, can occur when treating the Lyme germ, some co-infections, and types of yeast.

[27] Abx is a medical abbreviation for antibiotics.

[28] A Western blot test is typically used to confirm a positive HIV diagnosis. The Western blot test separates the blood proteins and detects the specific proteins (called HIV antibodies) that indicate an HIV infection.

Table 9.1: Topic modelling results

	Topic name	Top 20 words
1	Initial symptoms after exposure	Start, day, week, month, time, ago, rash, doctor, long, back, doxy[24], bite, bit, notice, year, experience, area, recently, idea, develop
2	Online patient communication	Post, disease, find, patient, support, information, share, great, group, site, call, read, article, info, make, story, cure, issue, news, forum
3	Mental state	Feel, bad, time, thing, make, sick, good, hard, work, lot, anxiety, give, life, today, back, night, live, part, year, lose
4	Outline of the disease	Disease, tick, chronic, find, doctor, treat, year, patient, treatment, infection, illness, medicine, people, include, diagnosis, bacteria, research, health, case, dr
5	Treatment modalities	Treatment, antibiotic, good, experience, llmd[25], hear, treat, work, read, med, year, eat, people, herx[26], put, supplement, abx[27], make, continue, give
6	Symptoms	Pain, symptom, feel, body, leg, muscle, joint, head, problem, eye, severe, hand, foot, leave, normal, fatigue, arm, feeling, back, headache
7	Diagnostic testing	Test, symptom, blood, positive, result, year, doctor, negative, diagnose, show, low, high, came_back, lyme, doc, band, lab, western_blot[28], question, genex

#1 (Initial symptoms after exposure). However, Lyme Disease can include a combination of symptoms and these can differ from person to person. Topic #6 highlighted symptoms such as pains in the legs, muscles and joints, fatigue and headaches.

Diagnostic testing was another major topic (#7), as laboratory blood testing (to identify antibodies to the bacteria) is the primary means to confirm a clinical diagnosis. However, some blood tests can produce a false-positive result, so that patients tend to undergo a more reliable test, called the "Western blot test", which was also found in the keywords.

Topic #5 addresses Lyme Disease treatment modalities. From this topic, it can be observed that the only medication mentioned is antibiotics, as this is the standard treatment for Lyme Disease. The posts related to the treatment topic demonstrate abx as a keyword, which is a medical abbreviation for antibiotics.

Topic #3 encompasses patients' mental state. Lyme Disease can affect sufferers physically and mentally – patients can experience anxiety, depression and a sense of loss, which can affect their lives, work and sleep patterns. In response, the community surrounding Lyme Disease patients often provides vital emotional support to help them cope with what they are facing. Therefore, support was

one of the topics most discussed in the posts, topic #2 highlighting online patient communication. This communication motivates others to post their experiences and stories and to create and participate in groups such as forum communities. Patients need to be motivated by being equipped with coping strategies, such as taking up a hobby, reading articles or books, or following the news.

From the above, it can be seen that topic analysis facilitates interpretations that would allow medical professionals and researchers to better understand citizens' concerns expressed via online posts in Lyme Disease support groups.

9.5 Discussion

Social media is a powerful tool that gives a new dimension to healthcare by acting as a peer-to-peer social interaction mechanism between a range of individuals. Analysing socially-generated medical information can offer valuable benefits, from the patient, professional and organizational perspectives. The benefits gained from various analysis techniques could help diverse healthcare audiences due to their focusing on different aspects of the information collected. A certain type of analysis could be suitable for a particular healthcare audience. For example, the aim of Zhan et al. [133] was to identify topics hidden amongst posts generated by e-cigarette customers, in order to provide effective support to policy makers. Furthermore, the benefits can cut across groups and overlap. There are also different ways of analysing the same dataset and a single analysis might not be equally relevant to each group of patients, professionals and organizations. However, mining discussions using different analysis technique or a combination of approaches could benefit one or a range of patient, professional and organizational perspectives.

To gain a thorough understanding of social interactions in Lyme Disease discussions, including the work in Chapter 5, we implemented a framework that utilizes three different techniques to analyse the same information: sentiment analysis, content analysis and topic analysis. The valuable conclusions presented by the different analysis techniques in this work suggest several practical benefits that could have both direct and indirect impacts on health communities, as well as health organizations and practitioners, due to the potential for enhancing health outcomes. We outline below the benefits for patients, professionals and organizations, which include, but are not limited to, the following.

9.5.1 Benefits to Various Stakeholders

Individuals can be overwhelmed by the sheer volume of information available online on any given topic. The notion of information overload [135] and related concepts, such as information anxiety [136], are significant here. If the outcomes of analyses (sentiment classes, concepts and topics) were available to individuals, it could provide a useful map of the territory or information space that would allow users to navigate the information with more confidence and ease. Visualizations of sentiment, content, and topic analysis can greatly aid understanding; see, for example, the work in Wang et al. [137].

These analyses are a potentially valuable activity for healthcare managers in transforming text data into quantifiable information that can add value to health professionals' work [138]. Sentiment information can alert professionals to particular areas causing patients distress. Content analysis can elucidate the knowledge held by patients, practitioners, doctors and organizations, so that they can understand the concerns aired by online members in Lyme Disease discussions. Mapping the concepts that most concern individuals to domain-specific semantic types can also help address the terminology gap that exists between health professionals and health consumers. Topic analysis provides insightful results that would allow medical professionals and researchers to understand citizens' most pressing concerns.

One scenario is that of a health organization, such as the Food and Drug Administration in the US, or a medical professional in public health wishing to collect specific data about a disease. They might want, for example, to be able to identify all the symptoms experienced by different patients with Lyme Disease. With our framework, sentiment analysis can automatically identify posts related to the "Lyme symptoms confusion" category, which includes informative discussions about all the symptoms experienced and patients' confusion in this respect. Furthermore, the topic analysis results identify the range of symptoms that have been discussed across community groups, and posts that relate to a specific symptom of interest, which could be extracted for further examination.

In general, there are various benefits that can be gained from the three forms of analysis proposed. For example, by helping health organizations to direct investment to those areas that are of most help to the community, it is possible to optimize investment in generating awareness in a cost-effective manner. Moreover, this type of work could help identify misinformation regarding health issues and disseminate pertinent health information to target communities.

Unfortunately, inaccurate medical information can be promulgated on social media such as online forums, be it inaccurate diagnostic criteria or bogus or unproven treatments. This may be unintentional or malicious. It is useful for medical professionals to be aware of what possibly erroneous ideas are circulating in the community at any time so as to alert and inform patients.

9.5.2 Health Information Leaflet and Expert Comparison

We also conducted a cross-comparison of the topic analysis results with several information leaflets on Lyme Disease intended for the general public. These leaflets were all produced by reputable health organizations and public health agencies: the World Health Organization (WHO) in the European Region[29], the Centers for Disease Control and Prevention (CDC)[30] in the US, the National Health Service (NHS) and Public Health England (PHE)[31] in the UK, and Health Canada[32]. As can be seen from Table 9.2, information contained in the leaflets and the topics identified in our analysis show largely similar themes.

An interesting observation is that there are two topics (see Table 9.2) – Location and Prevention – that are mentioned in the leaflets, but are not frequently discussed by patients on social media; correspondingly, there are two social media topics – Online Patient Communication and Mental State – that do not have a corresponding representation in the leaflets. We would speculate that location and prevention would be more appropriate prior to any exposure, and would be more likely found in information brochures; communication and mental state are of more concern to sufferers or would-be sufferers, and would therefore be more topical in online forums. Also, one could posit that organizations and public health agencies should consider including a description of psychological effects in their leaflets because these seem to be of considerable concern to those engaged in online media discussion about Lyme Disease.

We also consulted expert opinion as to the usefulness of the proposed sentiment analysis classification model for Lyme Disease posts (i.e., as one case of analysing medical user-generated discourse). This was done through email discussion

[29] http://www.euro.who.int/__data/assets/pdf_file/0008/246167/Fact-sheet-Lyme-borreliosis-Eng.pdf?ua=1

[30] https://www.cdc.gov/lyme/resources/toolkit/factsheets/FS-ChildrenLymeDisease-508.pdf

[31] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/694158/PHE_Tick_Leaflet.pdf

[32] http://healthycanadians.gc.ca/publications/diseases-conditions-maladies-affections/pamphlet-lyme-brochure/alt/pamphlet-lyme-brochure-eng.pdf

Table 9.2: Comparison of topics contained in leaflets and those identified on social media

	Topics	WHO	CDC	NHS	PHE	Health Canada
1	**Initial symptoms after exposure**	x	x	x	x	x
2	**Online patient communication**					
3	**Mental state**			Post-infectious Lyme Disease mentioned		
4	**Outline of the disease**	x	x	x	x	x
5	**Treatment modalities**	x	x	x	x	x
6	**Symptoms**	x	x	x	x	x
7	**Diagnostic testing**	x	x	x	x	x
A	**Location**	x	x			x
B	**Prevention**	x	x	Link	x	x

that included open-ended questions. An executive summary of the research was shared with the participants in order to make them aware of our work. From the 35 practitioners we initially contacted, a total of seven medical experts participated who are interested in Lyme or infectious disease and were from different regions (i.e., the UK, US and Ireland).

Our observations from the discussion revealed that, from the medical experts' point of view, our sentiment analysis model is of value from various perspectives, whether medical, patient, organizational or service provision (see Table 9.3). One of the participants was, however, unable to see any value in doing such analysis; that expert did not have any confidence in the quality of Internet social data and regarded information posted on the Internet as potentially misleading. In contrast, one of the advantages stated by the physicians is that the analysis could be a way for patients to access accurate data that would meet their needs and concerns. Another advantage mentioned was that it provides an effective and focused reading of online posts and can be used by medical experts as a means of 'listening to patients'. According to one of the physicians who participated, "medicine is a healing art which is garnished by a veneer of science. Science is subject to changes and is constantly updated. The healing art remains and requires active listening to the patient". According to another participant, "any number of services wanting to be patient-focused (e.g., those aimed at new or potentially diagnosed, those aimed at correcting misbeliefs

about Lyme Disease, etc.) could use this to generate FAQs covering these sort of themes". This type of analysis could also assist medical researchers by identifying what variations people have experienced in their Lyme Disease journey or what treatments have worked for them. This information could be used to supplement medical textbooks in order to expand and disseminate knowledge about Lyme Disease amongst medical practitioners.

Table 9.3: Benefits of our model from the physician participants' point of view

Main takeaway points	Perspective		
	Medical	Patient	Organization/ public health
A method of being able to "listen to the patient"; which is as important as laboratory data.	✓		
Healing and science are seen as separate entities. The model could help practitioners with active learning as part of healing (constant) as the science changes (variable).	✓		
Assists understanding of the conflict in diagnosing Lyme Disease.	✓	✓	✓
Supports psychiatrists in providing a direct link for interpreting thought patterns to assist in therapies such as cognitive behavioural therapy (CBT).	✓		
Plays a role in supporting experts to protect patients' health.	✓		
Helps to understand the hysteria and chaos surrounding this infection.	✓	✓	
Enhances patient-focused communication by providing relevant and needed information.			✓
Can be a way of disseminating the right data.	✓	✓	✓
Allows observation of the variations in Lyme Disease symptoms people have experienced or what treatments work for Lyme Disease amongst the public.			✓
There is no value in performing the classification as there is so much misinformation on the Internet about Lyme Disease.	–	–	–

9.6 Summary

We believe that understanding the sentiments, concerns and themes generated by participants on social media is very meaningful work. Our framework was used to investigate and comprehensively analyse an online health community, using various text mining techniques. In this chapter, we presented the methods we used to process a large amount of text from discussions on Lyme Disease in order to conduct content and topic analyses in addition to the sentiment analysis we had already carried out. We identified multiple concepts of most concern, which

we then mapped to medical terminology to achieve semantically meaningful concepts, and identified different hot topics of interest among the posts of online health communities related to Lyme Disease. The valuable conclusions would suggest that social media analysis is an influential source of knowledge for the comprehensive analysis of a specific disease. Directions for future work are many. One direction would be to identify the different stakeholder groups in more detail. It may be possible to identify different types of individuals or patients: for example, to distinguish between chronic sufferers with a confirmed diagnosis, those with an acute issue, and those seeking information who do not know if they have the condition. The 'expert patient' is one who often provides a lot of information to others online and is an interesting phenomenon in its own right [125]. It may also be possible to separate the concerns of family members and carers from those of the patients themselves.

Addressing areas such as inaccurate information and bias are other avenues of potential future work. Inaccurate information could be identified if it diverged significantly from established medical opinion. Various biases, be they socio-economic, age, race, etc., are inherent in any social network, as not all sufferers can or will participate in online forums. Methods to reduce bias should lead to more accurate and informative analyses.

The next and final chapter will conclude the findings and summarize the contributions presented in this thesis. Moreover, the chapter includes several possibilities for further expanding the work in this thesis that could be considered in future work.

The content of this chapter was included in a research paper [121] (under review).

Chapter 10

Conclusion and Future Work

This thesis addresses the classification of socially-generated data, in particular through the use of *sentiment analysis*, in which we classify user-contributed content with a range of sentiments (or affects) that might best characterize the online contributions. All classification developed in this thesis uses a feature-based model. We explored fully supervised learning approaches with short-form corpora classification (e.g., Twitter data) in Chapter 4 and long-form corpora in the form of online medical forums in Chapters 5, 6 and 7. To address the issue of the shortage of labelled data, which is a common problem in data analytics, we then amended the approach using an adapted semi-supervised learning algorithm for the long-form corpora. Finally, to exploit different aspects of the data, we adopted complementary analysis approaches, viz. topic analysis and content analysis.

This chapter summarizes the main findings, and points to possible directions for future work.

10.1 Main Conclusions of the Empirical Investigations

The summary of the main conclusions, findings and contributions can be stated as follows with regard to each of the research questions investigated.

RQ1: Can meta-level features improve binary sentiment analysis performance on short-form socially generated text, Twitter in particular?

In Chapter 4, Twitter was used for a sentiment analysis investigation. We

proposed an ensemble learning approach based on the meta-level features of seven existing lexicon resources for automated polarity sentiment classification. The ensemble employed four base learners (Two-Class Support Vector Machine, Two-Class Bayes Point Machine, Two-Class Logistic Regression, and Two-Class Decision Forest) for the classification task. Meta-level features were used with each of the base learners in the ensemble and three different labelled Twitter datasets were used to evaluate the effectiveness of this approach to sentiment analysis. Our experiments show that using meta-level features mitigated the problems associated with Twitter data, resulting in increased performance for the binary classification problem. The results suggest that the use of meta-level features outperformed the baseline by 2-14% for binary classification in accuracy and F score. The results indicate that, based on a combination of existing lexicon resources, the ensemble learners minimized the error rate by avoiding the poor selection of stand-alone classifiers.

RQ2: Can the meta-level feature, in conjunction with conventional features, improve multi-class sentiment analysis performance on more content-rich Lyme Disease medical forums?

For this, we extended the short-form binary classification to long-form multiclass classification. We considered medical posts related to Lyme Disease (Chapter 5) – and subsequently Lupus (Chapter 6) – from a cross-section of relevant medical forums. We analysed the data and identified and evaluated a set of categories that comprehensively covered the range of sentiments or affects expressed in specific medical discussions. We conducted our medical sentiment analysis as a multi-class classification problem to classify each user-generated post into one of the following classes: Depressed and frustrated, Lyme infection confusion, Lyme symptom confusion, Asking about treatment, Awareness and encouragement, and Seeking general information.

We then proposed a feature-based model for the classification of a Lyme Disease dataset. These features included a combination of three different feature sets: content-free features, meta-level features and content-specific features. Our experiments were conducted in an incremental manner using Multi-class Neural Network and Multi-class Logistic Regression classifiers and evaluated on high-quality annotated data to assess the feasibility and accuracy of our automated classification. Our experiment results show that multiple features used with a feature selection technique can maximize the performance of classifiers by

11-21% over the baseline using the hold-out evaluation method[33]. In Chapter 7, we conducted further evaluation using the cross-validation evaluation method, which resulted in a performance increase over the baseline of 11-14% using χ^2 feature selection and 12-13% using Fisher score feature selection[34]. The (Baseline + Content-free) and (Baseline + Content-free + Meta-level) feature sets outperformed the baseline in terms of accuracy and the Kappa statistic. The full feature-based model significantly outperformed all the proposed combinations of features in accuracy and Kappa. Overall, the experimental results demonstrate the effectiveness of our approach.

RQ3: Can the previous feature-based model adapt to an online medical forum discussing a different disease (i.e., Lupus)?

We further evaluated the feature-based model by assessing its ability to adapt to an online medical forum discussing Lupus which has similar characteristics to Lyme Disease (Chapter 6). As before, we identified and evaluated sentiments expressed in medical discussion forums related to Lupus. We observed that the categories identified were identical to the categories identified from the Lyme Disease dataset. Therefore, we conclude that the categories developed for the Lyme Disease dataset may also be adequate for data on other diseases.

We investigated the proposed feature-based model by experimenting with inductive learning algorithms to build a feature-based classification model. The feature-based model outperformed the baseline by 6-20% in medical posts related to Lupus using the hold-out evaluation method[35]. Those investigations led to further experiments using a cross-validation evaluation method (Chapter 7). The results indicated that performance increased over the baseline by 7-20% using χ^2 feature selection and by 8-18% using Fisher score feature selection[36]. When assessing the three feature sets, we found that Baseline combined with Content-free and Meta-level feature set outperform the performance significantly compared to the Baseline combined with only Content-free feature set using the Logistic Regression in terms of accuracy and Kappa, while this was not the case using Neural Networks. However, in both classifiers, the full feature-based model outperformed the baseline and (Baseline + Content-free + Meta-level) in accuracy and Kappa. We conclude that the feature-based multi-class classification model can be relevant to the analysis of other medical domains, in addition

[33]Referring to an average improvement in both classifiers.
[34]Referring to an average improvement in both classifiers.
[35]Referring to an average improvement in both classifiers.
[36]Referring to an average improvement in both classifiers.

to those concerned with Lyme Disease.

RQ4: Can the feature-based model be adapted for semi-supervised learning?

To address the issue of scarcity and the high cost of labelled data, particularly for medical-related discussions, we investigated a semi-supervised approach to multi-class sentiment classification in medical and health-related discourse in Chapter 8. We investigated the use of a semi-supervised learning technique, co-training, adapted for the feature-based model. In the adapted co-training, the features were divided into two distinct sets of views: Domain Independent and Domain Dependent.

We evaluated the approach using the Lyme Disease and Lupus datasets and compared the performance against the baseline and the initial co-training using a simple logistic classifier. Adaptive co-training outperformed the accuracy of each of the corresponding baselines trained using 20% and 25% of the training datasets by 11.1% and 7.4% respectively for the Lyme Disease dataset and by 7.8% and 4.6% for the Lupus dataset. As against the fully supervised model, the results were quite competitive: the fully supervised model trained on a labelled dataset scored accuracy only 2.8-4.7% higher than the adaptive model, which used substantially fewer labelled data for both datasets. The results of our experiments demonstrate the effectiveness of the approach.

RQ5: Can medical forums discourse be analysed by different methods to be useful in other ways?

In addition to sentiment analysis, which was explored in Chapters 5, 6, 7 and 8, we believe that alternative analytic methods can address the needs of a variety of healthcare stakeholders, who might be interested in differing aspects of the information. In Chapter 9, we performed content analysis to extract multiple concepts and themes in an attempt to obtain a general view of the medical content in online discussion. We further experimented with topic analysis as another text analysis method and used it to extract hot topics represented in socially generated medical discussions related to a specific disease.

In applying content analysis, we intended to map the concepts of most concern to users to domain-specific semantic types, which could be used to address the terminology gap that exists between health professionals and health consumers. For topic analysis, we extracted and dicussed the main topics mentioned in online medical posts related to Lyme Disease in order to offer the potential for medical professionals and researchers to better understand participants' current

concerns and hot topics.

10.2 Future work

The experiments and results outlined in this thesis show that the approaches taken enhance classification performance of socially generated text. Nonetheless, there are several possible directions for future work. In this section, we briefly point out some such directions.

10.2.1 Generalizing the Feature-based Model to Other Diseases and Platforms

In this thesis, we started by classifying sentiments expressed in medical forums relevant to Lyme Disease, and then a second disease, Lupus, which shared similar characteristics to the first. Thus, our approach could be tailored to data discussing other diseases and to other social platforms, as long as they share many of the same characteristics. In future work, more experiments could be performed to investigate empirically the performance and applicability of our approach to other data.

10.2.2 Improving and Exploring Semi-supervised Learning Approaches

An essential requirement for pursuing this type of investigation (Section 10.2.1) is high-quality labelled data for evaluation purposes: the scarcity of such data for socially generated text can prove to be a limiting factor. Generating a larger amount of high-quality labelled data would, although challenging, be necessary for further investigation. Alternatively, an interesting future research direction is to explore semi-supervised learning approaches or hybrid learning approaches that can deal with shortages of labelled data. For instance, large scale experiments are important to demonstrate the applicability of the modified co-training approach to different types of data. Furthermore, combining co-training with active learning to select the most informative data to be labelled could improve the performance of the classifiers.

10.2.3 Conducting more Experiments on Extracted Features

The feature space in user posts in online health communities is another area that could be explored further. One example might be the incorporation of closely related features, such as utilising the Unified Medical Language System (UMLS). A second example might involve exploring methods of reducing the number of features required to define an optimal set of features for medical forums, and undertaking comprehensive analysis of feature spaces by using, for example, dimensionality reduction and investigating the interaction and correlation of feature sets.

Investigating contextualised word embeddings is another approach that might be beneficial toward the text representation of semantic aspects of text data. Word embedding involves assigning a real-valued vector to each word according to the context in which it appears and then representing a single word of vocabulary as a point in a vector space. For instance, Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained word-embedding approach [139] that was trained on book data that include 800M words. Capturing the semantic or syntactical value of words based on their context could improve the performance of text classification.

10.2.4 Data Visualisation

Throughout the thesis, we have presented and justified several analytic techniques for the classification of data. For the most part, the system user is expected to invest effort in comprehending the classes generated, and is provided with little help in achieving this. Adding a data visualisation component to a classification that can summarise classes, and depict their inter-relationships, would be an extremely useful tool.

References

[1] C.-W. Hsu, C.-C. Chang, C.-J. Lin, *et al.*, "A practical guide to support vector classification," 2003.

[2] J. Brownlee, "Logistic regression for machine learning," *Link: https://machinelearningmastery. com/logistic-regression-for-machine-learning/. Udgivet*, vol. 1, 2016.

[3] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[4] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[5] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[6] K. Denecke, *Health Web Science: Social Media Data for Healthcare*. Springer, 2015.

[7] A. Montoyo, P. MartíNez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments," *Decision Support Systems*, vol. 53, no. 4, pp. 675–679, 2012.

[8] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 101–111, 2014.

[9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[10] A. Bermingham and A. F. Smeaton, "Classifying sentiment in microblogs: is brevity an advantage?," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1833–1836, ACM, 2010.

[11] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pp. 30–38, 2011.

[12] A. Abbasi and H. Chen, "Affect intensity analysis of dark web forums," in *2007 IEEE Intelligence and Security Informatics*, pp. 282–288, IEEE, 2007.

[13] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[14] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[15] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 492–499, IEEE Computer Society, 2010.

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

[17] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky, "Automatic extraction of opinion propositions and their holders," in *2004 AAAI spring symposium on exploring attitude and affect in text*, vol. 2224, 2004.

[18] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[19] F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer, "Determining word-emotion associations from tweets by multi-label classification," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 536–539, IEEE, 2016.

[20] P. Biyani, S. Bhatia, C. Caragea, and P. Mitra, "Using non-lexical features for identifying factual and opinionative threads in online forums," *Knowledge-Based Systems*, vol. 69, pp. 170–178, 2014.

[21] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Meta-level sentiment models for big social data analysis," *Knowledge-Based Systems*, vol. 69, pp. 86–99, 2014.

[22] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREc*, vol. 10, pp. 1320–1326, 2010.

[23] N. Yu, "Exploring c o-training strategies for opinion detection," *Journal of the Association for Information Science and Technology*, vol. 65, no. 10, pp. 2098–2110, 2014.

[24] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen, "Can i hear you? sentiment analysis on medical forums," in *Proceedings of the sixth international joint conference on natural language processing*, pp. 667–673, 2013.

[25] V. Hatzivassiloglou and J. M. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pp. 299–305, Association for Computational Linguistics, 2000.

[26] S. Melzi, A. Abdaoui, J. Azé, S. Bringay, P. Poncelet, and F. Galtier, "Patient's rationale: Patient knowledge retrieval from health forums," in *eTELEMED: eHealth, Telemedicine, and Social Medicine*, 2014.

[27] R. D. Waters and J. M. Williams, "Squawking, tweeting, cooing, and hooting: Analyzing the communication patterns of government agencies on twitter," *Journal of Public Affairs*, vol. 11, no. 4, pp. 353–363, 2011.

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

[29] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in *Fifth International AAAI conference on weblogs and social media*, 2011.

[30] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, p. 28, 2016.

[31] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," in *Extended Semantic Web Conference*, pp. 88–99, Springer, 2011.

[32] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[33] J. Brownlee, "Discover feature engineering, how to engineer features and how to get good at it," *Machine Learning Process*, 2014.

[34] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, pp. 415–463, Springer, 2012.

[35] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 12, 2008.

[36] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[37] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American society for information science and technology*, vol. 57, no. 3, pp. 378–393, 2006.

[38] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 67–75, 2005.

[39] J. Li, R. Zheng, and H. Chen, "From fingerprint to writeprint," *Communications of the ACM*, vol. 49, no. 4, pp. 76–82, 2006.

[40] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: An experiment on online product reviews," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 46–53, 2009.

[41] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 3, pp. 210–229, July 1959.

[42] V. Bobicev, M. Sokolova, and M. Oakes, "What goes around comes around: learning sentiments in online medical forums," *Cognitive Computation*, vol. 7, no. 5, pp. 609–621, 2015.

[43] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth, "What are people tweeting about zika? an exploratory study concerning its symptoms, treatment, transmission, and prevention," *JMIR public health and surveillance*, vol. 3, no. 2, p. e38, 2017.

[44] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," *arXiv preprint arXiv:1308.6242*, 2013.

[45] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL student research workshop*, pp. 43–48, Association for Computational Linguistics, 2005.

[46] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271, Association for Computational Linguistics, 2004.

[47] Y. Lu, "Automatic topic identification of health-related messages in online health community using text classification," *SpringerPlus*, vol. 2, no. 1, p. 309, 2013.

[48] T. Zhang, J. H. Cho, and C. Zhai, "Understanding user intents in online health forums," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 220–229, ACM, 2014.

[49] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in *Twenty-sixth aAAI conference on artificial intelligence*, 2012.

[50] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[51] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, Citeseer, 1998.

[52] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[53] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Systems with Applications*, vol. 116, pp. 209–226, 2019.

[54] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Investigating dynamic routing in tree-structured lstm for sentiment analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3423–3428, 2019.

[55] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on twitter with hybrid cnn and rnn models," in *Proceedings of the 9th International Conference on Social Media and Society*, pp. 226–230, 2018.

[56] P. Balage Filho and T. Pardo, "Nilc_usp: A hybrid system for sentiment analysis in twitter messages," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 568–572, 2013.

[57] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, ACM, 2004.

[58] H. Saif, M. Fernandez, Y. He, and H. Alani, "Senticircles for contextual and conceptual semantic sentiment analysis of twitter," in *European Semantic Web Conference*, pp. 83–98, Springer, 2014.

[59] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424, Association for Computational Linguistics, 2002.

[60] D. Kang and Y. Park, "based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1041–1050, 2014.

[61] P. Pugsee, V. Nussiri, and W. Kittirungruang, "Opinion mining for skin care products on twitter," in *International Conference on Soft Computing in Data Science*, pp. 261–271, Springer, 2018.

[62] P. Pugsee, P. Sombatsri, and R. Juntiwakul, "Satisfactory analysis for cosmetic product review comments," in *Proceedings of the 2017 International Conference on Data Mining, Communications and Information Technology*, p. 13, ACM, 2017.

[63] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from twitter data," *Social Science Computer Review*, vol. 31, no. 6, pp. 649–679, 2013.

[64] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Fourth international AAAI conference on weblogs and social media*, 2010.

[65] D. Gayo Avello, "Don't turn social media into another'literary digest'poll," *Communications of the ACM, 54 (10)*, 2011.

[66] J. E. Chung and E. Mustafaraj, "Can collective sentiment expressed on twitter predict political elections?," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[67] D. Gayo-Avello, "No, you cannot predict elections with twitter," *IEEE Internet Computing*, vol. 16, no. 6, pp. 91–94, 2012.

[68] A. Jungherr, P. Jürgens, and H. Schoen, "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment"," *Social science computer review*, vol. 30, no. 2, pp. 229–234, 2012.

[69] P. Biyani, C. Caragea, P. Mitra, C. Zhou, J. Yen, G. E. Greer, and K. Portier, "Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 413–417, IEEE, 2013.

[70] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," in

Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 201–213, Springer, 2013.

[71] J. Shen, P. Zhu, R. Fan, W. Tan, and X. Zhan, "Sentiment analysis based on user tags for traditional chinese medicine in weibo," in *Natural Language Processing and Chinese Computing*, pp. 134–145, Springer, 2015.

[72] K. Denecke and W. Nejdl, "How valuable is medical social media data? content analysis of the medical web," *Information Sciences*, vol. 179, no. 12, pp. 1870–1880, 2009.

[73] M. Z. Asghar, A. Khan, F. M. Kundi, M. Qasim, F. Khan, R. Ullah, and I. U. Nawaz, "Medical opinion lexicon: an incremental model for mining health reviews," *International Journal of Academic Research*, vol. 6, no. 1, pp. 295–302, 2014.

[74] J.-C. Na, W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, and Y.-L. Theng, "Sentiment classification of drug reviews using a rule-based linguistic approach," in *International conference on asian digital libraries*, pp. 189–198, Springer, 2012.

[75] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Use of sentiment analysis for capturing patient experience from free-text comments posted online," *Journal of medical Internet research*, vol. 15, no. 11, p. e239, 2013.

[76] F. Greaves, A. A. Laverty, D. R. Cano, K. Moilanen, S. Pulman, A. Darzi, and C. Millett, "Tweets about hospital quality: a mixed methods study," *BMJ Qual Saf*, vol. 23, no. 10, pp. 838–846, 2014.

[77] R. Alnashwan, A. P. O'Riordan, H. Sorensen, and C. Hoare, "Improving sentiment analysis through ensemble learning of meta-level features," in *KDWEB 2016: 2nd International Workshop on Knowledge Discovery on the Web*, Sun SITE Central Europe (CEUR)/RWTH Aachen University, 2016.

[78] J. Langford, L. Li, and A. Strehl, "Vowpal wabbit online learning project," 2007.

[79] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," in *Lrec*, vol. 10, pp. 2200–2204, 2010.

[80] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining.," in *LREC*, vol. 6, pp. 417–422, Citeseer, 2006.

[81] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," tech. rep., Citeseer, 1999.

[82] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[83] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 163–173, 2012.

[84] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

[85] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[86] R. Herbrich, T. Graepel, and C. Campbell, "Bayes point machines," *Journal of Machine Learning Research*, vol. 1, no. Aug, pp. 245–279, 2001.

[87] G. Andrew and J. Gao, "Scalable training of l 1-regularized log-linear models," in *Proceedings of the 24th international conference on Machine learning*, pp. 33–40, ACM, 2007.

[88] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning," *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, vol. 5, no. 6, p. 12, 2011.

[89] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[90] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka Jr, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170–179, 2014.

[91] A.-R. Ko, R. Sabourin, and A. de Souza Britto, "Combining diversity and classification accuracy for ensemble selection in random subspaces," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pp. 2144–2151, IEEE, 2006.

[92] Y. Su, Y. Zhang, D. Ji, Y. Wang, and H. Wu, "Ensemble learning for sentiment classification," in *Workshop on Chinese Lexical Semantics*, pp. 84–93, Springer, 2012.

[93] S. Clark and R. Wicentwoski, "Swatcs: Combining simple classifiers with estimated accuracy," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 425–429, 2013.

[94] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision support systems*, vol. 57, pp. 77–93, 2014.

[95] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Applied Soft Computing*, vol. 50, pp. 135–141, 2017.

[96] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," *Proceedings of SemEval*, pp. 1–18, 2016.

[97] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proceedings of the First workshop on Unsupervised Learning in NLP*, pp. 53–63, Association for Computational Linguistics, 2011.

[98] M. Bihis and S. Roychowdhury, "A generalized flow for multi-class and binary classification tasks: An azure ml approach," in *2015 IEEE International Conference on Big Data (Big Data)*, pp. 1728–1737, IEEE, 2015.

[99] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[100] R. Alnashwan, H. Sorensen, A. O'Riordan, and C. Hoare, "Multiclass sentiment classification of online health forums using both domain-independent and domain-specific features," in *Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pp. 75–83, ACM, 2017.

[101] R. Alnashwan, H. Sorensen, A. O'Riordan, and C. Hoare, "Accurate classification of socially generated medical discourse," *International Journal of Data Science and Analytics*, pp. 1–13, 2018.

[102] S. Fox, "The social life of health information. pew internet & american life project 2013," *17. https://www.pewinternet.org/2013/01/15/health-online-2013/*, 2013.

[103] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis," in *Twenty-Fifth International FLAIRS Conference*, 2012.

[104] J. Staiano and M. Guerini, "Depechemood: a lexicon for emotion analysis from crowd-annotated news," *arXiv preprint arXiv:1405.1605*, 2014.

[105] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[106] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[107] B. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," *IEEE Transactions on Systems, MAN, and Cybernetics-part a: Systems and Humans*, vol. 39, no. 1, pp. 36–46, 2008.

[108] T. R. Nichols, P. M. Wisner, G. Cripe, and L. Gulabchand, "Putting the kappa statistic to use," *The Quality Assurance Journal*, vol. 13, no. 3-4, pp. 57–61, 2010.

[109] M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, *et al.*, "Digital epidemiology," *PLoS computational biology*, vol. 8, no. 7, p. e1002616, 2012.

[110] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[111] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[112] C. Nadeau and Y. Bengio, "Inference for the generalization error," in *Advances in neural information processing systems*, pp. 307–313, 2000.

[113] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[114] R. Alnashwan, H. Sorensen, and A. O'Riordan, "Classification of online medical discourse by modified co-training," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, 2019.

[115] X. Ding and B. Liu, "The utility of linguistic rules in opinion mining," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 811–812, ACM, 2007.

[116] S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, "Employing personal/impersonal views in supervised and semi-supervised sentiment classification," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 414–423, Association for Computational Linguistics, 2010.

[117] Y. Zhou and S. Goldman, "Democratic co-learning," in *16th IEEE International Conference on Tools with Artificial Intelligence*, pp. 594–602, IEEE, 2004.

[118] M.-F. Balcan, A. Blum, and K. Yang, "Co-training and expansion: Towards bridging theory and practice," in *Advances in neural information processing systems*, pp. 89–96, 2005.

[119] M. Collins and Y. Singer, "Unsupervised models for named entity classification," in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[120] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[121] R. Alnashwan, H. Sorensen, and A. O'Riordan, "Medical social data mining and stakeholder perspectives,"

[122] H. A. Schwartz and L. H. Ungar, "Data-driven content analysis of social media: a systematic overview of automated methods," *The ANNALS of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 78–94, 2015.

[123] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[124] S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving, "A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication," *Journal of medical Internet research*, vol. 15, no. 4, p. e85, 2013.

[125] T. Greenhalgh, "Patient and public involvement in chronic illness: beyond the expert patient," *Bmj*, vol. 338, p. b49, 2009.

[126] R. G. Rodrigues, R. M. das Dores, C. G. Camilo-Junior, and T. C. Rosa, "Sentihealth-cancer: a sentiment analysis tool to help detecting mood of patients in online social networks," *International journal of medical informatics*, vol. 85, no. 1, pp. 80–95, 2016.

[127] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection—harnessing the web for public health surveillance," *New England Journal of Medicine*, vol. 360, no. 21, pp. 2153–2157, 2009.

[128] P. A. Ubel, C. Jepson, and J. Baron, "The inclusion of patient testimonials in decision aids: effects on treatment choices," *Medical Decision Making*, vol. 21, no. 1, pp. 60–68, 2001.

[129] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Harnessing the cloud of patient experience: using social media to detect poor quality healthcare," *BMJ Qual Saf*, vol. 22, no. 3, pp. 251–255, 2013.

[130] Y. Lu, Y. Wu, J. Liu, J. Li, and P. Zhang, "Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community," *Journal of medical Internet research*, vol. 19, no. 4, p. e109, 2017.

[131] S. Stemler, "An overview of content analysis," *Practical assessment, research & evaluation*, vol. 7, no. 17, pp. 137–146, 2001.

[132] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program.," in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.

[133] Y. Zhan, R. Liu, Q. Li, S. J. Leischow, and D. D. Zeng, "Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms," *Journal of medical Internet research*, vol. 19, no. 1, p. e24, 2017.

[134] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, ACM, 2015.

[135] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf, "Quantifying information overload in social media and its impact on social contagions," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[136] R. S. Wurman and R. S. Wurman, *Information anxiety: What to do when information doesn't tell you what you need to know*. Bantam New York, 1990.

[137] C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang, "Sentiview: Sentiment analysis and visualization for internet popular topics," *IEEE transactions on human-machine systems*, vol. 43, no. 6, pp. 620–630, 2013.

[138] A. M. Hopper and M. Uriyo, "Using sentiment analysis to review patient satisfaction data located on the internet," *Journal of health organization and management*, vol. 29, no. 2, pp. 221–233, 2015.

[139] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[140] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics*, p. 1367, Association for Computational Linguistics, 2004.

[141] B. Liu, "Opinion mining and summarization," in *Tutorial at the World Wide Web Conference (WWW), Beijing, China*, 2008.

[142] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[143] J. Short, E. Williams, and B. Christie, *The social psychology of telecommunications*. John Wiley & Sons, 1976.

[144] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[145] K. J. Petrie and J. Weinman, *Perceptions of health and illness: Current research and applications*. Taylor & Francis, 1997.

[146] K. P. Davison, J. W. Pennebaker, and S. S. Dickerson, "Who talks? the social psychology of illness support groups.," *American Psychologist*, vol. 55, no. 2, p. 205, 2000.

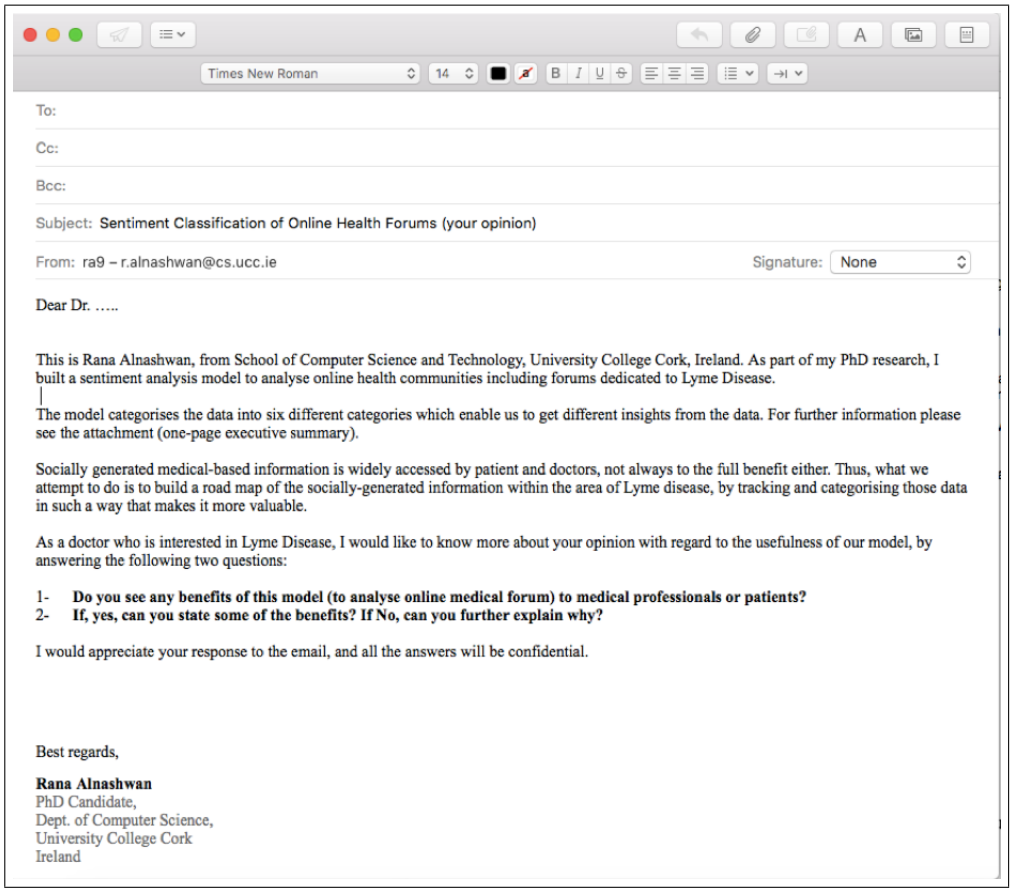
[147] S. Bhatia and P. Mitra, "Adopting inference networks for online thread retrieval," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[148] R. L. Daft and R. H. Lengel, "Organizational information requirements, media richness and structural design," *Management science*, vol. 32, no. 5, pp. 554–571, 1986.

Appendix A

Expert Opinion

In this appendix we send an email that includes open-ended questions. The aim was to consulted expert opinion as to the usefulness of the proposed sentiment analysis classification model for Lyme Disease posts (i.e., as one case of analysing medical user-generated discourse).



To:

Cc:

Bcc:

Subject: Sentiment Classification of Online Health Forums (your opinion)

From: ra9 – r.alnashwan@cs.ucc.ie Signature: None

Dear Dr. …..

This is Rana Alnashwan, from School of Computer Science and Technology, University College Cork, Ireland. As part of my PhD research, I built a sentiment analysis model to analyse online health communities including forums dedicated to Lyme Disease.

The model categorises the data into six different categories which enable us to get different insights from the data. For further information please see the attachment (one-page executive summary).

Socially generated medical-based information is widely accessed by patient and doctors, not always to the full benefit either. Thus, what we attempt to do is to build a road map of the socially-generated information within the area of Lyme disease, by tracking and categorising those data in such a way that makes it more valuable.

As a doctor who is interested in Lyme Disease, I would like to know more about your opinion with regard to the usefulness of our model, by answering the following two questions:

1- **Do you see any benefits of this model (to analyse online medical forum) to medical professionals or patients?**
2- **If, yes, can you state some of the benefits? If No, can you further explain why?**

I would appreciate your response to the email, and all the answers will be confidential.

Best regards,

Rana Alnashwan
PhD Candidate,
Dept. of Computer Science,
University College Cork
Ireland

Figure A.1: Email survey for expert opinion