The Pennsylvania State University

The Graduate School

College of Information Sciences and Technology

USING SUPERVISED LEARNING TO IDENTIFY DESCRIPTIONS OF PERSONAL EXPERIENCES RELATED TO CHRONIC DISEASE ON SOCIAL MEDIA

A Thesis in

Information Sciences and Technology

by

William Murphy

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2014

The thesis of William Murphy was reviewed and approved* by the following:

John Yen
Professor of Information Sciences and Technology
Thesis Supervisor

Prasenjit Mitra
Associate Professor of Information Sciences and Technology

Lee Giles
Professor of Information Sciences and Technology

Peter Forester
Senior Lecturer, Security Risk and Analysis
Graduate Program Administrator

*Signatures are on file in the Graduate School.

ABSTRACT

Patients are increasingly turning to online communities for health information and emotional support. In 2012, a study by the Pew Research Center found that more than 70% of Internet users in the United States, or 180 million adults, have searched the web for medical information [34]. According to the same study, 18% of Internet users have sought others online with similar medical conditions, and 3-4% have posted about their medical treatments [33]. Healthcare providers are also using the Internet to deliver various types of health interventions, including stress management courses, breast cancer coping groups, anti-smoking treatments, and weight loss therapy [6]. These trends have led to a surplus of patient data on the web, including patients' descriptions of their experiences of different ailments and the effects of treatment.

Sentiment analysis and social network analysis are powerful computational tools with which to make sense of this ever-growing corpus of medical data that is accumulating in online communities and social media. With sentiment classification algorithms, researchers can aggregate thousands or even millions of pieces of text to perform tasks such as predicting stock market movements [11], aggregating product reviews [87], and even gauging national mood [56]. These same methods can also be applied to healthcare to improve the quality of healthcare services. Some researchers are already advocating for more data mining in the healthcare domain, arguing that this will create a new "digital epidemiology" that will improve the healthcare system [110].

Nevertheless, there are significant technical challenges involved in mining social media data. This data is often difficult for text mining systems to parse due to its disorganized nature and the presence of slang, and developing useful features to accurately classify texts in this domain is an open problem. Additionally, before measuring the sentiment of online texts about healthcare, it is important to understand whether these messages represent attitudes or descriptions of personal experiences.

This thesis examines a relatively unexplored supervised machine learning task in the healthcare domain, automatic identification of social media messages pertaining to cancer-related personal experiences. We demonstrate that supervised learning methods can be used to accurately predict whether Twitter posts contain descriptions of personal experiences using four datasets of tweets about breast cancer, lung cancer, prostate cancer, and diabetes. Despite the unbalanced nature of this classification problem (of 4,821 labeled tweets, fewer than 20% of Twitter posts contain descriptions of personal experiences), these methods are able to classify with high F-Measure (>70%). We also show that content-based are more effective than context-based features.

This thesis also discusses novel data filtering techniques and natural language processing-based feature engineering methods that significantly improve classification of these short Twitter messages. These features take advantage of slang and other information that is typically ignored by text mining systems. Finally, this thesis demonstrates that this personal experience identification task is amenable to a transfer of learning approach, as knowledge about social media post content from one type of cancer can be transferred to another type of cancer or another type of chronic disease.

This technology has a number of applications in today's information-driven healthcare industry, including aggregating experiences with different treatments and medications, which could lead to more patient-centric delivery of healthcare.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge the faculty members who made this thesis possible, including my advisor Dr. John Yen, my undergraduate honors advisor Dr. Lisa Lenze, and my committee members, Dr. Prasenjit Mitra and Dr. Lee Giles. I would also like to acknowledge Yafei Wang and Emily Stang, who provided valuable guidance and assistance during the research process.

Chapter 1

Introduction

1.1 Health Information on Social Media

Many researchers in medical fields are looking forward to the development of data mining techniques for healthcare, arguing that this treasure trove of publically available data on the Internet may improve medicine in the long term [110]. Social media data has the potential to give healthcare professionals new ways of studying both treatment and diagnosis by allowing them to passively study the health of thousands or millions of patients. Social media data also has the potential to unlock new insights about how patients support each other emotionally via the Internet. Moreover, the prevalence of health social networks such as PatientsLikeMe [91] and the online health communities present on mainstream social networks such as Facebook and Twitter suggest that a large repository of data about public health will be publicly available for researchers to study.

Nevertheless, many questions remain about the effect of the Internet and online communities on patient outcomes, and in some cases the presence of medical misinformation on the web can be harmful [150]. Indeed, medical information on the web is often inaccurate, and according to a 2009 poll, 75% of users who look up medical information on the web do not check the source or validity of the information they find online [6]. The usefulness of web-based interventions has also been called into question [6]. Other scholars argue that the dynamics of online health communities are not fully understood, particularly their effect on patients' emotional states [145].

In addition, extracting information from social media is a difficult task even for modern data mining and machine learning approaches. Social media data is heterogeneous, and often includes open-form texts, links, images, and multimedia, which presents a challenge for machine learning methods designed for more homogenous data [33]. Social media posts are also embedded in a very particular context, which includes the online community they are a part of as well as the personal history of the poster. Finally, social media posts contain large amounts of "noise" and useless information as well as inaccurate information, which makes it difficult to extract meaningful knowledge or draw strong conclusions [33].

One largely unexplored open research problem in this area is experience mining, or the use of machine learning methods to automatically detect social media posts that contain descriptions of personal experiences. Currently, it is difficult or impossible to automatically detect whether a social media post is describing the a personal experience of the author, which makes it difficult to identify the frequency of real-world phenomena such as adverse drug effects (ADRs) using social media mining.

Despite these setbacks and open problems, in the last several years there has been a concerted effort to apply machine learning to the healthcare domain and study online interactions related to public health, most prominently using automated sentiment analysis. Sentiment analysis methods have been used to track public health, understand how patients respond emotionally to particular diseases, medications and symptoms, and even gauge the quality of the healthcare system. Other researchers have used these same methods to study health information outside of social media, such as suicide notes and academic opinions about the effectiveness of particular treatments. In addition, researchers in both computer science and the social sciences have been using machine learning methods to study how individuals interact in online health communities, including how patients support each other and share health information.

1.2 Twitter

Twitter is a popular microblogging service that allows users to post short text messages, or tweets, in real time [129]. A tweet is limited to 140 characters, and can include text, URLs linking to outside web pages, images, videos and other content [129]. Other users can reply or re-post these short messages to their own Twitter feeds. This action is referred to as retweeting; the duplicated tweet is referred to as a retweet. Twitter users can read and write posts by visiting the Twitter site via web browser or through a mobile application. This accessibility and ubiquity has made Twitter a popular platform for many forms of content such as news, personal blogging, sports, pop culture, and health information.

The figure below shows a tweet from the College of Information Sciences and Technology's official Twitter account and several retweets by other Twitter users.

Figure (1). A Twitter post and several retweets.

As a social media platform, Twitter is massively popular in the United States and around the world. Twitter has more than 800 million users [128], including 100 million users who are active daily. Twitter has published over 300 billion tweets, and approximately 500 million new tweets are created each day [128]. Twitter also has a large international audience, with 77% of Twitter users who are active monthly located outside of the United States [128]. This popularity makes Twitter an ideal platform to study online health information, as its large userbase allows researchers to study how individuals discuss their own personal health and how health-related information content travels through a social network.

However, Twitter's restricted format and emphasis on information sharing also makes Twitter difficult to extract data from—in particular, Twitter's 140 character limit poses unique challenges for machine learning researchers. As a result of this restriction, Twitter posts are significantly "messier" than other forms of writing, as posters frequently use slang and abbreviated words. These nonstandard words are more difficult to parse and more difficult for a learning algorithm to learn from, especially because their usage changes over time [33]. In addition, the 140-character limit restricts the amount of information that can be learned from each tweet, making it harder to train a text classification algorithm on a dataset of Twitter posts. For example, it is difficult to automatically classify the sentiment of Twitter posts, as there are only 140 characters worth of information for a machine learning algorithm to use to make this sentiment prediction. The abundance of retweets also creates a data filtering problem, as most tweets contain redundant information rather than new content. As a result, it is still very difficult to extract meaningful information from Twitter posts even in aggregate, and the challenges associated with learning from this messy data are still considered open research problems.

These challenges to automated classification make it difficult to understand the nuances of health-related content posted to Twitter, particularly whether this content reflects the personal experiences of real users. This is a major obstacle when attempting to leverage Twitter's large

community to study patients' emotional reactions to health conditions, as it is not always clear whether a Twitter post is discussing a personal experience with a disease or merely expressing an opinion. Though some research efforts in the area of experience mining have been effective on long-text formats such as blog posts, attempting to classify personal experiences on a short-text medium such as Twitter is an unexplored problem.

This thesis reviews existing research on extracting health information from social media, particularly on Twitter, and explores novel methods for feature extraction and data filtering designed to make the most of the unstructured nature of Twitter posts. In particular, we will discuss the use of natural language processing methods to filter a dataset to remove redundant data as well as novel methods of feature extraction designed to make use of Twitter's unique linguistic properties. These techniques are then applied to the problem of identifying personal experiences on Twitter, where we demonstrate that they can be effectively used to locate Twitter posts related to chronic disease.

Chapter 2

Background

2.1 Machine Learning

Machine learning, also known as statistical learning or data mining, is a collection of statistical and computational methods that can automatically identify patterns in a dataset and make predictions based on these patterns [41,77]. Machine learning techniques have been applied to a wide variety of domains, including email spam filtering, handwriting recognition, and image classification [41,77]. Machine learning methods are typically divided into two categories, supervised machine learning and unsupervised machine learning [41,77].

2.1.1 Supervised Learning

Supervised learning is the use of statistical inference to learn a hypothesis from a set of observed examples, and then applying this hypothesis to make predictions about a set of unobserved examples. The set of observed examples, called the training set, is a collection of records for which some output variable is known. The set of unobserved examples, called the testing set, is a collection of data with unknown output values. Consider the following example from Murphy (2012), which consists of a collection of shapes that have two classes:

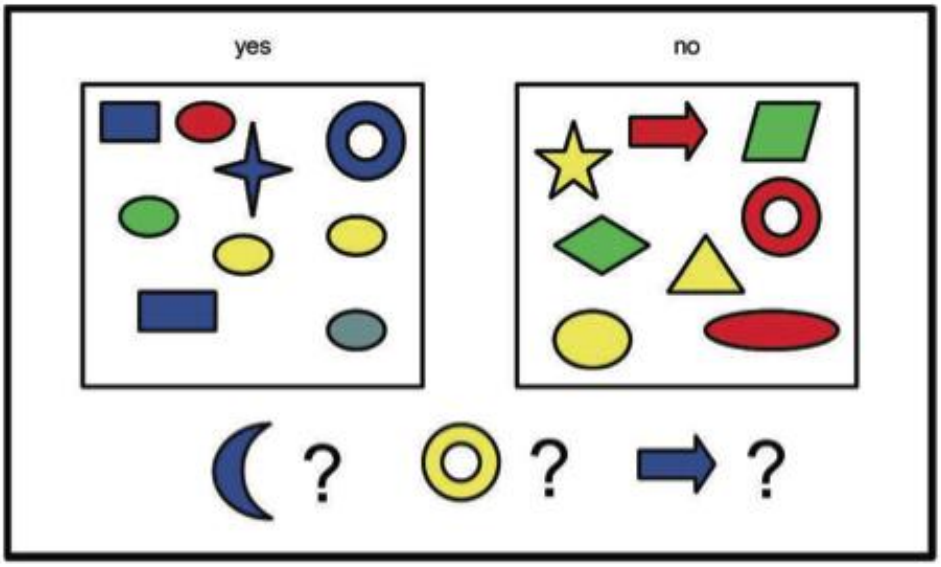

Figure (2). Example of a supervised learning problem (Murphy 2012).

In this diagram, we have a training set of shapes, some belonging to the "yes" class and some to the "no" class, and a testing set of three shapes of unknown class. The output variable in this example is the class—either "yes" or "no." Each shape can also be represented as a feature vector, or an ordered vector of the attributes of each shape, including the class:

Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

Figure (3). Feature vectors for example supervised learning problem (Murphy 2012).

A supervised learning algorithm can then perform statistical inference on this matrix of features vectors and learn some hypothesis about the shapes and the two classes, which can then be used to predict the classes of the testing set shapes. When the set of outcomes in a supervised learning problem is categorical or nominal, as in the example above, the problem is referred to as a classification problem. A simple example of a classification problem is image classification, in which there are a fixed number of classes of images, some training images with known class values, and a test set containing images of unknown class.

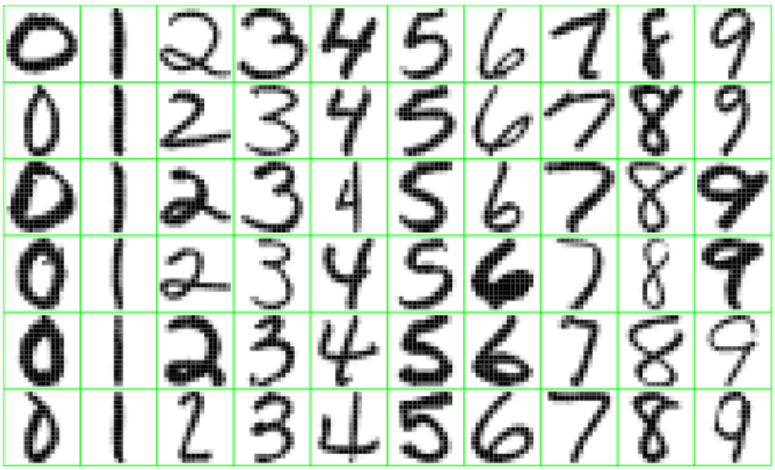

Figure (4). Handwritten digit recognition problem (Hastie 2004).

The diagram above, adapted from Hastie (2004), shows handwritten digits from U.S. postal envelopes. Digit recognition is a classic example of a supervised image classification problem—these examples could be used as the training set for a supervised image recognition classifier, which could then predict the class of new images of handwritten numerals [41]. When the output to a supervised learning problem is continuous, such as a real number value, the problem is referred to as a regression problem [41].

A typical example of training a supervised learning model is shown below:

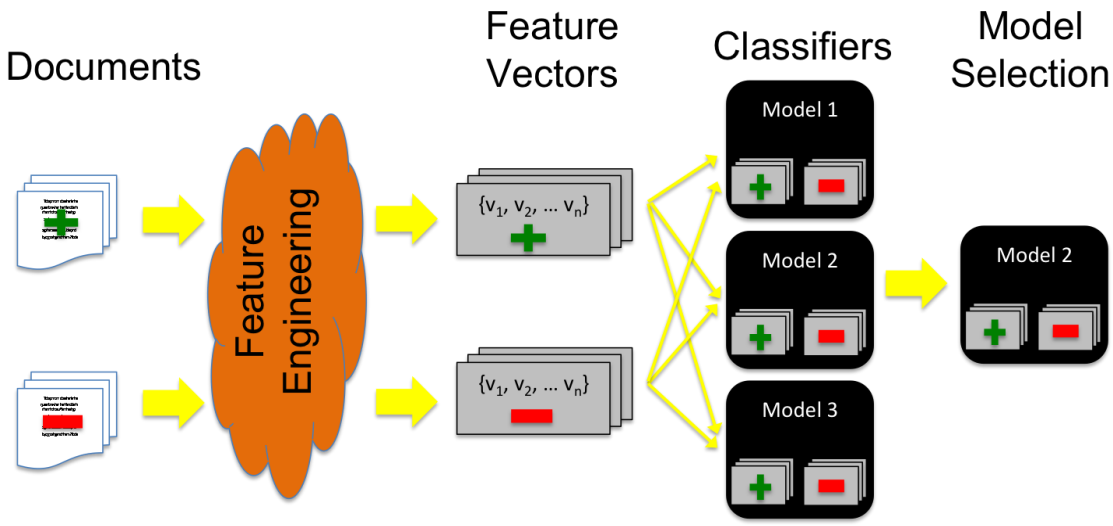

Figure (5). Process of training a supervised learning model.

The diagram above shows the process of training a supervised learning classifier. On the left, documents of two classes are transformed into feature vectors, and several classification algorithms are trained to learn about these features. The specific features used in the models depends on the process of feature engineering, in which the machine learning researchers use their domain knowledge to identify properties that are useful for classification. Finally, the models are compared by examining how well they can predict the class of items in the training set, and the best model is selected.

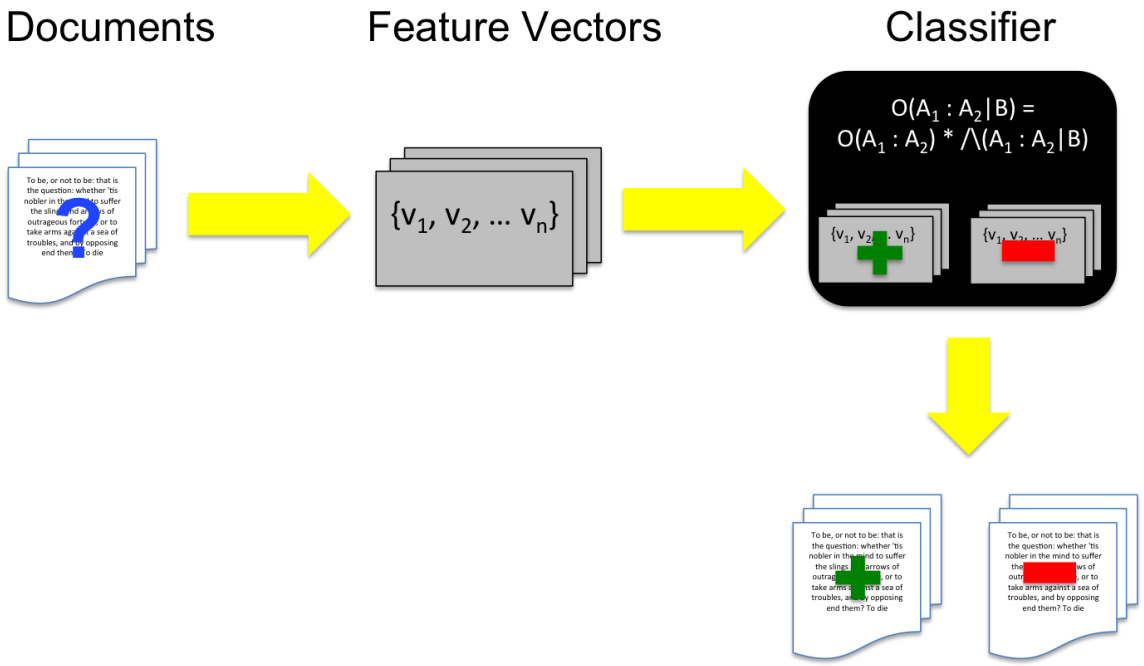

Figure (6). Process of applying a supervised learning model.

The diagram above shows the process of predicting the class of a collection of new documents using a trained model. The documents are transformed into their respective feature vectors, and the classifier uses these vectors to predict the class of each document.

2.1.2 Unsupervised Learning

Unsupervised learning, by contrast, is the use of statistical techniques to find interesting patterns in unlabeled data. One common unsupervised learning problem is clustering, in which a dataset must be grouped into some number of clusters based on the distribution of the data points [77]. For example, consider a dataset that contains information about individuals' heights and

weights. The diagrams below, from Murphy (2012) demonstrate how this dataset can be clustered to create groupings that may indicate an interesting pattern [77].

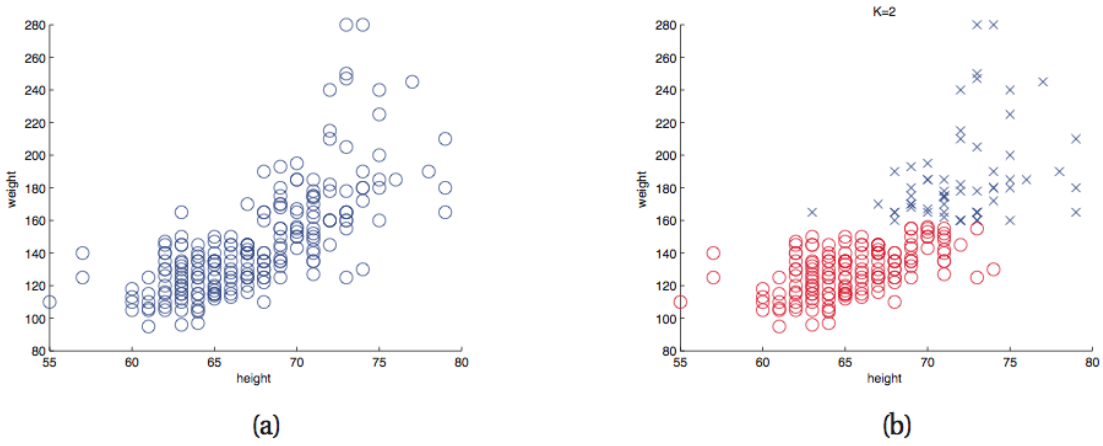

Figure (7). Example clustering problem (Murphy 2012).

Unsupervised learning is also referred to as "knowledge discovery," as there is no defined output in an unsupervised learning problem, only a collection of data that may contain some interesting or hidden structure [77]. Unlike supervised learning, predefined categories or classes are not used in unsupervised learning.

2.1.3 Text Mining

The application of supervised and unsupervised learning methods to text documents is known as text mining [77]. One common text mining problem is document classification, in which a document, such as an article or a web page, must be classified into a set of defined types [77]. One common example of document classification is email spam filtering, in which a classifier must predict whether an email is "spam" or "not spam" [77]. Similarly, in document

clustering, a set of documents is grouped into clusters based on some measure of similarity between the documents. A more advanced application of machine learning to text documents is probabilistic topic modeling, in which the topics of documents are modeled as a mixture model with some unknown parameters [77].

Text mining also includes natural language processing (NLP), which uses statistical methods to develop models of language and extract fine-grained information from text, such as parts-of-speech [26].

When applying machine learning methods to text documents, the documents are traditionally reduced to feature vectors using the bag of words model. In this model, a document is represented by the set of words it contains and the frequency of each word [77]. This is typically achieved by transforming a set of documents into a document-term matrix, in which each row represents a document and each column represents a word, and each cell in the matrix contains a value corresponding to the number of occurrences of that word in each document [77].

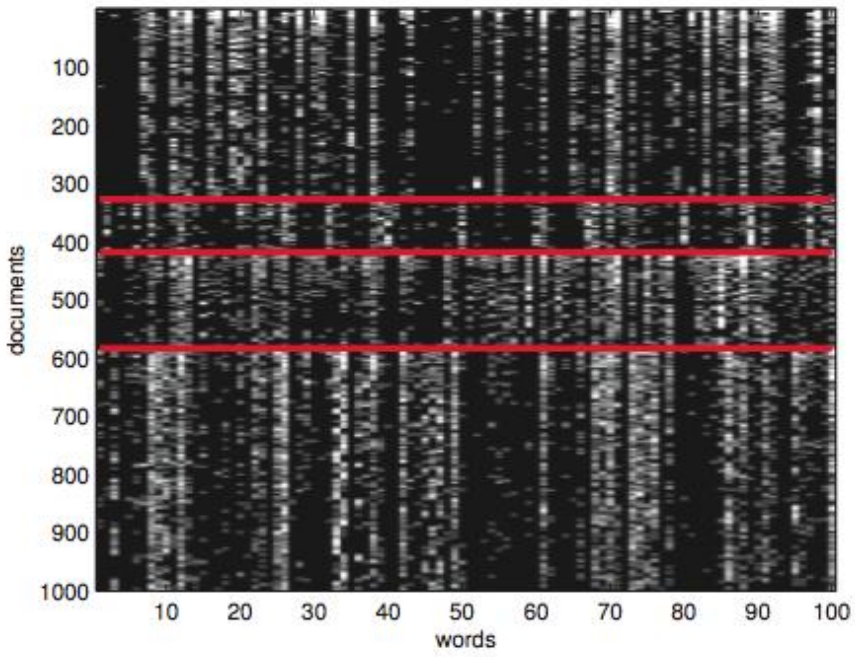

Figure (8). Example document-term matrix for a text classification problem (Murphy 2012).

The diagram above, from Murphy (2012), shows the document-term matrix for a collection of online forums; the red lines represent the boundaries between different forums [77]. However, word frequencies are far from the only features used in text classification tasks, and other features that incorporate the syntactic properties of documents, such as parts-of-speech, are also common [77].

2.2 Sentiment Analysis

Sentiment analysis is the use of computational methods to study the emotions, opinions, and attitudes expressed in text. Sentiment analysis is also referred to as sentiment mining, opinion mining, and subjectivity analysis [60,61,87]. The field combines methods from machine learning,

natural language processing, and affective computing, and it draws heavily from supervised learning methods [87].

2.2.1 Definition

Sentiment analysis consists of several computational tasks; the primary task is sentiment polarity classification. Sentiment polarity classification, also known simply as polarity classification, is a binary classification task in which a document must be labeled as expressing either a positive opinion or a negative opinion [87]. This task is also referred to as document-level sentiment classification, as it treats a document as a single unit of information [61]. Polarity classification is typically performed using supervised learning, wherein a set of documents with known sentiment values is used to train a classifier, which can then be used to predict the polarity of unlabeled documents [87]. Any classification algorithm can be used for this task, though Support Vector Machines and Naïve Bayes have traditionally been used in the literature [87]. Like other text mining fields, sentiment analysis makes use of NLP methods for feature extraction, primarily part-of-speech tagging and other methods of incorporating syntactic information into a feature set [87].

Though binary classification is commonly used for sentiment analysis, multi-class models for sentiment classification have also been used, most notably the Profile of Mood States model (POMS), which measures emotion using six categories: tension, depression, anger, vigor, fatigue, and confusion [12]. More recent research has attempted to ground sentiment analysis in psychological theories and existing models of emotion. For example, Bollen (2011), an often-cited sentiment analysis paper on stock market prediction, argues that, "[W]e stress the importance of measuring mood and emotion using well-established instruments rooted in decades of empirical psychometric research" [11]. Other authors have adapted models from other fields,

such as Lansdall-Welfare (2012), which used a four-emotion model from marketing research [56], and Quercia (2012), which combines a variety of emotional models into a general happiness metric [103]. According to Pang & Lee (2008) and Liu & Zhang (2012), commonly used features for document-level sentiment classification include the presence and frequency of individual words, parts of speech, the use of negation, the presence of affect-laden words and phrases (such as "good" or "hate"), and the syntax of text [61,87].

In sentiment strength detection, sentiment labels are treated as continuous values rather than discrete classes; this transforms the sentiment classification task into a regression task [61]. When applied to product reviews, this task is also referred to as rating inference, as the sentiment value being inferred represents the consumer's rating of a particular product [87]. However, polarity analysis and sentiment strength detection can be viewed as two sides of the same coin. Consider a model that outputs the probability that a document belongs to the "positive" class—this probability can be treated as a sentiment score for this document [61]. Using this probability as a continuous measure of sentiment, turns the two-class polarity classification problem into a regression problem, where the sentiment score is bounded by 0 and 1 [61]. Multi-class models such as POMS have been transformed into regression problems in this way, and the probability that a document belongs to each emotional category is treated as a measure of emotional content [12].

Sentence-level sentiment classification is the application of these same techniques to classify individual sentence rather than documents. A related problem is sentence subjectivity classification, in which sentences are classified into "objective" and "subjective" categories; an objective sentence contains only factual information, whereas a subjective sentence contains sentiment or opinion [61]. In sentence-level sentiment analysis, sentences are typically classified by subjectivity first, and then the subjective documents are classified by polarity [61].

As sentiment analysis aims to extract opinions about particular products, persons, or events, entity and attribute extraction methods are also a part of sentiment analysis research. The primary entity extraction task in sentiment analysis is referred to as named entity recognition (NER), which is the task of identifying entities, opinion holders, and the time at which an opinion was expressed [60,61]. Entity recognition methods led to the development of aspect-based sentiment analysis, in which opinions about the features of two entities, typically consumer goods, are summarized using polarity classification [61]. For example, Liu & Zhang (2012) described how two cellphones might be compared using aspect-based sentiment analysis: The two products would be reduced to their features, such as their voice quality and battery life, and then the number of positive and negative opinions about these features are extracted from product reviews [61].

Other areas of sentiment analysis research focus on problems related to sentiment classification or extensions of the methods discussed above. Opinion spam detection is the problem of identifying bogus or fake opinions, such as advertisements disguised as product reviews [60]. Opinion spam detection is an important research issue for sentiment analysis on social media, as social media platforms can contain large volumes of spam or unrelated text [86,92]. In the Twitter domain, researchers have attempted to remove opinion spam by ignoring social media posts containing certain keywords, ignoring posts that contain URLs, or using supervised machine learning to classify posts as spam or non-spam [86,92,109].

Another growing research area is the identification and mining of comparative opinions, which are opinions that compare the features of two entities, such as the reception on two different cellphones [60]. Mining comparative opinions consists of three tasks: The identification of a comparative opinion phrase, extracting the objects and object features from the text, and identifying the preferred feature or object [60]. Finally, cross-lingual sentiment analysis is the application of sentiment mining methods to non-English texts, and also includes the task of

learning from a corpus in one language and making predictions about a corpus in another language.

2.2.2 History: Supervised Sentiment Classification

Automated sentiment classification began with attempts to determine stock market movements from discussion on web forums and extract consumers' opinions about products and services from the Internet review sites [87]. These early attempts at sentiment mining used rudimentary sentiment analysis algorithms, such as searching for affect-laden keywords [87]. An early seminal work in the field is Turney (2002), which attempted to aggregate reviews of automobiles, banks, movies, and travel destinations by categorizing each review as "thumbs up" or "thumbs down" [126]. The algorithm used to classify reviews was relatively simple: Each review was scanned to identify adjectives and adverbs, and each adverb or adjective was assigned a semantic content value (a positive or negative number) based on its synonymy to other words with known affect values. These scores were then averaged to yield an overall score for the review, allowing it to be classified as either positive or negative [126]. Despite the simplicity of this approach, Turney's algorithm achieved accuracy as high as 84% for automobile reviews and 66% for movie reviews [126]. Another early and independent attempt at sentiment analysis was Dini (2002), which outlined an opinion classification system based on a semantic framework that assigned sentiment values to individual words and phrases [32]. Other researchers devised similar naïve algorithms for sentiment analysis in 2002-2003, and the convergence of these research efforts led to the creation of sentiment analysis as a field [87].

Pang and Lee (2002) expanded on this work and performed sentiment analysis on movie reviews using the bag-of-words model and three machine learning algorithms: Naïve Bayes, maximum entropy classification, and Support Vector Machines [88]. This research by Pang and

Lee was the first to employ machine learning algorithms for sentiment classification, and this innovation greatly influenced the field—after 2003, sentiment analysis research abandoned the semantic word-scoring algorithms that characterized earlier research and focused on the application of machine learning and natural language processing techniques to sentiment classification [61,87,112].

Interest in sentiment analysis among data mining practitioners and academics grew rapidly in the mid-2000's as blogging and online reviewing became ubiquitous, as blogs are laden with opinions about products, services, public figures, and other entities of interest to academics and industry professionals [87]. This data made it possible to understand public sentiment about particular products and events, as well as the emotional state of the public more generally [87]. The explosive growth of social media near the end of the decade further accelerated sentiment analysis research, as it provided even more data about products and services from which opinions could be mined [61,86]. In the past several years, sentiment analysis has successfully measured product sales [88], national mood [87], and it has even predicted fluctuations in the Dow Jones Industrial Average [11].

2.2.3 New Developments: Semantics and the Web

Recent research in the sentiment analysis field has worked towards solving the technical challenges described above and pushing the field in new directions.

One growing area is the combination of sentiment analysis methods with semantic methods. Saif (2012) showed that adding semantic concepts as features can improve the accuracy of a Twitter sentiment classifier [107]. Mukherjee (2012) also designed experiments to show the effectiveness of features derived from semantic tags [78]. Mohtarami (2012) and Mohtarami (2013) took a different approach, explaining that particular phrases can be mapped to particular,

known semantic meanings, which can be used to make inferences about the sentiments of two different texts [72,73]. This, in turn, can improve classification performance compared to traditional methods. Overall, these efforts have shown that blending traditional supervised learning approaches with semantics has proven very effective.

Other authors have examined feature extraction methods for sentiment analysis of social media. For example, Aisopos (2012) drew a distinction between content-based and context based feature extraction for Twitter sentiment analysis, and compared the effectiveness of both methods [3]. Barbosa (2010) also discussed the use of metadata and other sources of features for Twitter sentiment prediction [5]. Other authors have proposed wholly new approaches, such as Agarwal (2011), which devised a feature extraction method using tree kernels to measure the similarity between the part-of-speech tags and parse trees of Twitter posts [2]. These feature extraction innovations are typically driven by new types of data from social media, such as Liu (2012b), which examined the use of emoticons to smooth out n-gram models [62]. However, some of these methods are still applicable to more traditional texts. For example, Liu (2012c) outlined a feature extraction model that learns about expression sentiment using simple heuristics; the method is applicable to both social media short texts and longer, more traditional movie reviews [63]. Similarly, Zhai (2011) designed a method for clustering features that is applicable to any type of text [146].

Other efforts are harder to classify, including new definitions of sub-problems in the field and development of systems for specific platforms. For example, Kucuktunc (2012) applies standard supervised learning methodologies to perform sentiment analysis on a large, messy web dataset from Yahoo! Answers, and Mejova (2013) focuses on predicting political sentiments within Twitter [54,68]. Some of these contributions include redefinitions and new types of classification problems. For instance, Silva (2011) suggests that the problem of estimating sentiment over time should be viewed as a data streaming problem [116]. Kessler (2012) and Kim

(2012) examined some classification problems tangentially related to sentiment analysis: Kessler (2012) attempted to identify words that are used inconsistently, and Kim (2013) used supervised learning to predict whether sentences contain explanatory content [52,53].

Nearly all of these new developments have taken advantage of advances in natural language processing as well as semantics, and many modern approaches to sentiment analysis incorporate part-of-speech tagging [61]. However, despite these trends, sentiment analysis is still a monumentally difficult and unsolved problem [61].

Nevertheless, some progress has been made on the canonical problems in sentiment analysis described by Pang (2008) and Liu (2012). Mukherjee (2013) presented a new method for detecting opinion spam in Amazon product reviews, and Cheng (2012) developed a new type of n-gram model for multilingual sentiment analysis of large social media datasets [21,75]. This method makes it possible to use n-gram models even on languages such as German, which have flexible grammar that determines word ordering [21]. Next, Xu (2012) expanded on the classic problem of subjectivity identification by showing how the problem of identifying subjective statements is essential to performing sentiment analysis on blog posts [137]. These efforts imply that the framing questions established by Pang (2008) and others remain relevant, and have guided many research initiatives.

2.3 Transfer Learning

However, one of the largest research initiatives in the machine learning field focuses on cross-domain classification, also known as transfer learning. In transfer learning, the training set consists of documents from one domain, and the testing set consists of documents from another, possibly unknown domain [9].

In the text mining domain, this problem was first formalized in Blitzer (2007) and Tan (2007). Tan (2007) introduced a naïve solution, to this problem, namely using a classifier trained in one domain to label examples in another domain, and then retraining the classifier on these examples [122]. Similarly, Blitzer (2007) outlined some algorithms for finding mutual information between feature spaces in different domains [9].

Since 2007, a variety of approaches to transferring knowledge, particularly sentiment knowledge, have been proposed. Tan (2009) demonstrates a feature selection method designed to work with a Naïve Bayes classifier for transfer learning, in which the model selects generalizable features that work well in both the target and source domains [121]. Working along similar lines, Xia and Zong (2011) suggested that features based on part-of-speech tags could be used for cross-domain learning, as grammatical characteristics are often similar even in different domains [136]. Wu (2009) proposed a graph-based model, in which the sentiment of a user and their PageRank is incorporated into a transfer of learning model [134, 135].

More recently, researchers in the transfer learning/sentiment analysis domain have focused on building a sentiment lexicon, or a corpus of sentiment-laden words that can be used in multiple domains. For example, He, Lin, and Alani (2011) demonstrates that a joint topic and sentiment model can be used to locate general indicators of polarity by finding words that have similar polarity across multiple topics [42]. Jialin Pan (2010) and Wu & Tan (2011) developed similar methods involving a word-clustering model that can act as a "bridge" between two domains by finding words with similar sentiment in both domains [47,133]. Yoshida (2011) also focused on building bridges between domains, though the authors accomplish this at the level identifying individual words rather than word clusters [144]. Next, Chetviorkin (2012) described an approach in which sentiment lexicons are learned in specific domains, and then generalized to "meta-domains" that include multiple domains with similar lexicons [23]. Finally, Ponomavera (2012), acknowledging the difficulty of cross-domain classification, developed a method for

predicting the loss in accuracy when transferring a model between domains. Despite more than half a decade of progress in cross-domain classification, transfer of learning is still considered a difficult problem [99].

2.4 Experience Mining

A small research area of data mining research that is also heavily related to sentiment analysis is the field of experience mining. Experience mining is the use of computational methods to identify and extract descriptions of personal experiences from user-generated content on the Internet [in45 Experience mining is related to the problem of opinion spam, as both research areas aim to identify a subset of text messages that are considered relevant and filter out messages that are considered irrelevant. Inui (2008) was the first to define experience mining and differentiate it from sentiment analysis, observing that "subjective information in sentiment analysis…is only half of the possible harvest from UGCs [user generated content]. UGCs contain not only subjective material but also a vast range of factual, objective statements describing such personal experiences" [45]. Though experience mining incorporates sentiment analysis, it emphasizes the identification of personal experiences rather than the identification and summarization of opinions. Inui et al. argued that experience mining consists of four technical challenges: Event mention extraction, entity-event relation extraction, factuality analysis, and experiencer identification [45].

Inui (2008) also demonstrated the utility of experience mining by implementing an experience search engine. In this system, supervised machine learning methods are used to classify Japanese web blogs into experience classes based on the types of personal experiences they contain, with categories such as "shampoo," "beverage," and "automobile" [45]. A search engine allows users to search this database of experiences and only receive results from

categories of their choosing [45]. Smyth (2009) suggested that experience mining can be used to create a case-based reasoning system, in which descriptions experiences are collected from the web and reused by a person or decision agent when making a decision [118]. Smyth (2009) outlined three core challenges to creating such a system: Capturing personal experiences from the web, coping with the noise that results from using non-expert descriptions of experiences, and selecting relevant experience descriptions from a database [118].

Other researchers have extended this work and developed applications that attempt to solve the four research goals outlined in Inui (2008) and Smyth (2009). Jijkoun (2010) applied supervised learning to classify forum posts based on whether each post contains descriptions of "successful" or "unsuccessful" personal experiences, or no personal experiences at all [49]. Jijkoun et al. found that linguistic features, including bag of words and part-of-speech tag counting, are well-suited for this task [49]. Park (2010) continued this line of research, and explored the usefulness of features based on the semantics of verb usage, which can be used to predict the semantic meaning of sentences based on their verb tense and placement [89]. Ryu (2010) focused on extracting daily life events from web data, and compares several statistical algorithms for extracting instructions on how to perform household tasks from web articles [106]. Abe (2011) better defined the problem of factuality analysis, and used a machine learning-based model to predict whether a document describes an event in the past or discusses a hypothetical event [1]. Abe (2011) also outlined an experience mining application that automatically collects experience descriptions from Japanese blogs, extracts events, and stores them in a database with semantic tags [1].

More recently, Myaeng (2012) proposed several new methods for event extraction and identification of temporal links between events [78]. Myeang (2012) also studied several approaches for extracting events from a stream of social media data and inferring the context in which these events took place [78]. Similarly, Sauer (2012) introduced SEASALT, an application

that performs case-based reasoning by extracting experiences from Internet communities and web articles and selecting relevant cases to answer domain-specific questions in the Java programming domain [113].

Despite the small volume of literature on experience mining relative to other areas of sentiment analysis, authors in this field have successfully defined the problem of identifying personal experiences on the web, and demonstrated several approaches to experience extraction. However, the field is still in its infancy, and none of the existing literature attempts to apply the problem of experience mining to short texts or use advanced NLP techniques to improve feature extraction. Moreover, the existing literature demonstrates that experience mining is still a very difficult problem, as classifying subjective statements without context is a challenging task for any classification algorithm.

Chapter 3

Sentiment Analysis in Healthcare

Researchers in medical fields and computer science have written enthusiastically about using sentiment analysis and other text mining methods to improve the healthcare system. Salathé (2012) argued that these technologies can create a new "digital epidemiology" that will give medical professionals a better understanding of health risk factors and disease outbreaks [110]. Online sentiment analysis, Salathé contended, allows researchers to "study individuals and groups in the rich contexts in which their lives unfold, and to study person-to-person spread of disease and behaviors at the level at which it actually occurs" [110]. Brownstein (2009) also claimed that sentiment analysis and text mining in general can provide pharmaceutical companies with valuable information about the effects of new drugs by mining online texts containing patients' reactions to these medicines [13]. Brownstein et al. also suggested that this web data could serve as supplement to clinical trials when new medicines are being evaluated [13].

In practice, however, sentiment analysis has yet to be widely applied to the healthcare domain, and most of the research conducted thus far has been carried out in isolation. Though the idea of sentiment is intrinsically linked to emotional health and public well-being, few studies acknowledge this relationship, and even fewer perform sentiment analysis in the context of public health. Existing sentiment analysis research in healthcare-related areas can be divided into roughly three categories: The study of sentiment for public health surveillance, analysis of online health social networks, and sentiment analysis of suicide notes. A related field in the medical domain, content analysis of health-related texts, has also been exchanging ideas with the sentiment analysis community, which may influence the direction of sentiment analysis research.

3.1 Public Health Surveillance

Public health surveillance is the collection and analysis of health-related information [123]. Public health surveillance is typically performed by organizations such as the Centers for Disease Control (CDC) and World Health Organization (WHO), generally using surveys, polls, and news reports [123]. According to the World Health Organization, public health surveillance has three principal goals: Serving as an early warning system against health emergencies, tracking progress towards health goals, and setting policy agendas for lawmakers [123].

In the past several years, text mining and sentiment analysis techniques have been applied to online health data in an attempt to build automated systems to assist in public health surveillance. Most sentiment analysis applications in the public health domain have been designed to mine data about specific problems and specific medical conditions, but a few general-purpose tools for mining arbitrary public health-related opinions have been created. For example, Bobicev (2012) and Yoon (2012) used supervised polarity classification methods to discover health opinions about a wide variety of medical conditions in Twitter posts [10, 143]. Bhattacharya (2012) extended this work and developed a sentiment and knowledge discovery system that dynamically creates statements of the form "X causes Y" using words that co-occur in tweets with medical terms, and then collects new Twitter posts and classifies their polarity to determine the public's sentiment towards each statement [7]. Parker (2013) applied similar methods to the problem of predicting and tracking public health trends, and shows that polarity classification of tweets can track the public's interest in health topics [90]. Additionally, Goeuriot (2011) considered the problem of developing a sentiment lexicon for health topics, or a collection of health-related words and their polarity values [36]. Such a lexicon, Goeuriot et al. argued, could be used to improve opinion mining systems in the healthcare domain, particularly those that

target patients [36]. Nonetheless, the majority of research in this area has avoided the problem of developing general-purpose lexicons and had instead focused on specific healthcare sub-domains.

3.1.1 Specific Diseases and Disease-Related Events

Sentiment analysis has been used to track attitudes towards particular medicines, vaccines, and diseases. Some of these research efforts focus on pubic opinion related to specific medical conditions and how patients suffering from these conditions are coping. For example, Jamison-Powell (2012) analyzed tweets related to insomnia and classified how Twitter users discuss insomnia with each other into two themes, "experiencing" and "coping" [46]. Prabhu (2012) tracked the change in sentiment of tweets about prostate cancer following an announcement by the United State Preventative Services Task Force and argues that policy makers should use sentiment studies when examining public reaction to policy [100].

Aggregate social media data has also be used to discern public sentiment about specific healthcare topics, most notably influenza. For example, Signorini (2011) found that the number of tweets about the H1N1 ("swine flu") virus decreased in May 2009 even as the number of cases increased, suggesting a decrease in public interest in the outbreak [115]. Marcel Salathé and his colleagues have published a suite of papers on sentiment analysis and vaccines: Salathé et al. (2011) collected tweets related to H1N1 vaccinations, finding that sentiment towards vaccination by region are strongly correlated with the CDC's statistics on vaccination rates: Regions with predominantly negative sentiment towards vaccines tend to reject vaccinations [109,111]. Salathé et al. also studied the connection between Twitter uses using the "followers" and "friends" features, and observed that users with a particular sentiment are more likely to associate with users who share that sentiment, be it positive or negative [109,111]. Additionally, Salathé et al. (2012) found that negative sentiment is contagious and can lower vaccination rates over time

[111]. Previous work by Salathé suggests that communities with low vaccine acceptance are significantly more likely to experience outbreaks of vaccine-preventable diseases such as measles, mumps, and rubella [108]. Lamb (2012) also explored public response to influenza outbreaks by studying sentiment and information sharing behavior on Twitter during an outbreak [55].

In addition to influenza, sentiment analysis has been used to explore other medical conditions, including respiratory conditions, pregnancy, and mental illness. Myslin (2013) performed sentiment analysis on tweets related to smoking, in order to study smoking behavior and sentiment to several types of tobacco products [79]. Gillingham (2012) studied respiratory conditions by developing a polarity classification system for tweets about asthma as means of collecting data for public respiratory health surveillance [35]. The intersection between mental illness and sentiment analysis has also been explored in recent years by works such as Huang (2007) and Li (2012) [44,59]. Huang (2007) used keyword-based queries to locate MySpace users who may be suicidal; Li (2012) used supervised sentiment analysis methods to identify emotional distress among bloggers in order to identify bloggers who many need therapy or emotional support [44,59]. Working along similar lines, Brubaker (2011) analyzed how MySpace users expressed grief after the loss of a loved one using a variety of natural language processing methods [14].

More recently, these efforts have expanded to lifestyle health factors such as diet and pregnancy. Crouch (2012) applied supervised learning methods to sentiment analysis of speech, as part of an automated system that advises individuals about their diets and records their emotional reactions [27]. Additionally, Choudhury (2013a) and Choudhury (2013b) used sentiment analysis to predict post-partum depression in pregnant women in social media, based on the posting patterns of these women and their use of language [24,25]. Sentiment analysis has also been incorporated into knowledge extraction systems, such as a system described in Denecke

(2009), which examined content differences between sources of online health information including blogs and wikis [29].

Despite the proliferation of sentiment analysis research in public health, research in this area is fragmented rather than united, and the machine learning systems described above are highly specialized to particular domains and platforms.

3.1.2 Healthcare Quality

Other research has used sentiment analysis to predict and estimate the quality of healthcare services at the level of individual doctors, hospitals, and nation-wide healthcare systems. These studies use data from a variety of sources, including social media, patient feedback forms, and physicians' notes.

Greaves (2012) assessed the quality and cleanliness of hospitals in the United Kingdom using sentiment analysis of free-text responses from an NHS online survey [38]. The authors reported that sentiment score can predict whether respondents rated a hospital as "clean" with 81% accuracy, whether they stated they were treated with dignity with 83% accuracy, and whether they would recommend the hospital to others with 89% accuracy [38]. Greaves (2013) expanded upon this work, showing that social media can also be used as an indicator of hospital quality in the United Kingdom [37]. Working along similar lines, Cambria (2010) developed an ontology-based system for classification of patient-related texts, and used to rank NHS hospitals [16]. Cambria (2012) used these findings to develop a software application that tracks patient health status as uses this data in aggregate to measure healthcare quality [15].

In the United States, Paul (2013) attempted to determine which factors influence patients to give positive and negative ratings of their doctors, using a dataset of 50,000 online reviews of physicians [93]. Using a joint topic and sentiment model, Paul et al. examined how particular

topics differ between positive and negative reviews, including “knowledgability,” “staff,” and “helpfulness” [93]. Xia (2009) also performed polarity classification of online patient reviews using topic models; Smith (2012) and Alemi (2012) applied similar methods to patient emails and hospital satisfaction surveys [4,117]. In addition, Zhang (2012) developed a join topic-sentiment method to resident evaluations of medical trainees, suggesting that sentiment analysis can be used as a measure of trainee performance [147].

At a more fine-grained level, Murff (2011) built a classification system to predict whether a medical procedure was completed successfully or developed complications, using polarity classification and a dataset of clinical notes [76]. At the healthcare system level, Steele (2012) suggested a broad and ambitious approach: Retrieve text from multiple online sources, combine it, and then use sentiment analysis to measure the overall “zeitgeist” of how the public feels about healthcare services [119]. However, this dream has yet to be realized, as sentiment-based measures of healthcare quality have not been widely validated or popularized.

3.1.3 Drugs and ADRs

Sentiment analysis has also been used to gauge public opinion about particular medications, and to identify adverse drugs reactions (ADRs), or unintended harm associated with taking the prescribed dose of a medication. Detection of ADRs is strongly related to polarity classification, in that it is a binary classification task on a set of documents with a “positive” and “negative” outcome; many of the same methods are used for both tasks [140]. Research in this area began with Melton (2005), which used text mining methods on patient discharge summaries from New York hospitals and concluded that supervised learning can identify adverse drug reactions from the discharge summaries [69]. Along with other research efforts, Chee (2011) extended this work, and outlined a system that searches for negative sentiment towards particular

drugs in order to uncover any adverse side effects of these drugs that might not have appeared in clinical trials [18]. Yates (2013) also created a text mining system to identify social media posts that signify ADRs [142].

In addition to ADR recognition, sentiment analysis has been used to mine user review of drugs from social media posts. Leaman (2010) was among the first to demonstrate that blog comments that contained medical reviews could be classified to predict the sentiment that an individual has toward a particular drug [57]. More broadly, Yalamachi (2011) outlined the technical specifications for SideEffective, a sentiment mining system that crawls the web and analyzes patient reviews of arbitrary medications [139]. Na (2012) also described a rule-based system to that also has the ability to classify sentiment in drug reviews [80].

Other methods for processing language to uncover attitudes about medicine have also been suggested. Neustein (2007) proposed a semantics-based model called Sequence Package Analysis to identify patient opinions about medicine [81], and Goeuriot (2011) and Kaiser (2012) developed aspect-based sentiment analysis models for web forum discussions of popular drugs, in order to summarize patient feedback related to these medications [36,51]. Goeuriot et al. demonstrated that forum posters are more likely to discuss negative aspects of medications, and that anxiety, weight loss, and pain relief drugs are most frequently discussed [36]. Other methods for aggregating patient feedback about medications have also been explored: Jiang (2012) described the specifications for an unsupervised learning system to cluster patient responses to medications, as extracted from online forum posts [48]. Researchers have also studied external factors that influence the public's perception of drugs: Chee (2009) used sentiment analysis on Twitter posts to explore the effect of FDA announcements on public sentiment; according to Chee (2009)'s analysis, these announcements can have a significant effect on how a medicine brand is perceived [17].

Though a comprehensive approach to mine patient feedback has yet to be developed, existing research clearly demonstrates the value of sentiment analysis in predicting and summarizing patient feedback about medications.

3.1.4 Academic Opinions

Sentiment analysis has also been used to mine academic and professional opinions about healthcare topics. For example, sentiment analysis has been applied to academic research in the medical field as a means of summarizing this literature and predicting the efficacy of medications. For example, Niu (2004), Niu (2005), and Sarker (2011) applied polarity analysis to academic articles about clinical trials, attempting to predict whether a given paper suggests that a particular drug is recommended for a particular ailment [82,83,112]. Swaminathan (2010) and Miao (2012) applied polarity classification and sentiment strength detection to a much larger classification task, identifying relationships between foods and medical conditions [70,120]. In this classification task, each relationship between a type of food (e.g. “soybeans” or “green tea”) and a medical condition (e.g. “gastric cancer”) has some unknown polarity value that represents the effect of the food item on the condition, which is estimated using supervised learning on biomedical publications [70]. This estimation summarizes the medical literature’s understanding of the effects of food on particular diseases, and determines which foods are healthy and which are harmful [70].

Other research has used sentiment analysis to study medical professionals themselves. Desai (2012) used Twitter sentiment analysis to examine how academics communicated at an American Society of Nephrology (ASN) conference, and Lewis (2012) applied sentiment analysis to determine what medical professionals think about the increasing popularity of the Doctor of Nursing Practice (DNP) medical degree [30, 58]. Though this sub-field is relatively new, it

suggests that sentiment analysis can be used to perform automated meta-analysis of the medical field and help scholars perform the difficult task of aggregating expert knowledge.

3.1.5 Public Happiness

Despite the abundance of health-related data in social media, there are few studies that use sentiment analysis to understand public health and general well-being [92]. However, a few recent pieces of research have used sentiment analysis of microblogging websites to create maps of happiness and general public well-being. Bollen (2011) performed a large-scale sentiment analysis of all public tweets published between August and December 2009, observing that major economic and political events strongly influence well-being as measured by Twitter sentiment [12]. Mishne (2006) reported similar results from analysis of blogs, and demonstrates that catastrophic events like the London train bombings of 2005 have a tremendous effect on public sentiment [71].

Paul (2011) performed keyword-based analysis on Twitter data to determine the incidence of common illnesses and sentiment about them [92]. The incidence of disease estimated from Twitter weakly correlates with CDC statistics [92]. Lansdall-Welfare (2012) also used Twitter data to track incidence of the flu and public mood over the span of several months, and experimented with visualization methods to transform this data into a facial expression that represented how the public's feelings changed as time passed [56]. The free online tools WeFeelFine [131] and the Twitter Well-Being Tracker [127] currently provide users with the ability to carry out similar analyses with live data; the latter specifically searches for keywords related to physical and emotional health and provides an aggregated score for each category [127].

Some surveys of Twitter sentiment have focused on particular regions and developed more thorough metrics to evaluate public feeling. Quercia (2012a) studied Twitter use in London and aggregates sentiment scores by region: sentiment scores are assigned to each profile based on a sample of tweets, and the scores of Twitter users in a particular region are then averaged to create a "Gross Community Happiness" score for that community [103]. Quercia et al. found that this GHC score is strongly correlated with measures of socio-economic well-being and with other metrics of emotional health [103]. Quercia (2012b) continued this work using topic modeling, finding that healthy and well-off London communities discuss different topics compared to poor and unhealthy communities [104].

Overall, these studies suggest that microblogging is a rich data source for assessing emotional health and well-being. Despite the biases inherent in using Twitter as a data source [92,109], millions of people use the service to talk about their health and well-being, making it an attractive data source to study public health.

3.2 Health Social Networks

Online health communities are online platforms that support interpersonal exchanges where patients may "talk to friends, relatives, and professionals about what their diagnosis and treatment may entail" [28]. Prior to the application of automated sentiment analysis to online communities, social scientists and medical professionals explored online health communities and their ability to provide social support [150].

Several recent works have applied sentiment mining to understand patient interactions in online support groups. Yu (2011) performed a preliminary study of emotions in online health communities by quantifying emotional content in WebMD discussion threads [145]. Yu (2011)

used affective word counting to compare the emotional content of discussions of breast cancer, prostate cancer, and cancer treatment, but concluded that these findings are only preliminary due to the simplicity of the algorithm used and difficulties of accurately classifying text from online forums [145]. Additionally, Denencke (2009) developed a topic-based model to identify contradicting opinions on health-related web blogs, which can be used to track a blogger's opinions over time, and demonstrated its effectiveness on WebMD blog posts [29].

Sentiment analysis has also been used to summarize the prevailing social behaviors of large online health communities. Qui (2011) performed a much deeper analysis of forum posts on the American Cancer Society's Cancer Survivors Network [102]. Qiu et al. (2011) classified posts using the meta-learning algorithm AdaBoost, and found that 75%-85% of posters were classified as having negative sentiment have positive sentiment after interacting with other community members [102]. Additionally, Qiu et al. demonstrated that the addition of classifiers for how frequently posters use slang and whether posters refer to each other by name can improve classification when working with online health community texts [102]. Zhao et al. (2011) and Zhao et al. (2014) expanded on this work by performing further sentiment analysis on the Cancer Survivors Network in an attempt to identify community leaders [148,149]. Zhao et al. also studied the emotional effect leaders have on their communities, observing that community members experience noticeable sentiment changes when community leaders' health deteriorates [149]. Ofek et al. (2013) demonstrated that a dynamically constructed sentiment lexicon can improve the performance of supervised learning algorithms for the task of polarity classification on the Cancer Survivors Network [84].

Other research has studied the effectiveness of online health forums used as a supplement to other forms of social support and treatment. Chen (2011) analyzes data from an online forum that was used as part of a cancer survivors workshop at Stanford University [20]. After comparing a variety of supervised and unsupervised learning methods, these sentiment

classifications are used to assess the overall effectiveness of the workshop [20]. Chen (2011) concludes with a discussion of the potential to use workshop data to predict health outcomes, but argues that it would not be sufficient to track a patient's progress beyond the short term [20]. In the diabetes domain, Chen (2013) presented DiabeticLink, an online health social network that uses sentiment analysis to aggregate users' opinions about specific drugs and incorporates this information into its search functionality [19].

Despite the challenges associated with studying online communities, these studies have already led to important insights about how health communities grow and support their members, and they make a strong case for the role of sentiment analysis in studying emotional support.

3.3 Suicide Note Classification

Sentiment analysis has also been used to analyze suicide notes. This subfield of sentiment analysis builds upon studies in clinical psychology and suicidology, which used qualitative methods to discuss the linguistic properties of suicide notes. For example, in a well-known paper, Ogilvie (1966) compared the semantic and linguistic properties of genuine and simulated suicide notes, using content analysis to gauge authenticity [85]. Statistical methods were first applied to suicide notes in 2007: Jones (2007) used manually coded suicide notes to construct statistical prediction rules (SPRs) for suicide note classification using average sentence length, frequency of parts-of-speech, and other criteria [50]. Similarly, Handelman (2007) assembled a dataset of suicide notes from male and female attempters and completers of suicide, and used two-way ANOVA on a set of manually coded linguistic variables, including the presence of pronouns, the use of future tense, and references to religion or metaphysics [40].

Sentiment mining of suicide notes began with Pestian (2008), which applied supervised machine learning to distinguish between real suicide notes and simulated suicide notes written by

not-at-risk participants [95]. To extract features related to semantic content, Pestian (2008) created a classification schema of emotions often present in suicide notes, which included concepts such as affection, anger, depression, and worthlessness [95]. Each class was further divided into multiple concepts, e.g. the class affection included the sub-classes "love, concern for others, and gratitude" [95]. Pestian et al. compared the classification accuracy of a variety of machine learning algorithms [95]. Pestian (2008) reported that a linear SVM classifier achieved 78% classification accuracy, 7 percentage points higher than classification by healthcare professionals [95]. Further research by Pestian (2011) expanded on this work, and demonstrated that a logistic model trees classifier trained on a large featureset consisting of part-of-speech and other linguistic features can identify genuine and simulated suicide notes more accurately than psychiatric trainees and healthcare professionals [97].

Research on suicide notes within sentiment analysis research community continued to expand with a contest held by the American Medical Informatics Association in 2011, in which participants were asked to classify emotions in suicide notes using a much larger dataset than previous studies [96]. In total, 106 scientists composing 24 teams submitted results; different researchers from different countries applied a variety of different learning approaches [96]. The contest also led to a number of publications in 2012 by participants summarizing the algorithms and feature extraction methods used [31,66,96,130,132]. This research effort led to the creation of more complex tools for sentiment analysis and emotion identification in suicide notes. For example, Wicentowsky (2012) and Desmet (2012) classified suicide notes based on 15 pre-defined emotional classes, and discuss heuristics for text preprocessing to fix spelling and grammatical errors that might hamper sentiment analysis [31,132]. Wang (2012) and McCart (2012) both proposed hybrid models classifying suicides notes that combine several existing NLP-based feature extraction methods and a rule-based model, and argued that combining multiple models provides more accurate classification than individual learning [66,130]. Yang

(2012) also conducted experiments with a hybrid model, using a voting mechanism to combine the output of several types of classifiers [141]. Much like a Random Forests classifier, each model was given a variety of different semantic and linguistic features [141]. All of these publications reported higher classification accuracies compared to Pestian et al.'s efforts in 2008 and 2011, suggesting that these hybrid models and new approaches are indeed very powerful text classification tools.

Suicide note research has also motivated new approaches to fine-grained sentiment analysis. Luyckx (2012) explored fine-grained sentiment extraction using the 15-emotion model, and Cherry (2012) presented a model based on latent sequence modeling, which divides sentences into segments, or "emotion regions," and predicts the emotion contained in each region [22,65]. Read (2012) discussed features that can be used to determine fine-grained sentiments within sentences of suicide notes, and compared the performance of these features using several types of machine learning algorithms [105].

Other approaches to classification of suicide notes have also been developed, such as Xu (2012), which demonstrated that supplementing a training set of suicide notes with text extracted from web blogs can improve classification performance [138]. In an attempt to apply these methods to real at-risk patients, Pestian (2013) performed a prospective clinical trial that incorporated machine learning [94]. In this trial, suicidal and non-suicidal adolescent patients were asked questions about their mental health status, and machine learning methods were used to predict whether the patients should be classified as suicidal; Pestian (2013) reported that the 60 patients involved in the study were classified into suicidal and non-suicidal groups with 93% accuracy [94].

Overall, many of these studies represent the cutting edge of sentiment analysis research, and they may lead to the construction of systems that can reliably outperform humans at the task of distinguishing authentic suicide notes from simulations [96].

3.4 Content Analysis

Social science researchers have also examined emotional and affective content in healthcare-related texts, generally without the use of machine learning methods. Instead, these researchers study the role of emotion in these texts using content analysis, a collection of qualitative research methods that explores how individuals communicate and express their feelings in writing. Content analysis has been applied to health-related posts on social media, and studies using content analysis methods tend to explore their source material more deeply than sentiment analysis, though often at the expense of scale. For example, Scanfeld (2010) used multiple human coders to identify negative sentiments towards influenza vaccines in a sample of 1000 Twitter posts [114]. Scanfeld et al. also categorized these tweets and noted that many of the tweets sampled contain inaccurate information or misinformation about vaccines [114]. Content analysis also allows authors to examine media other than text, such as Lo (2010), which attempted to gauge public perception of epilepsy by studying YouTube videos [64]. McNeil (2012) supplemented this research by performing content analysis of tweets about epilepsy, finding that epilepsy is often used as the subject of jokes or discussed in a negative light [67].

Other studies use content analysis to provide qualitative and descriptive commentary on the role of social media in healthcare and patient support. For example, Greene (2010) performed a qualitative analysis of 15 Facebook groups about diabetes and analyzed how their members use Facebook to communicate [39]. Similarly, Prochaska (2011) examined Twitter accounts that spread information about smoking cessation, and concluded that the advice that these Twitter users post is often inconsistent with clinical recommendations about quitting smoking [101]. Finally, Heaivilin (2011) used content analysis to explore information sharing behaviors related to dental pain on Twitter, using a sample of 772 tweets [43]. Despite the lack of scale found on machine learning-based approaches to extracting sentiment from social media posts, content

analysis offers an alternative set of tools with which to more deeply examine discussion of healthcare on the web.

Chapter 4

Data Collection

4.1 Twitter API Collection

The four datasets used in this thesis consist of posts published in the public domain on the Twitter platform. These four datasets of tweets were collected between January and March 2013 using Twitter's Streaming API, which allows developers to query Twitter's servers to retrieve tweets that contain particular keywords in real time. In this thesis, four datasets corresponding to four different chronic diseases were collected. The keywords used to collect each dataset are shown in the table below:

	Breast Cancer	**Prostate Cancer**	**Lung Cancer**	**Diabetes**
Keywords used to collect	"breast cancer" "breastcancer"	"prostate cancer" "prostatecancer"	"lung cancer" "lungcancer"	"diabetes" "diabetic"

Table (1). Twitter API collection keywords.

The collection process for each dataset was performed by a script that opened a persistent connection via the Twitter API and then collected tweets over a period of 10 hours. The tweets were then saved in JSON format for filtering and tagging.

4.2 Data Filtering

After the tweets were collected from the Streaming API, the corpus of tweets was filtered based on two criteria. First, the corpus was filtered by language, in order to eliminate non-English language tweets. This was achieved using the metadata from each tweet: Twitter user accounts include an editable field that denotes the user's language. Tweets that were published by users whose language was not set to "English" were removed from the corpus. The table below shows the reduction in size of each dataset when this filter was applied:

	Breast Cancer	**Prostate Cancer**	**Lung Cancer**	**Diabetes**
Corpus Size Before English Filter	4459	2302	1489	2537
Corpus Size After English Filter	4026	2171	1449	1933

Table (2). Effects of English-language filter.

Second, the collection of tweets was filtered to remove tweets that are duplicates or very similar to other tweets in the corpus. Given that retweeting is the predominant method of spreading information on Twitter, any collection of tweets containing the same keywords will likely contain a large number of tweets that are identical or nearly identical. Many retweets contain the same text as the tweet based on, and others will consist of the same text with small alterations. Two common methods of signifying that a tweet is a retweet are: a) the addition of the string "RT" plus a username appended to the beginning of the tweet, and b) the addition of the string "via" plus a username to the end of the tweet. The retweeting user may also add hashtags or a short comment to the end of the retweet, depending on the number of characters remaining. For example, consider the following two tweets from the breast cancer corpus:

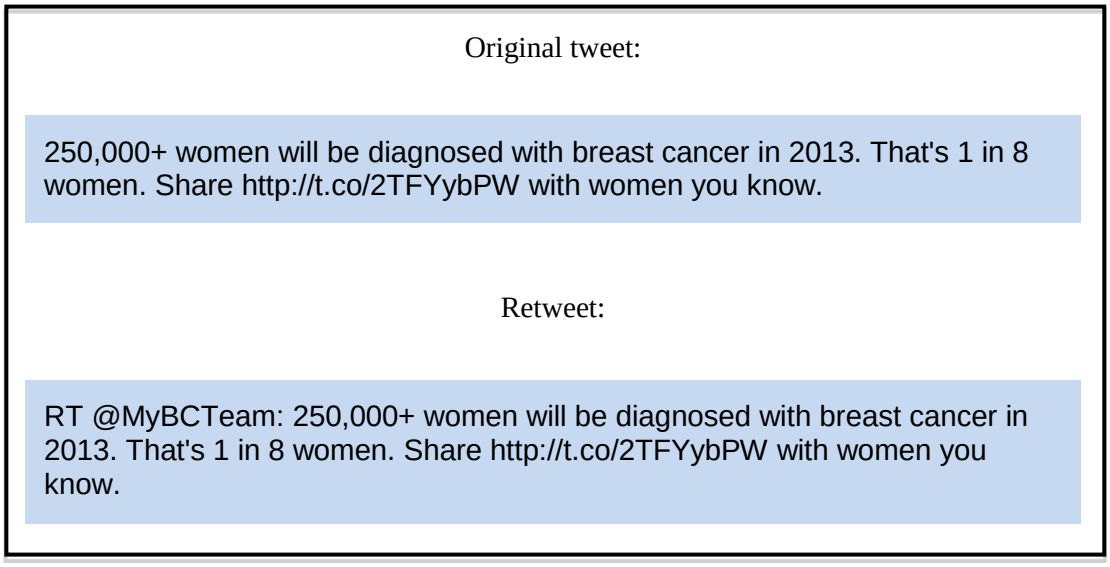

Figure (9). Tweet versus retweet text for a breast cancer tweet.

These tweets contain the same informational content; the only difference is 14 characters added to the beginning of the tweet to signify where the message originated. For the purposes of a text classification experiment, these near-duplicates represent a threat to experimental validity, as their presence will artificially improve supervised classification performance. As a result, tweets with very similar text must be removed from the corpus prior to preprocessing.

However, traditional string-matching algorithms are suboptimal for this task, as they are based on the concept of minimal edit sequences, or the number of characters that must be modified to transform one string into another [98]. For example, one commonly used metric to compute the distance between two strings is the Levenshtein distance, in which the distance between two strings is equal to the minimum number of one-character substitutions, insertions, or deletions required to change one into the other [98]. Levenshtein distance and similar algorithms are ill-suited for detecting retweets because of the way in which retweets are denoted. As discussed above, retweets are commonly denoted by “RT @Username:” or “via @Username,” where “@Username” is the Twitter username of the author of the original tweet. Retweets can also include additional hashtags or short (one or two word) comments, depending on the length of

the tweet. As Twitter usernames can be up to 15 characters long, a retweet with these signifiers may contain 20 more characters than the original tweet. As a result, the Levenshtein distance between a tweet and a retweet will be very high,

As a result, the Ratcliff-Obershelp pattern matching algorithm was used to remove duplicate and similar tweets. The Ratcliff-Obershelp algorithm defines the distance between two strings using the following algorithm [8]:

Given two strings A and B,

1) Find the "anchor," or the largest substring contained in both A and B and record its length.
2) Repeat this procedure recursively for the strings to the left and to the right of the anchor in both A and B.
3) Recursion stops when there are no more anchors to be found.
4) The Ratcliff-Obershelp similarity is two times the sum of the lengths of all the anchors, divided by the total number of characters in A and B.

In order to remove duplicate or near-duplicate posts from the corpus of tweets using the Ratcliff-Obershelp function as a measure of similarity, the following algorithm was used:

1) Starting with the first tweet in the corpus, find all other tweets that have a similarity higher than 0.8.
2) Remove these tweets from the corpus.
3) Repeat this process iteratively for the remaining tweets.

This algorithm has a worst-case running time of $O(n^2)$, as it computes all pairwise comparisons. This makes it inappropriate for larger datasets, but it is sufficient for this small corpus. The table below shows the size of the four datasets after retweets are removed:

	Breast Cancer	**Prostate Cancer**	**Lung Cancer**	**Diabetes**
Corpus Size Before Similarity Filter	4026	2171	1449	1933
Corpus Size After Similarity Filter	2159	644	699	1299

Table (3). Effects of retweet filtering using Ratcliff-Obershelp on four datasets.

4.3 Tagging

The author of this thesis and one other volunteer tagged the four tweet datasets, classifying posts as either "Personal" or "Impersonal." Posts tagged as "Personal" contain some description of personal experience related to a chronic disease. Specifically, in order for a post to be classified as Personal, it must satisfy all of the following conditions:

1. The post must be predominantly in English, such that its meaning is clear to an English-language reader.
2. The post must NOT contain only spam, i.e. advertizing for a particular product, service, fundraiser, or organization.
3. The post must explicitly and clearly describe or make reference to a real-world event related to a real individual's experiences related to chronic disease. This includes diagnosis of a chronic disease, treatment or management of a chronic disease, a health event such as surgery or death caused by a chronic disease, a fundraising/charitable event related to a chronic disease, emotional support or individuals with chronic disease, and general discussion about individuals suffering from a chronic disease. These descriptions can apply to the poster themselves or to a friend or relative of the poster.

Some examples of Personal and Impersonal tweets from the breast cancer dataset are shown in the table below:

Personal	Impersonal
@ZBusch34: So proud to say that today is my moms 4th year of being a survivor of breast cancer. Couldn't ask for a better mother.	EEOC Sues Law Firm for Firing Employee with Breast Cancer | @scoopit http://t.co/a2IRfPrT
#1000thtweet goes out to my Mom! Almost a year ago she was diagnosed with breast cancer and overcame it. She is always there for me #loveyou	RT @MyBCTeam: 250,000+ U.S. women will be diagnosed with breast cancer in 2013. Share http://t.co/2TFYybPW with women you know.
I was sooooo happy to hear from my aunt tonight .. She was in such good spirit. She's not letting breast cancer get the best of her.	Avon is the numba 1 fundraiser for the breast cancer cause! Buy Avon ladies! :D
I'm getting this on my forearm with a breast cancer ribbon around the top. Support for my grandparents. http://t.co/BIEacVgd	If you want to donate to breast cancer you only have to donate a dollar! Your name will be written on the banner at the game tomorrow!

I've had a breast cancer scare this week, I'm relieved to say that all is ok, but can I just say #checkyourbreasts	RT @MegRobertson: How breast #cancer survivors are redefining what recovery means: http://t.co/dkb21lpC9F with @personal_ink @annmarieg4 &
@JSCinnamon Good. I'm glad. Good luck with the walk for breast cancer, it's nice your doing for your mom. God bless her!	open access: Risk of breast cancer in Lynch syndrome: a systematic review http://t.co/astbuq2Ahp
RT @runningislife4: Sitting here with my granny that has survived breast cancer twice. she is laughing and joking.	A simple blood test may predict if breast cancer will return after treatment. http://t.co/ckcxL42K

Table (4). Examples of personal and impersonal tweets from breast cancer dataset.

All Twitter posts from all four datasets were labeled by the two taggers working as a pair, reading each post silently or aloud and discussing the content before agreeing upon a label. Any disagreements about labels were resolved by discussion during the tagging process; no significant disagreements about the tagging process or how different types of tweets should be classified occurred. The tagging was carried out on three separate sessions in April of 2013, which totaled approximately 10 hours.

After the tagging process was completed, it became apparent that the four datasets are highly unbalanced in terms of content—the majority of tweets collected do not contain descriptions of personal experiences. In all four datasets, fewer than 25% of posts were tagged as Personal; the proportion is relatively consistent across all four datasets. This is shown in the table below:

	Breast Cancer	**Prostate Cancer**	**Lung Cancer**	**Diabetes**
Percent of Dataset Tagged as Personal	20.8%	14.9%	24.4%	15.2%

Table (5). Percentage of personal posts in each dataset.

There are roughly four categories of tweets that do not contain personal experiences: First, there are tweets that contain news article headlines, such as the first two examples of

impersonal tweets in the right column of the table above. These tweets consist simply of the headline to a news story and a link to that story. Although the new article is related to chronic disease, the tweet itself contains no description of personal experiences. Second, there are tweets that share information related to public health, usually containing a link to a more detailed description.

Third, there are tweets promoting fundraising initiatives or attempting to raise awareness of chronic diseases. These posts may imply the existence of a real-world event, such as a breast cancer walk or a fundraiser, but do not explicitly discuss it. The third and fourth tweets in the right column of the table above are examples of this type of impersonal tweet—both attempt to solicit donations or raise awareness of upcoming charity events. Fourth, there are tweets containing jokes and other content that is not related to discussion of chronic disease. These types of tweets were much more common in the diabetes dataset, as the word "diabetes" is often used in a joking context on Twitter. For example, the tweet *"@N_HeWinsBy_KO lol this is a fact... Shit taste like liquid diabetes lol"* uses the word diabetes in a joking context to describe a sugary food, not the chronic disease. Fifth, and finally, there are tweets written primarily in languages other than English that were not removed by the language filter discussed above. Fortunately, this only a few tweets in each dataset belong to this category, which suggests that using metadata to remove non-English tweets is relatively effective.

Overall, the prominence of these types of tweets demonstrates that Twitter is primarily a medium for information sharing rather than experience sharing. This is likely due to the 140-character limit on tweets, as evidenced by the fact that many impersonal information-sharing tweets contain links to longer articles. In addition, the broadcast-like nature of the medium makes it ideal for spreading awareness of specific events, which explains the large number of fundraising-related posts. Nevertheless, many users do share their own experiences on Twitter; the next several chapters discuss novel methods for locating these posts.

Chapter 5

Feature Extraction

After the data collection and filtering process, the remaining corpus of tweets was preprocessed and reduced to a set of categorical and numeric features. Two types of features were extracted. First, there are context-based features, which are features derived from metadata about the tweet or about Twitter user who published the post. Second, there are content-based features, which are based on wordlists, part-of-speech tagging, syntax, special characters, and other properties of text. A total of 14 context-based features and 54 content-based features were extracted. The preprocessing and feature extraction process is outlined in the diagram below.

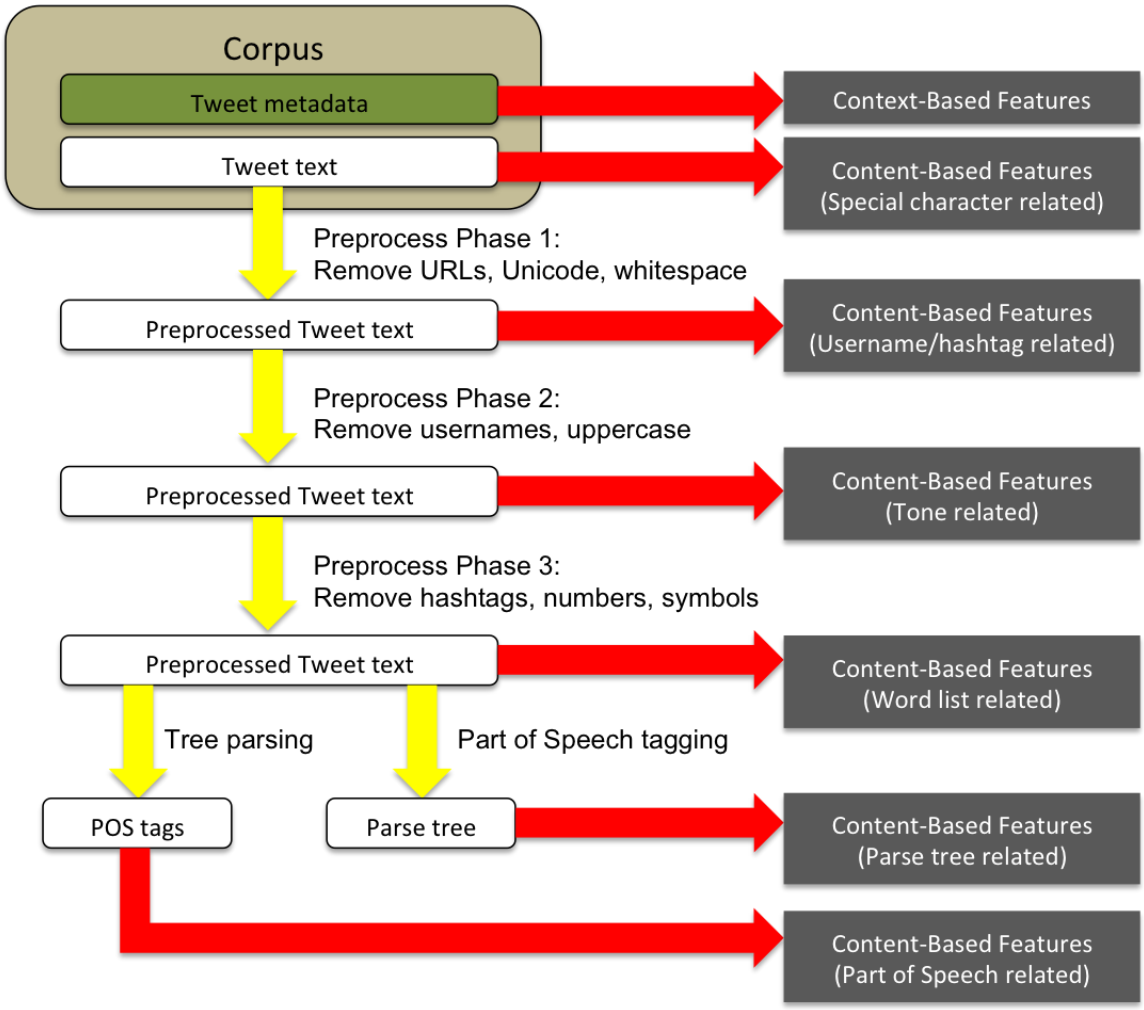

Figure (10). Feature extraction pipeline for Twitter datasets.

As the diagram shows, the feature extraction process consists of four preprocessing stages. Features are extracted before and after each of the four stages. First, context-based features and some content-based features are collected from the raw JSON tweet data. Next, the tweets undergo three phases of preprocessing using regular expressions to remove elements such as username, URLs, and other extraneous symbols. Before these elements are removed, they are used to create features about how Twitter users write posts about cancer. Finally, the tweets are

run through a Part of Speech tagging algorithm and a tree parser, and the POS tags and tree representation are used to develop the final two sets of features for this dataset.

5.1 Context-Based Features

Tweets collected from the Streaming API contain two types of metadata: Metadata about each tweet and metadata about the author of each tweet. These two types of metadata were then converted into two types of features, post metadata features and user metadata features. The first set of features to be extracted were the post metadata features, which capture some of the basic context surrounding each individual tweet. These three features are described in the table below:

Feature	Description	Type
Reply	Did the author select another tweet and write this tweet as a reply?	Categorical {Y, N}
Retweet	Did the user select another tweet and re-tweet it to create this tweet?	Categorical {Y, N}
Photo	Does the tweet contain an embedded image or link to a photo-sharing site recognized by Twitter?	Categorical {Y, N}

Table (6). Tweet context features.

Next, user-related features designed to capture basic information about the author of each tweet were extracted from each tweet's metadata. A wide variety of types of Twitter accounts post about cancer, including news organizations, cancer survivors, charities, and individuals with friends or family who have cancer. Different types of users post different types of content; some users are more likely than others to post about personal experiences. As a result, user-related features provide information about the content of a tweet by providing information about the type of user who posted it. These features are enumerated in the table below:

Feature	Description	Type
Followers	How many followers does the author's Twitter account have?	Numeric
Follows	How many Twitter users does the author follow?	Numeric
Posts Favorited	How many tweets has the author selected and added to their list of favorite tweets?	Numeric
Number of Posts	How many public tweets has the user posted to their tweet stream?	Numeric
Number of Lists	How many custom lists (lists created by Twitter users to organize the accounts they follow) does the author's account appear on?	Numeric
Profile Capitals	How many capital letters does the author's Twitter profile title contain?	Numeric
Profile Special Chars	How many non-alphanumeric characters does the author's Twitter profile title contain?	Numeric
Profile Length	How many characters are contained in the author's Twitter profile description?	Numeric
Profile Background	Has the author changed their Twitter profile background from the default background?	Categorical {Y, N}
Geolocation	Does the author's Twitter account have geolocation enabled?	Categorical {Y, N}
Profile Link	Does the author's Twitter profile include URL, such as a link to a personal web site?	Categorical {Y, N}

Table (7). User context features.

These user-based features serve several purposes. First, some of the features, including Follows, Number of Posts, Posts Favorited, Profile Background, and Geolocation, were designed to measure the author's level of usage of Twitter. For example, if a tweet author has a high number of public posts and a non-default background, this suggests that the author users Twitter very frequently and put a considerable amount of effort into their profile. Next, the Followers and Number of Lists features were designed to measure the user's popularity within the Twitter social network. Finally, the Profile Link, Profile Capitals, and Profile Special Chars features can be used to estimate whether a user account represents an individual or an organization. For example, a profile title with many capital letters and a URL suggests that the Twitter account represents and organization with an official name and website rather than an individual. Given that individuals

are more likely to tweet about personal experiences compared to Twitter accounts representing news organizations, these features are a form of indirect information about the content of the tweet.

5.2 Content-Based Features

After the context-based features were extracted from the tweet metadata, content-based features were extracted from the text itself. These content-based features are designed to capture the tone, message, and structure of each tweet. This is achieved using word lists of known categories of words, extraction of Twitter objects such as usernames and hashtags, and with the application of natural language processing methods. As a result, these features provide more fine-grained detail about each tweet compared to the context-based features. The preprocessing and feature extraction steps used to extract these content-based features are detailed below.

First, features related to special characters and punctuation were extracted from the tweet text using simple regular expressions.

Feature	Description	Type
Non-ASCII Characters	How many non-ASCII characters (characters that are defined by Unicode but not ASCII) are in the tweet?	Numeric
URLs	How many URLs (“http://” followed by some string) are in the tweet?	Numeric
Quotations	How many Twitter usernames (“@” followed by a username) are in the tweet?	Numeric
Colons	How many hashtags (“#” followed by a word or phrase) are in the tweet?	Numeric
Numbers	How many numeric characters (0-9) are in the tweet?	Numeric
Periods	How many periods (not including period characters that are part of ellipses) are in the tweet?	Numeric
Ellipses	How many ellipses (“…”) are in the tweet?	Numeric
Emoticons	Are there more positive-afffect or negative-affect emoticons (based on a list of 41 positive emoticons and	Categorical {POS, NEG,

	31 negative emoticons) in the tweet? If there are more positive than negative emoticons, the feature value is POS. If there are more negative emoticons, the feature value is NEG. If there are an equal number of positive and negative emoticons or no recognized emoticons, the feature value is OTHER.	OTHER}
Money	How many dollar signs (“$”) are in the tweet?	Numeric
Dates	How many times is a date (any number between 1900-2099 and/or any Gregorian calendar month name) mentioned in the tweet?	Numeric
Capitalized Words	How many words in the tweet have their first letter capitalized?	Numeric
Capital Letters	How many capital letters are in the tweet?	Numeric
All-Caps Words	How many words in the tweet have all of their letters capitalized?	Numeric

Table (8). Special character and punctuation content-based features.

At first glance, these features appear to capture small, insignificant information about punctuation and grammar, but these features capture very important details about the content of each tweet. For example, the number of quotations indicates whether or not a tweet is quoting another person or another news source. Additionally, the number of capitalized words provides information about whether a tweet is a sentence or a title. The presence of question marks and/or exclamation points is also a strong indicator of the tone of a tweet. Similarly, the types of emoticons in the tweet are strong indicators of its affective content, and the presence of either positive or negative emoticons can indicates that it has an informal tone.

After these features were extracted, the first preprocessing step was performed. In this preprocessing phase, URLs, non-ASCII characters, and extra whitespace were removed. URLs are defined as a string beginning with “http://”, which is sufficient to remove URLs from tweets as Twitter appends all URLs with “http://” by default when a tweet is published, even if this prefix is not shown. Non-ASCII characters are characters that are defined in the UTF-8 standard but not in the ASCII standard. Thus, these characters are a subset of the Unicode character set, but not the ASCII character set. Finally, extra whitespace is defined as tabs (“\t”), newlines

("\n"), carriage returns ("\r"), or more than one consecutive space (" ") character. These elements were removed using regular expressions. For example, consider the following tweet:

> Support my mum, who is doing a marathon next week to raise money for breast cancer care charities.
> Thank you :) ♥ http://t.co/ldXP86E7CR
>
> After preprocessing, this tweet contained the following text:
>
> Support my mum, who is doing a marathon next week to raise money for breast cancer care charities. Thank you :)

Figure (11). Example of initial phase of tweet preprocessing.

Note that the URL and Unicode heart have been removed, as has the newline between the two sentences and the extra spaces before the emoticon.

After this initial preprocessing phase, two features related to Twitter usernames and hashtags were extracted. These features are enumerated in the table below.

Feature	Description	Type
Username Mentions	How many Twitter usernames ("@" followed by a 1-15 character username) are in the tweet?	Numeric
Hashtags	How many hashtags ("#" followed by a word or phrase) are in the tweet?	Numeric

Table (9). Username and hashtag content-based features.

The username feature captures the some of the content and intended message of a tweet—for example, was the tweet directed at a particular person or group of people, or was it

intended for a general audience? Similarly, the presence of a hashtag can indicate the topic of a tweet, as hashtags are frequently used on Twitter to summarize and denote topics [33].

Following this extraction step, the second preprocessing step was performed: Twitter usernames were removed from the corpus, and all of the text was converted to lowercase text. For example, the tweet shown above was converted to the following text after this preprocessing step:

> support my mum, who is doing a marathon next week to raise money for breast cancer care charities. thank you :)

Figure (12). Example of second phase of tweet preprocessing.

Note that the uppercase letters have been removed and replaced with lowercase letters. Next, six more features related to the tone and emotional content of the text were extracted, using wordlists of tone and emotion words used Qiu (2011) to classify post sentment on an online cancer forum [102].

Feature	Description	Type
Exclamation Points	How many exclamation points (“!”) are in the tweet?	Numeric
Question Marks	How many question marks (“?”) are in the tweet?	Numeric
Slang Words	How many Internet slang words (based on a list of 200 slang words such as “omg, “lmfao, and “lol”) are in the tweet?	Numeric
Positive Words	How many positive-affect words (based on a list of 2005 positive-affect words, such as “happy”, “thrive”, and “zest”) are in the tweet?	Numeric
Negative Words	How many negative-affect words (based on a list of 4775 negative-affect words, such as “hate”, “upset”, and “vile”) are in the tweet?	Numeric
Pos/Neg Word Ratio	What is the ratio of positive words to negative words (# positive words divided by # negative words) in the tweet?	Numeric

Table (10). Tone-related content-based features.

After this extraction step, the final text preprocessing step was performed. Hashtags (“#”), numbers (0-9), and any other non-alphabetic (not “a” through “z”) characters were removed. Words containing apostrophes were combined into one word. For example, the string “don’t” would become “dont”. Words joined by m-dashes were split into two words. For example “fat-free” would become “fat free.” As an illustration, the tweet shown above was converted to the following after this preprocessing step:

support my mum who is doing a marathon next week to raise money for breast cancer care charities thank vou

Figure (13). Example of third phase of tweet preprocessing.

Note that the emoticon and period have been removed from the tweet. After this final preprocessing step, the following features based on word lists were extracted:

Feature	Description	Type
Self Words	How many words referring to self (based on a list of 10 words, such as “I”, “my”, and “our”) are in the tweet?	Numeric
Family Words	How many words referring to family (based on a list of 46 words, such as “mother”, “brother”, and “cousin”) are in the tweet?	Numeric
Personal Pronouns	How many personal pronouns (based on a list of 17 personal pronouns, such as “he”, “herself”, and “they”) are in the tweet?	Numeric
News Words	How many words commonly found in news headlines about cancer (wordlist: “against”, “money” “news”, “research”, “women”) are in the tweet?	Numeric
Conversation Words	How many conversational words (using a list of 24 words such as “”) are in the tweet?	Numeric
English Names	How many English-language names (based on a list of 5163 names) are in the tweet?	Numeric
Average Word Length	What is the average word length (sum of word length of each word divided by number of words) of the	Numeric

	tweet?	
Post Length	How many words are in the tweet?	Numeric
Words Before Keyword	How many words occur before the keyword in the tweet? (E.g. for the tweet "My aunt has breast cancer" the feature value is 3.)	Numeric
Negation Words	How many negation phrases (based on a list of 286 negation phrases, such as "no", "not", "absence of") are in the tweet?	Numeric

Table (11). Word list-based content-based features.

These features provide significant insight into the content and tone of the tweet. For example, the presence of self, personal, and conversation words are strong indicators of the topic of each tweet, and the presence of positive and negative words is a strong indicator of its sentiment. Similarly, the presence of conversation words, family words, names, and news words can indicate whether a tweet is a headline or summary of a news article or a more informal personal message.

After this feature extraction step, additional linguistic features were extracted using natural language processing methods. First, each word in each tweet was assigned Part of Speech tags using a Part of Speech tagging algorithm. The Stanford NLP Core library, a popular natural language processing library that includes a Part of Speech tagger and parser, was used for this task. First, each tweet is converted into a sequence of Part of Speech tags using the Stanford tagger. The image below shows the transformation from plaintext to Stanford NLP tags for a single tweet:

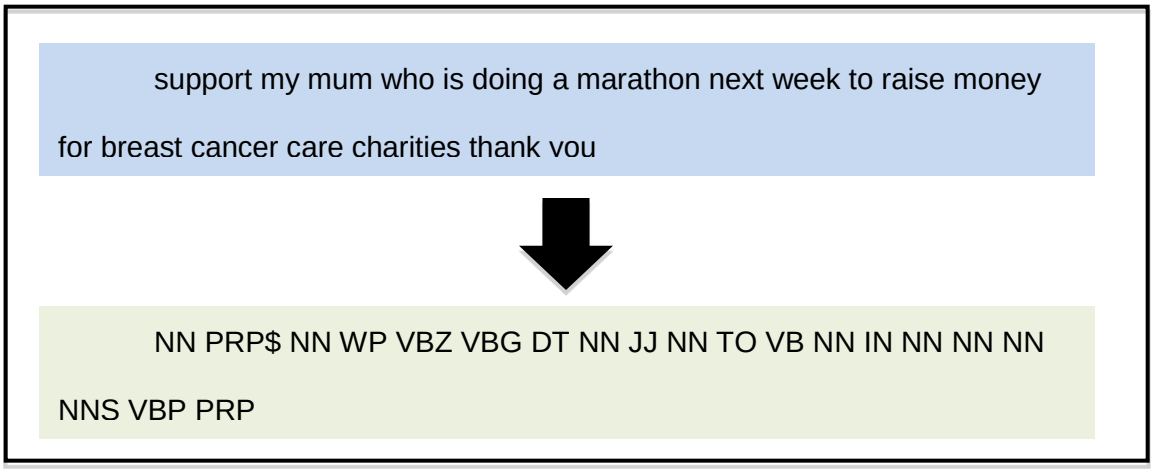

Figure (14). Part of speech tagging for tweets.

The full list of Stanford tags is shown in the figure below:

CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition/subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Figure (15). Stanford NLP Core part of speech tag descriptions.

While the Stanford tagger is not always accurate, especially on short, informal texts containing many slang words, it can provide a useful estimation of the grammatical structure of

each tweet, including whether the tweet was written in past or present tense. After each tweet was tagged, the following part of speech-related features were extracted from the POS tag strings:

Feature	Description	Type
% Verbs	What percent of the words tweet are tagged as verbs? (The number of VB, VBG, VBN, VBP, and VBZ tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
Majority Verb Tense	Do past tense or present tense verbs appear more frequently in the tweet? If past tense verb tags (VBD, VBG, VBN) are more common than present tense verb tags (VB, VBP, VBZ) the feature value is PAST. Otherwise, the feature value is PRESENT.	Categorical {PAST, PRESENT}
% Singular Nouns	What percent of the words tweet are tagged as singular nouns? (The number of NN tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
% Plural Nouns	What percent of the words tweet are tagged as plural nouns? (The number of NNS tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
% Proper Nouns	What percent of the words tweet are tagged as proper nouns? (The number of NNP and NNPS tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
% Adjectives	What percent of the words tweet are tagged as adjectives? (The number of JJ, JJS, and JJR tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
% Possessive Pronouns	What percent of the words tweet are tagged as personal or possessive pronouns? (The number of PRP and PRP$ tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
% Interjections	What percent of the words tweet are tagged as interjections? (The number of UH tags in the tweet is recorded and divided by the total number of tokens in the tweet.)	Numeric
POS of Word Before Keyword	What is the POS tag of the word directly before the keyword used to collect it? If the POS tag is NN, NNS, NNP, or NNPS, the feature value is NOUN. If the POS tag is VB, VBD, VBN, VBP, or VBZ, the feature value is VERB. If the POS tag is JJ, JJR, or JJS, the feature value is ADJECTIVE. If the POS tag is PRP or PRP$, the feature value is PREPOSITION. If the POS tag is DT, the feature value is DETERMINER. If the keyword is the first word in the tweet or the word	Categorical {NOUN, VERB, ADJECTIVE, PREPOSITION, DETERMINER, OTHER}

	before the keyword has a POS tag other than the ones mentioned, the feature value is OTHER. (E.g. for the tweet "My aunt has breast cancer" the feature value is VERB.)	
POS of Word After Keyword	What is the POS tag of the word directly before the keyword used to collect it? If the POS tag is NN, NNS, NNP, or NNPS, the feature value is NOUN. If the POS tag is VB, VBD, VBN, VBP, or VBZ, the feature value is VERB. If the POS tag is JJ, JJR, or JJS, the feature value is ADJECTIVE. If the POS tag is PRP or PRP$, the feature value is PREPOSITION. If the POS tag is DT, the feature value is DETERMINER. If the keyword is the last word in the tweet or the word before the keyword has a POS tag other than the ones mentioned, the feature value is OTHER. (E.g. for the tweet "Breast cancer treatments" the feature value is NOUN.)	Categorical {NOUN, VERB, ADJECTIVE, PREPOSITION, DETERMINER, OTHER}

Table (12). NLP content-based features.

These features capture a number of syntactic properties of each tweet, which often reflect its content. For example, the majority verb tense of a tweet indicates whether it is describing a past or current event. Additionally, the relative number of nouns, adjectives, verbs, and other Parts of Speech will vary depending on whether the sentence is a detailed description of an event or a simple list of persons or medications in the breast cancer domain.

After these features were extracted from the Part of Speech tags, the final preprocessing step was performed. This step is the conversion of the text from each tweet into a parsing tree representation of the text. The Stanford NLP Parser, which is part of the Stanford NLP Core library [125], was used for this task. The parser generates both a parse tree, which is a tree structure representing the sentence, and a list of relationships between words, called typed dependencies. Each typed dependency has a particular type, which corresponds to a particular syntactic construct, such as the use of auxiliary verbs, possessive verbs, and direct objects. For example, the typed dependencies of the tweet above are shown below:

```
poss(mum-3, my-2)
dobj(support-1, mum-3)
nsubj(doing-6, mum-3)
aux(doing-6, is-5)
rcmod(mum-3, doing-6)
det(marathon-8, a-7)
dobj(doing-6, marathon-8)
amod(week-10, next-9)
tmod(doing-6, week-10)
aux(raise-12, to-11)
xcomp(doing-6, raise-12)
nsubj(thank-19, money-13)
nn(charities-18, breast-15)
nn(charities-18, cancer-16)
nn(charities-18, care-17)
prep_for(money-13, charities-18)
ccomp(raise-12, thank-19)
dobj(thank-19, you-20)
```

Figure (16). Typed dependencies for example tweet.

Each typed dependency represents a relationship between two words in the tweet. The type of the relation is based on the parts of speech of the two words and their location within the sentence. For example, the first relation signifies that the second word in the tweet, “my,” is a possession modifier for the third word, “mum.”

Next, the Stanford NLP Parser is used to generate a syntax tree for each tweet. The parsing tree for the tweet above tweet is as follows:

```
(ROOT
  (S
    (VP
      (VB support)
      (NP
        (NP
          (PRP$ my)
          (NN mum))
        (SBAR
          (WHNP
            (WP who))
          (S
            (VP
              (VBZ is)
              (VP
                (VBG doing)
                (NP
                  (DT a)
                  (NN marathon))
                (NP
                  (JJ next)
                  (NN week))
                (S
                  (VP
                    (TO to)
                    (VP
                      (VB raise)
                      (S
                        (NP
                          (NP
                            (NN money))
                          (PP
                            (IN for)
                            (NP
                              (NN breast)
                              (NN cancer)
                              (NN care)
                              (NNS charities))))
                        (VP
                          (VB thank)
                          (NP
                            (PRP you)))))))))))))))
```

Figure (17). Stanford NLP parse tree for example tweet.

Even without examining the particular tags in this tree, it is apparent that the parser has correctly identified the hierarchical structure of the sentence, including the fact that “thank you” is divorced from the rest of the content, and created a tree accordingly. The parser also recognized

that the phrase “breast cancer care charities” is a compound noun phrase, which is a relatively difficult distinction to make. The parsing tree and the typed dependency relations are then used in the final feature extraction phase. Features based on these representations are designed to record properties of individual syntactic structures found in the text, which provides information about the sentence complexity and linguistic style of each tweet. These features are enumerated in the table below:

Feature	Description	Type
Parse Tree Depth	What is the maximum depth of the parsing tree?	Numeric
Depth of Keyword	How deep in the parsing tree is the relation containing the keyword used to collect the tweet?	Numeric
Relation Types	How many distinct types of syntactic relationships are present in the tree?	Numeric
Nominal Subject	How many nominal subject phrases, or phrases in which a noun phrase is the syntactic subject (e.g. “The baby is cute”), are in the tweet?	Numeric
Noun Compound Modifier	How many noun compound modifier phrases, or phrases in which a noun modifies another noun (e.g. “Oil price futures”), are in the tweet?	Numeric
Possession Modifier	How many possession modifier phrases, or phrase in which a possessive determiner changes the meaning of a noun (e.g. “their offices”), are in the tweet?	Numeric
Direct Object	How many direct objects, or phrases in which a noun is the object of a verb (e.g. “She gave me a raise”), are in the tweet?	Numeric
Clausal Complement	How many clausal complements, or complimentary phrases with their own subjects (e.g. “I am certain that he did it”), are in the tweet?	Numeric
Adjectival Modifier	How many adjectives are used to modify nouns (e.g. “Sam eats red meat”) in the tweet?	Numeric
Determiner	How many noun phrases containing determiners (e.g. “The man is here”) are in the tweet?	Numeric
Auxiliary	How many auxiliary verbs, or non-main verbs accompanying a main verb in a phrase (e.g. “He should leave”), are in the tweet?	Numeric
Copula	How many copular verbs, or verbs using “to be” to link themselves to a subject (e.g., “James is honest”), are in the tweet?	Numeric
Relative Clause Modifier	How many relative clause modifiers, or phrases containing a verb that change the meaning of a noun (e.g. “I saw the book you bought”), are in the tweet?	Numeric

Table (13). Linguistic properties and sentence complexity-based content-based features.

Like the part of speech-based features discussed above, these features are useful to the extent that descriptions of personal experiences tend to contain certain types of syntax or grammatical patterns. For example, the depth of the parsing tree and the number of unique typed dependencies are both indicators of the complexity of a sentence. Others, such as Direct Object and Adjectival Modifier, identify the presence particular sentence structures that may help indicate the content of a tweet.

After the feature extraction process, the feature vectors for each tweet are concatenated into a feature matrix. Each column of this matrix corresponds to a tweet and each row corresponds to a feature. This matrix represents the feature space for the original dataset of tweets.

Chapter 6

Personal Experience Classification

6.1 Metrics

In order to validate the performance of a supervised learning algorithm on a particular dataset, a statistical procedure called cross-validation is used. In this procedure, a dataset is divided into *k* groups. One of the *k* groups is used as a testing set, and the other $k - 1$ groups are used as the training set. This procedure is repeated for each of the *k* groups, such that each group is used as the testing set exactly once. The output of a cross-validation experiment can be viewed as a confusion matrix:

Predicted Class			
Positive	Negative		
TP True Positive	**FN** False Negative	Positive	Ground Truth
FP False Positive	**TN** True Negative	Negative	

Figure (18). Example confusion matrix for a supervised learning experiment.

Each cell of the confusion matrix represents a possible classification outcome for each data point. The rows of the matrix correspond to the instance's true class value, and columns correspond to its predicted class. For example, if the true of an instance is Positive and the supervised learning model predicts that it is Positive, it is placed in the first row and first column, which corresponds to True Positive. If the true class of the instance is Negative and the supervised learning model predicts that it is Positive, it is placed in the first row and second column, which corresponds to False Positive.

From this confusion matrix, it is possible to define several metrics to evaluate the performance of a classifier. In the machine learning community, there are three prominent measures of classification performance: Accuracy, F-Measure, and ROC area. The first and most intuitive measure of classifier performance is accuracy, which is calculated as follows:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Intuitively, accuracy is the number of correctly classified instances divided by the total number of instances. However, accuracy is not sufficient measure of performance, as it is possible to simple assign all instances to a single class and still obtain high accuracy if the dataset is unbalanced. Therefore, other measures such as Precision, Recall, and F-Measure are also important.

Precision is the percentage of instances classified positive that are True Positives, as is calculated as follows:

$$\text{Precision} = \frac{tp}{tp + fp}$$

A related measure is recall, which is the percentage of positive instances that are classified as positive.

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-Measure, also known as F1 score, combines both Precision and Recall into one measurement. F-Measure is the harmonic mean of Precision and Recall, and is calculated as follows:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Another measure of performance is ROC area, which measures the ability of a classifier to discriminate between positive and negative instances. The receiver operating characteristic, or

ROC, is computed by varying the threshold used to classify an instance as positive or negative and graphing the true positive rate and false positive rate as this threshold varies. For example, the graph below shows the ROC curve for a Logistic Regression classifier on the breast cancer dataset, using the content and context-based features.

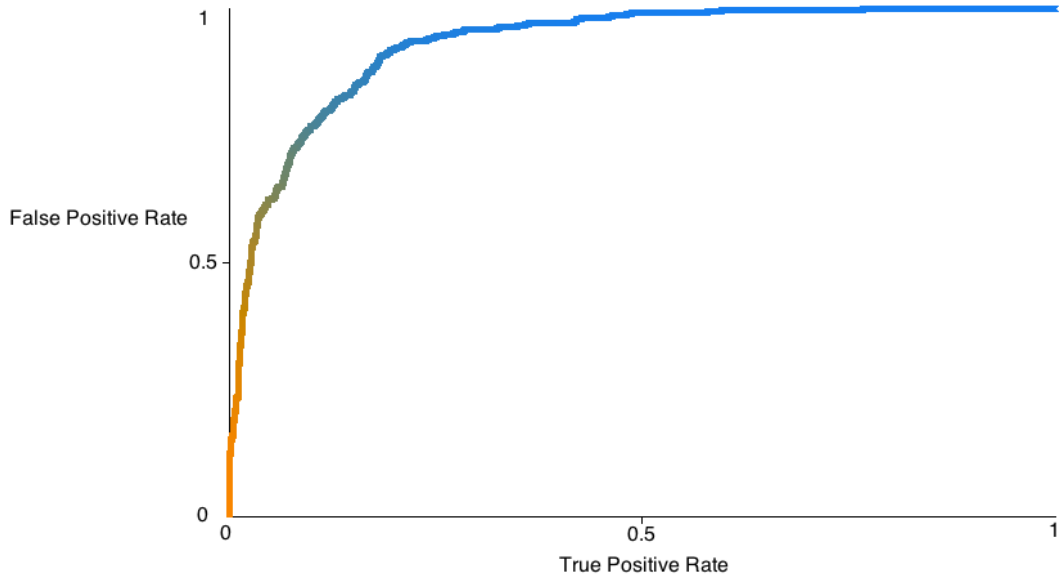

Figure (19). ROC area for logistic regression classifier on breast cancer dataset, using content and content-based features.

The ROC area is defined as the area under this curve, which will be between 0 and 1; a value of 1 represents a perfect classifier, and a value of 0 represents a classifier that is always wrong.

It is also important to study the effectiveness of different features, as not all features provide the same amount of information about a dataset's classes. This can be estimated using Information Gain, which measures the amount of information provides about the probability distribution of a dataset's classes. The Information Gain of a feature is the decrease in entropy of

this marginal probability distribution from the original distribution, conditioning on the feature. Given a feature *T* and probability distribution of classes *a*, this is calculated as follows:

$$IG(T, a) = H(T) - H(T|a)$$

Note that this is equivalent to computing the mutual information between the feature *T* and the distribution of classes [41]. In the sections below, these metrics are applied to evaluate classification of the four Twitter datasets and evaluate the effectiveness of each feature.

6.2 Classification Procedure and Results

In order to evaluate the effectiveness of the context-based and content-based features, twelve different classification algorithms were applied to the four datasets using k-fold cross-validation. This set of twelve classifiers includes several different types of supervised learning models, including Bayesian methods, boosting, decision trees, and Support Vector Machines. As shown in the table below, each classifier has a different set of parameters that can be modified.

Classifier	Parameters	Range
AdaBoost	Number of boosting iterations	50 – 250
Bagging	Number of bags Bag size (as % of training set)	10 – 50 50 – 100
Bayesian Network	N/A	N/A
J48 Decision Tree	Confidence factor for pruning Minimum leaf node size	0.1 – 0.5 1 – 20
K-Nearest Neighbors	Number of neighbors (K)	1 – 20
Logistic Model Tree	Minimum leaf node size	5 – 35
Logistic Regression	N/A	N/A
LogitBoost	Number of boosting iterations Shrinkage parameter	10 – 50 0.1 – 1
Naïve Bayes	N/A	N/A
RandomForest	Number of random trees Number of features per tree	5 – 20 1 – 20
SVM (linear kernel)	Cost of misclassification (C)	0.1 – 10
SVM (RBF kernel)	Cost of misclassification (C) Gamma	0.1 – 10 0.1 – 1

Table (14). Parameters for supervised learning classifiers.

In order to optimize the parameters of each classifier during supervised learning, a nested k-fold cross-validation procedure was used for validation. This procedure is as follows:

First, the dataset was split into *n* equal-size groups, or folds, which are used to perform *n* supervised learning experiments to classify part of the dataset. In each fold, one of the *n* groups is used as the testing set, and the other ($n - 1$) groups are used as the training set. Note that this is the standard procedure for cross-validation. However, before each classifier is built on one of the training sets, *n*-fold cross-validation is performed on that training set in order to optimize the parameter values for each classifier. This cross-validation is used to perform a grid search of parameter values within the ranges shown in the table above, and the best parameter values (based on some criteria) are selected. Then, the classifiers are trained on the training set using these parameter values, and then each model is applied to the testing set. This procedure is repeated for each of the *n* folds, so that optimal parameter values are learned for each of the *n* training sets. Cross-validation was also used on each fold for feature selection via the backtracking feature selection algorithm prior to the parameter optimization phase.

First, in order to assess the usefulness of the content-based and context-based features described above, it is necessary to establish a baseline classification accuracy using more standard feature extraction methods, namely the n-gram feature extraction model. After the text preprocessing phases discussed above, 1-grams and 2-grams were extracted from each dataset; only n-grams that occurred in at least 5 posts were considered.

To compare the classification accuracy of the n-gram model to the content-based and context-based features, six supervised learning experiments were conducted using nested 5-fold cross-validation. Within each fold, 5-fold cross-validation was performed on the training set to optimize the value of each classifier parameter. The parameters that maximized the accuracy of

the classifier were selected as the best parameters for that fold. The sets of features used in these six experiments are as follows:

1. 1-grams and 2-grams, with backtracking feature selection on each fold.
2. Context-based features only, with no feature selection.
3. Content-based features only, with no feature selection.
4. Context-based and content-based features, with no feature selection.
5. Context-based and content-based features, with backtracking feature selection on each fold.
6. Context-based and content-based features and 1-grams and 2-grams, with backtracking feature selection on each fold.

These experiments were conducted using the 12 classifiers enumerated above. The full results for each classifier on each dataset are shown in Appendix A. The average performance for each set of features (aggregating across all 12 classifiers), with classifier parameters optimized to minimize the error rate, is shown in the graph below

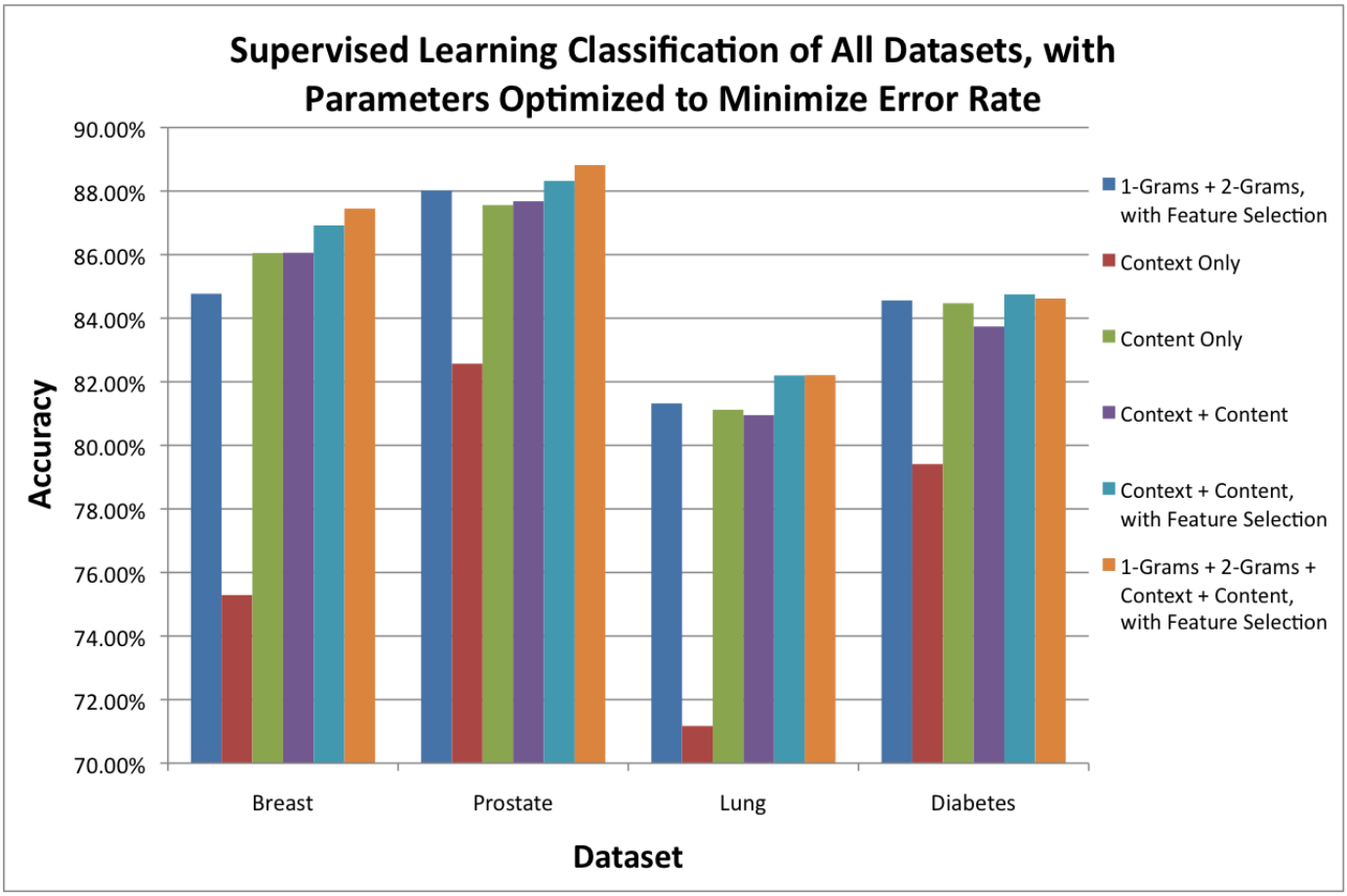

Figure (20). Classification accuracies for supervised learning experiments on all datasets, using six feature spaces, with parameters optimized to minimize error rate.

This comparison of featureset performance reveals a number of interesting properties about these dataset and the effectiveness of various features. First, we see that our Context + Content features are competitive with the traditional text mining featureset, the bag of words model—in all four cases, the Context + Content featureset (with feature selection) outperformed the n-gram model. This suggests that there is indeed valuable information hidden in the structure of these short texts posts that can be captured with the new kind of feature engineering discussed in the previous chapter. In addition, we observe that combining this approach with the traditional approach yields a classifier that is even more powerful, at least in the case of the three cancer datasets.

However, it is also obvious from the graph that the context-based features are vastly inferior to the content-based features; for the breast cancer and lung cancer datasets, the classifier

that was only given these features is just as powerful as simply guessing the majority class 100% of the time. Nevertheless, in all four cases the Context + Content featureset is more powerful than just the Content featureset. This implies that the Context features do contain some useful information that is uncorrelated with the content-based features.

Next, we observe that there are large differences in classification accuracy between these four datasets. The breast and prostate datasets can be classified with high accuracy; the lung cancer and diabetes datasets lag slightly behind. In the case of the diabetes dataset, it is possible that the large number of posts containing jokes and other content that the NLP-based features cannot distinguish from genuine descriptions of personal experiences is creating this discrepancy. However, this hypothesis does not explain the similar drop in performance for the lung cancer dataset.

It is also worth noting that feature selection produced a noticeable increase in accuracy for all four datasets. In all four cases, the Context + Content featureset with feature selection was superior to the same featureset with out feature selection.

As the graph shows, the average accuracy on all datasets for the sets of featuresets with backtracking feature selection is well above 80%. However, while the accuracy and ROC area for these supervised learning models is high, the F-Measure is very low. This is because some of the classifiers are classifying most instances as Impersonal (the majority class) in order the increase accuracy at the expense of Precision and Recall. For example, consider the confusion matrix for the Logistic Model Trees classifier on the breast cancer dataset, with Content + Context features, which was the highest accuracy classifier for that dataset and feature set:

Predicted Class			
Personal	Impersonal		
283	167	Personal	Ground Truth
93	1616	Impersonal	

Figure (21). Confusion matrix for LMT classifier on breast cancer dataset, with Content + Context features, with parameters optimized to minimize error rate.

As shown in this confusion matrix, the unbalanced nature of this dataset leads to a large number of false negatives when classifier parameters are selected to minimize error rate. In this example, more than one-third of the Personal posts are misclassified as Impersonal.

This lack of discriminatory power can also be seen for all six featuresets when we examine the f-measure for these six sets of experiments. This is shown in the figure below:

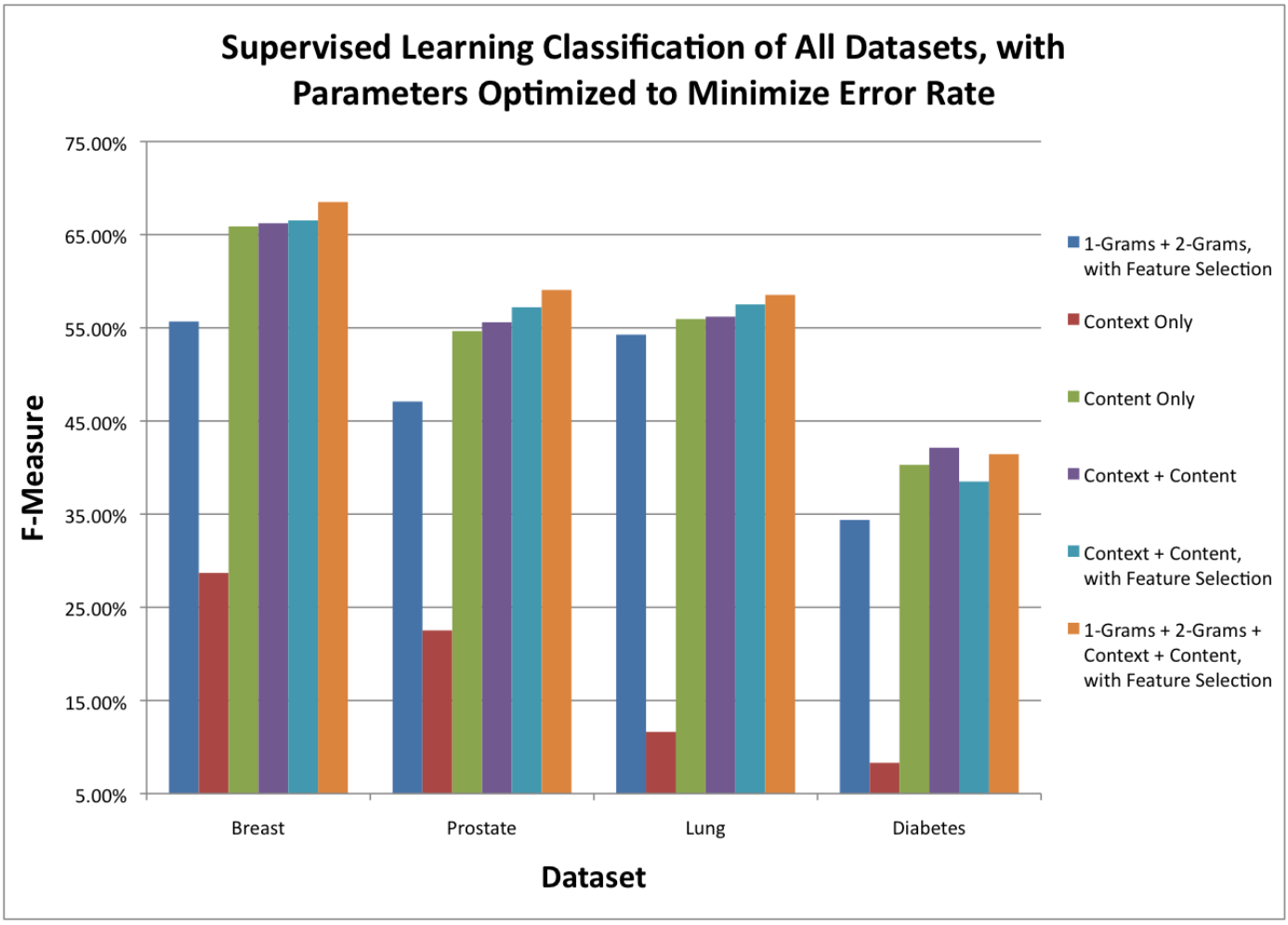

Figure (22). Classification accuracies for supervised learning experiments on all datasets, using six feature spaces, with parameters optimized to minimize error rate.

As the graph shows, despite the high accuracies reported above, the f-measures for these experiments are less impressive. In particular, the Context classifiers simply assign nearly all of their inputs to the (Impersonal) majority class, resulting in poor precision and recall. The other sets of classifiers perform slightly better, but in general none of these models are very robust.

Interestingly, the f-measures for the diabetes dataset are lower than the lung cancer dataset, which confirms the hypothesis discussed above that this dataset's Impersonal posts'

informal tone makes them harder to classify compared to impersonal posts from the other three datasets.

In order to improve precision and recall for the Personal class, the above experiments were repeated, but classifier parameter values were optimized to maximize f-measure (rather than accuracy) using cross-validation. The graph below shows the average performance of all classifiers for each of the four datasets, when optimizing classifier parameters to maximize F-Measure instead of minimizing error rate:

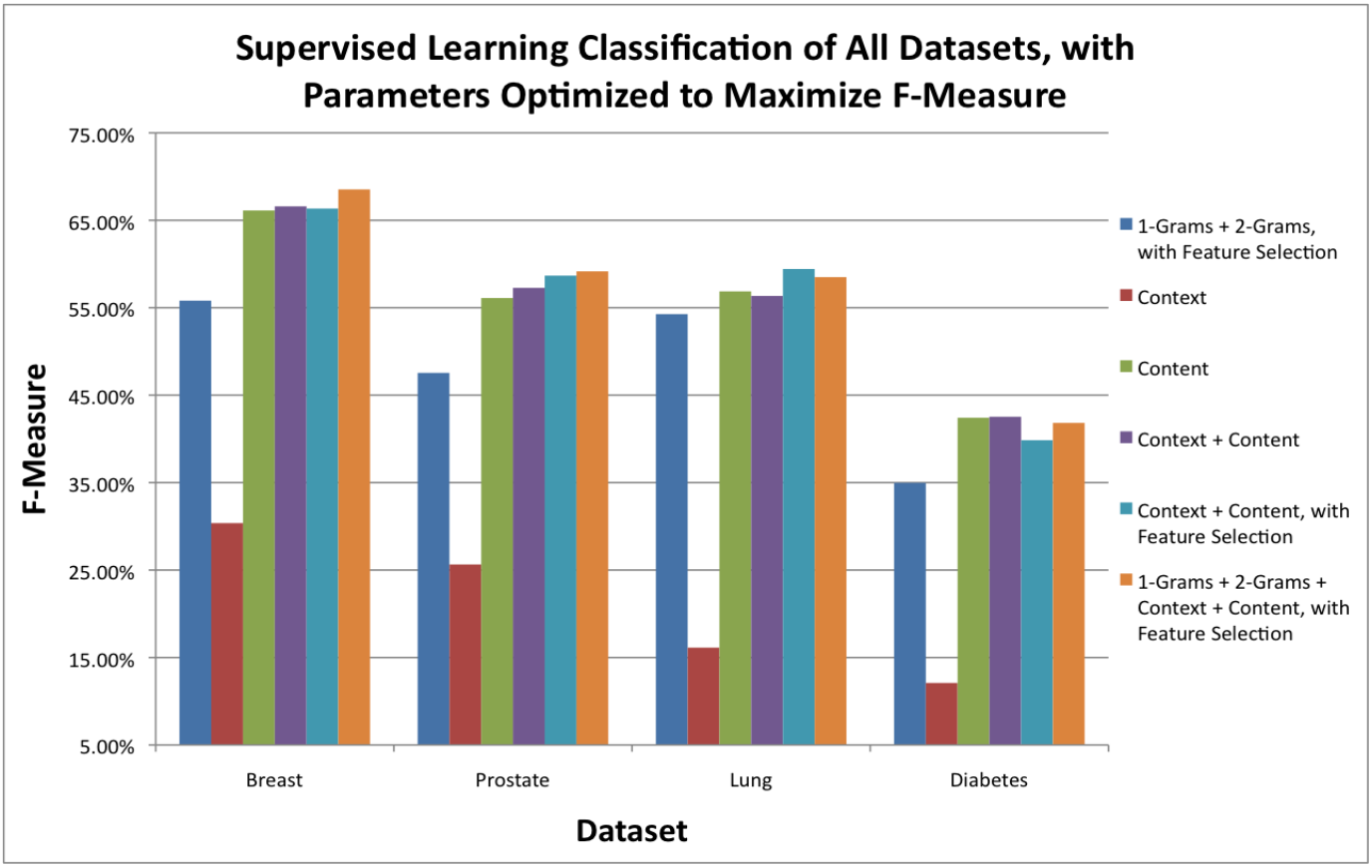

Figure (23). Classification F-Measures for supervised learning experiments on all datasets, using six feature spaces, with parameters optimized to maximize f-measure.

In this round of experiments, we see that the f-measures for most of the classifiers have slightly increased, most notably the Context featureset classifiers, which were very poor when the classifiers were optimized to decrease error rate. Overall, however, the effectiveness of the six different sets of feature is essentially the same as the previous set of experiments; even the differences in classification f-measures between the four datasets is the same as in the set of experiments above. As before, the classifiers for the diabetes dataset have particularly poor discriminatory power. Note that there was no changes to the underlying featurespace; the classifiers' precisions and recalls have been improved because different sets of parameters were chosen.

This improvement is also reflected in the confusion matrices for individual experiments. Above, we discussed the confusion matrix for the Logistic Model Trees classifier on the breast cancer dataset, using the Content + Context feature, when the parameters of each classifier were optimized for accuracy. The equivalent confusion matrix when the parameters of each classifier were optimized to maximize f-measure is shown below:

Predicted Class			
Personal	Impersonal		
232	128	Personal	Ground Truth
74	1293	Impersonal	

Figure (24). Confusion matrix for LMT classifier on breast cancer dataset, with Content + Context features, with parameters optimized to increase f-measure.

Note that the number of false negatives has decreased by almost 40, and the number of false positives has decreased by more than 20. Therefore, we can conclude that this classifier is far more robust, as it is better at discriminating between the two classes rather than simply classifying the majority of instances it sees as Impersonal to obtain a low error rate.

In addition, this increase in F-Measure did not lead to a significant decrease in accuracy across all four datasets. This is shown in the figure below, which shows the accuracy for this second set of experiments:

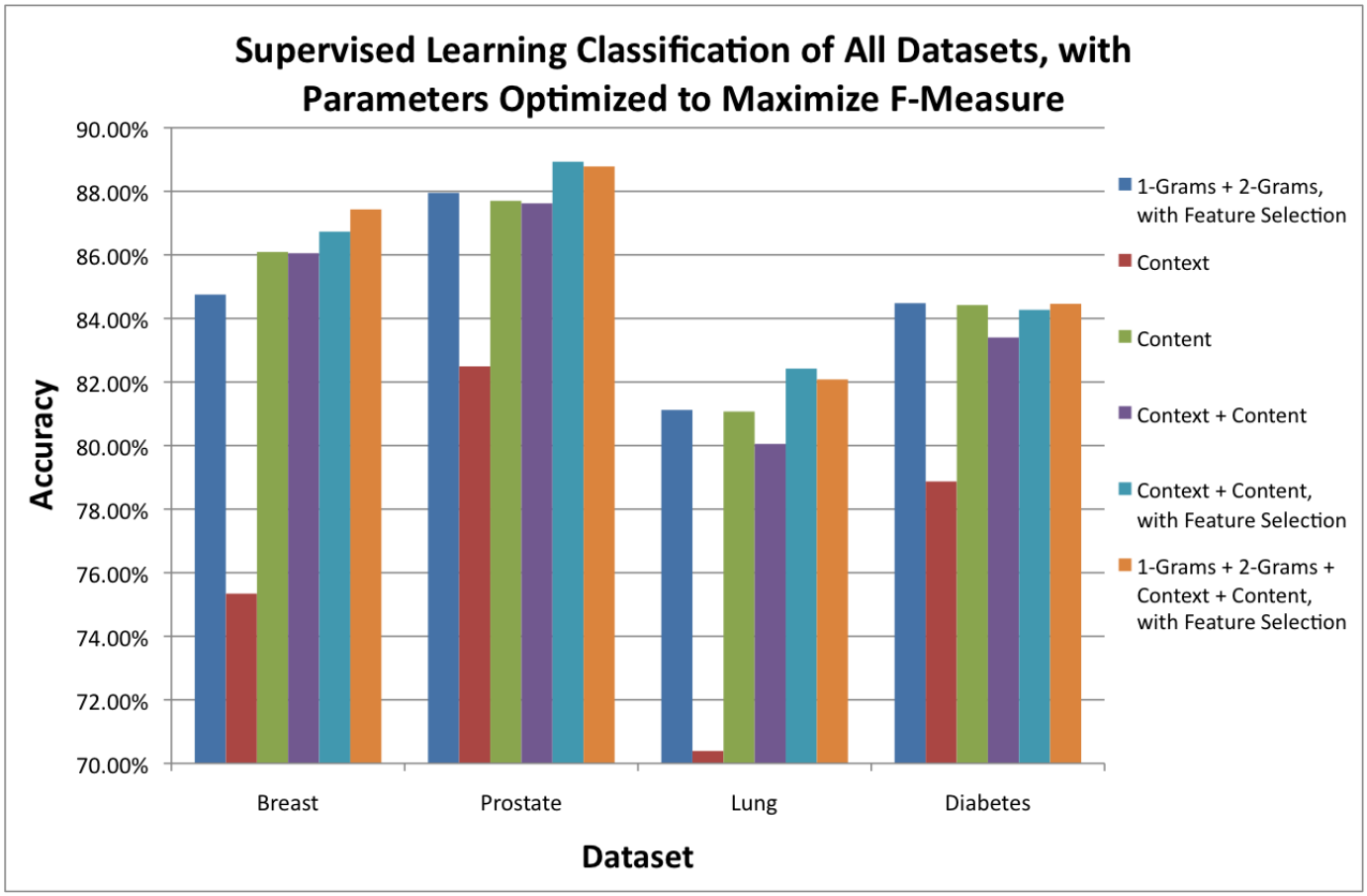

Figure (25). Classification accuracies for supervised learning experiments on all datasets, using six feature spaces, with parameters optimized to maximize f-measure.

The fact that accuracy did not significantly decrease in this round of experiments suggests a simple strategy for optimizing parameters when dealing with unbalanced data: Select classifier parameters to increase the precision and recall rather than the accuracy, as this will lead to a generally more robust classifier.

6.3 Feature Ranking and Analysis

To gain a better understanding of why certain featuresets perform better than others, it is help to rank these features for each dataset. The tables below show the top 15 features for each dataset, as determined by the Information Gain algorithm. Content-based features are in standard text, and context-based features are in red and in italics.

Breast Cancer		**Prostate Cancer**	
Feature	**Info Gain**	**Feature**	**Info Gain**
Self Words	0.22089	URLs	0.16694
URLs	0.16351	Self Words	0.12347
Possessive Pronouns	0.12388	*Posts Favorited*	0.11548
Average Word Length	0.12344	Conversation Words	0.09932
Capital Letters	0.11551	Average Word Length	0.09361
Family Words	0.09471	Numbers	0.08772
Conversation Words	0.09471	Possessive Pronouns	0.08319
Possession Modifier	0.08566	Family Words	0.07952
Posts Favorited	0.08373	Reply	0.07895
Capitalized Words	0.07620	Parse Tree Depth	0.06857
Numbers	0.06906	Depth of Keyword	0.06825
Profile Link	0.04966	Post Length	0.06237
Relation Types	0.04903	Nominal Subject	0.06087
Post Length	0.04634	*Follows*	0.05782
Parse Tree Depth	0.04550	Relation Types	0.05482

Lung Cancer		**Diabetes**	
Feature	**Info Gain**	**Feature**	**Info Gain**
Possessive Pronouns	0.15037	Self Words	0.15037
Family Words	0.14330	URLs	0.14330
Self Words	0.13968	*Posts Favorited*	0.13968
URLs	0.10278	Possessive Pronouns	0.10278
Possession Modifier	0.09313	Average Word Length	0.09313
Post Length	0.08986	Conversation Words	0.08986
Nominal Subject	0.08845	Relation Types	0.08845
Relation Types	0.08759	Capitalized Words	0.08759
Personal Pronouns	0.08253	Parse Tree Depth	0.08253
Singular Nouns	0.07786	Singular Nouns	0.07786
Conversation Words	0.07074	Possessive Pronouns	0.07074
English Names	0.06353	*Profile Link*	0.06353
Parse Tree Depth	0.05146	Capitalized Words	0.05146
Average Word Length	0.05053	Post Length	0.05053

Posts Favorited	0.04878	Clausal Complement	0.04878

Table (15). Information Gain ranking of features.

This ranking suggests several important insights about these features. Firstly, the table shows that several of the novel NLP-based features in this work, including Parse Tree Depth, Relation Types, Depth of Keyword, and other features based the presence of typed dependences are indeed useful for short text classification. This suggests that there is still potential for innovation in terms of the application natural language processing methods to classify short texts, at least in the healthcare domain. This feature ranking also confirms that more traditional POS-based NLP features, such as counts of Singular Nouns and Possessive Pronouns, are still applicable to this domain.

Second, this feature ranking shows that these four domains have many linguistic properties in common—the URL and Self Words features are ranked in the top 5 in all four domains, and the Possessive Pronouns feature is ranked in the top 7 in all four domains. The Family Words, Average Word Length, and Capital Words features are also consistently highly ranked. This is a useful result for researchers attempting to identify personal experiences in multiple domains or performing transfer of learning between domains, as it suggests that the short text medium encourages users to adopt similar writing styles even when talking about different diseases.

Next, it is apparent that content-based features are generally more powerful than context-based features. Content-based features dominate all four of these lists, and of the 13 context-based features described in the previous chapter, only *Post Favorited* and *Profile Link* appear on more than one top-15 lists. Notably, both of these features are indicators of the activity of a Twitter user—a user with more favorites is likely to be more active, and a user with a link in their

profile has put more effort into customizing their Twitter account. This is shown in the graph below:

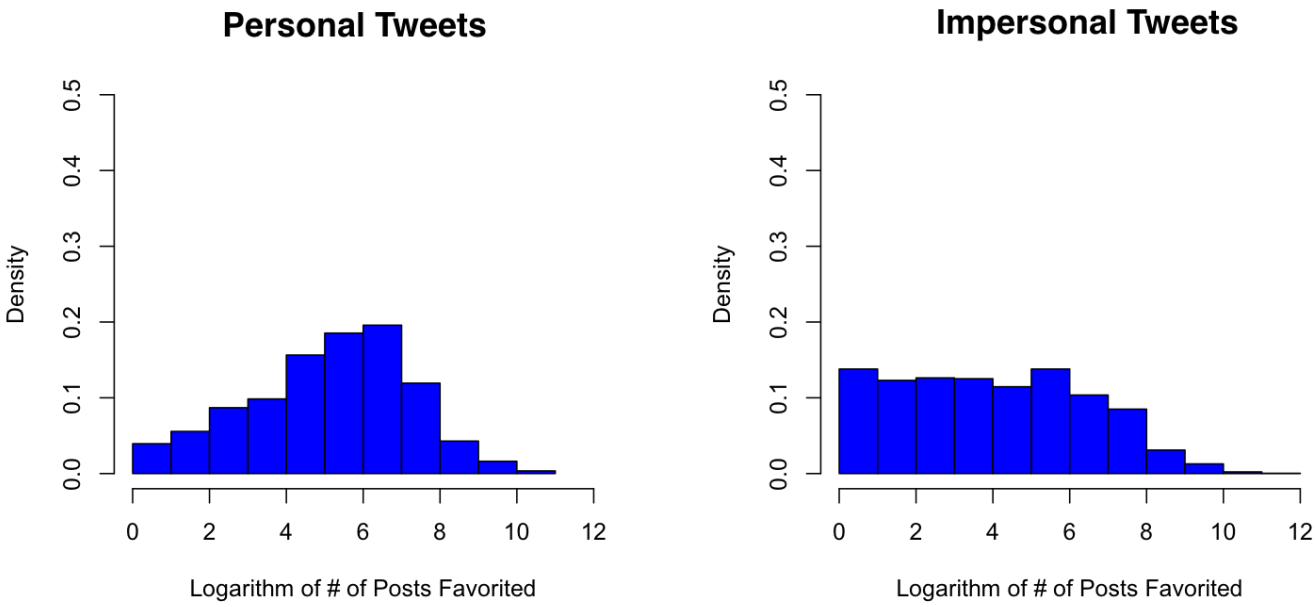

Figure (26). Probability distributions of Posts Favorited feature, on a $\log_2$ scale..

Though the difference between these two distributions is noticeable, it is not particularly strong, which is why this feature, along with the other context-based features, are only weak classifiers. The main reason for the failure of these features is their "long tail" properties within the population of Twitter users. The majority of Twitter users are not very active and do not have a large number of friends or followers, but some users are very active and have thousands or millions of followers. This is apparent when examining the means and standard deviations of the top context-based features across the two classes of tweets. This is shown in the table below.

Feature	Personal Tweets	Impersonal Tweets
# of Friends	Mean: 480.19 SD: 1004.99	Mean: 1685.77 SD: 9028.56

# of Followers	Mean: 1169.56 SD: 8672.83	Mean: 2812.84 SD: 16154.34
# of Favorites	Mean: 897.94 SD: 3065.25	Mean: 485.37 SD: 2459.01
# of Statuses	Mean: 13207.70 SD: 37540.47	Mean: 20790.02 SD: 54619.06
# of Lists User is Mentioned In	Mean: 15.91 SD: 171.85	Mean: 37.37 SD: 174.20

Table (16). Means and standard deviations of some context-based features.

Despite the significant difference in means, the standard deviations are even larger, making it difficult for a classifier to learn from these features. For example, consider the *Posts Favorited* feature, which was described above as having the highest Information Gain—for both classes, the standard deviation is several times larger than the mean. This, again, is because of distribution of followers across Twitter accounts has a very long tail—while approximately 25% of the users in this dataset have 10^6 followers, some have as many as 10^{14}. These distributions are shown in the histograms below, on a $\log_2$ scale:

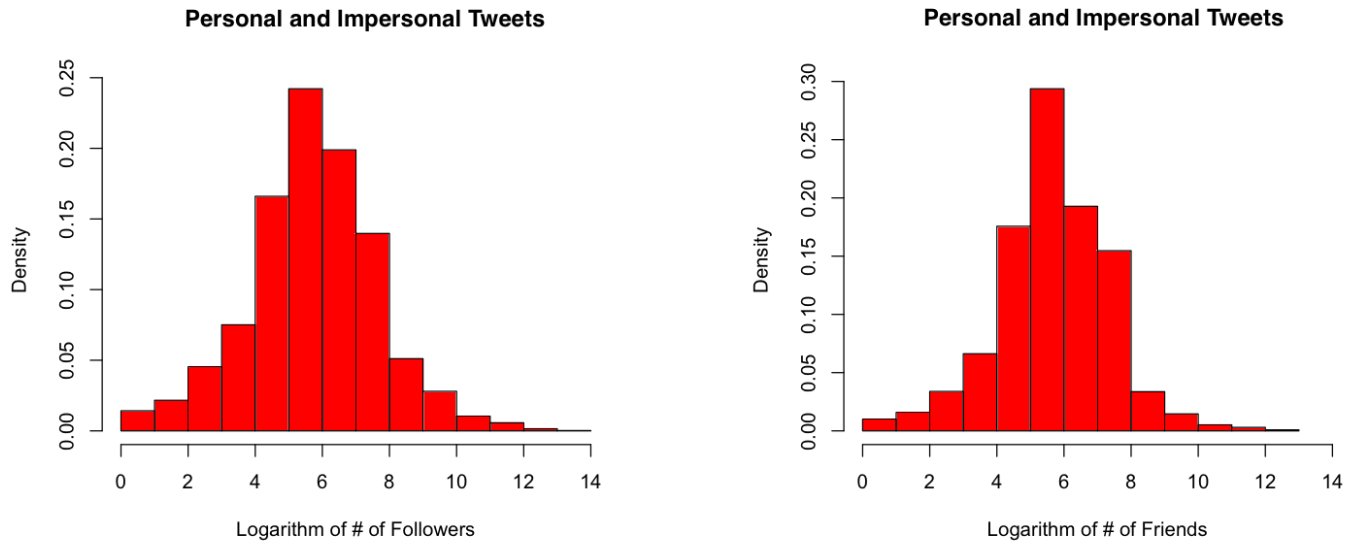

Figure (27). Probability densities of Twitter followers and friends, on a $\log_2$ scale..

These graphs clearly demonstrate the "long tail" properties of the Twitter social graph. Though more than 50% of Twitter users have fewer than 10^6 friends and 8 followers, a small percentage of very popular users have millions of friends and millions of followers.

In the four Twitter datasets used in this thesis, the most Twitter account with the most followers was @TheEllenShow, the Twitter account of celebrity talk show host Ellen DeGeneres, at the time of data collection had more than 20 million followers. By contrast, several hundred user accounts in the dataset had fewer than 100 followers.

Interestingly, this property is observed in both the Followers and Friends distributions, even though these represent two aspects of Twitter's directed graph. Not only do some users have a disproportionately large audience, others have a disproportionately large information diet. These Twitter accounts with large number of friends may represent users who follow other users as a sign of appreciation rather than as a means of following other users' posts, or simply users who do not check their friends list and do not notice the rapid accumulation of content. At the other end of the spectrum, Twitter accounts that do not follow anyone may represent organizations that only use Twitter for broadcasting, or they may simply represent users who prefer to read individual Twitter feeds manually. Whatever the cause, this extremely high variance makes it very difficult for a classifier to learn a useful decision boundary using these contextual features.

Despite these limitations, these features still reveal interesting properties about the population of Twitter users who post about chronic disease. Users who post about personal experiences are less likely to contain a profile with a URL, and have fewer followers on average. However, they follow *more* users on average than users who posted content that did not contain personal experiences. This suggests that users who are more likely to post personal content are more interested in consuming content on Twitter and interacting other users, and less interested in broadcasting to a wide audience. This may be because Twitter accounts that post personal

experiences are more likely to be accounts representing individuals, as opposed to accounts representing organizations. This hypothesis is particularly plausible given that organizations are very likely to include a link to their website in their Twitter profile. Overall, though, due to the ineffectiveness of these features, it is difficult to draw conclusions about the types of users who post about their personal experiences.

Among the content-based features, the most generalizable features are *URLs* and *Self Words*, which are the two most effective features on three of the four datasets. The presence of words referring to self is positively correlated with personal experience, as the presence of URLs is negatively correlated with personal experiences. The probability density functions for these two features on all Personal and Impersonal posts are shown in the figure below:

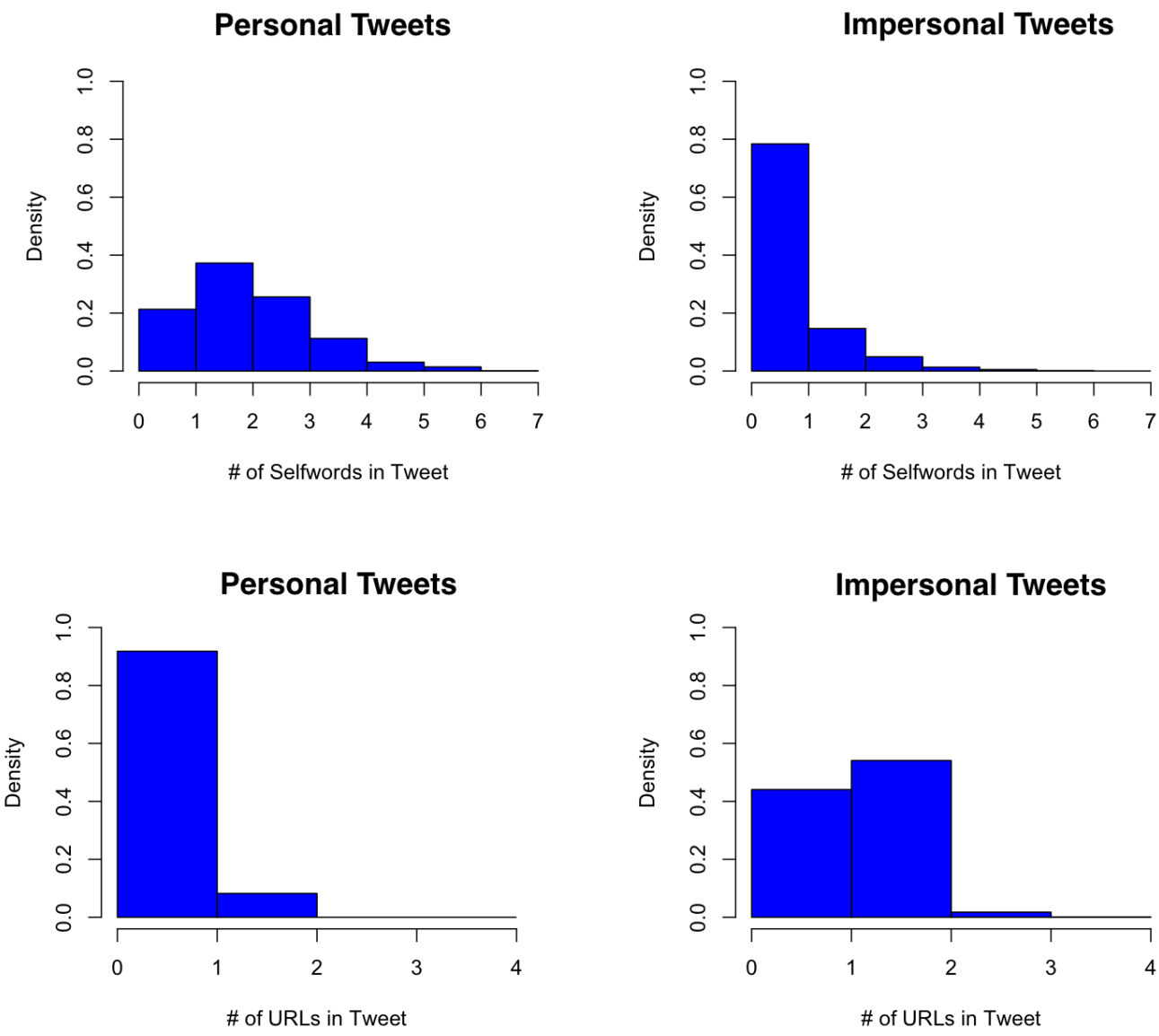

Figure (28). Probability distributions of URL and self word features.

This result is intuitive, as a tweet containing a link presumably describes the contents of that link rather than a personal experience, and tweets containing self-referential words presumably describes the poster. However, analysis of the other top content-based features

reveals some less intuitive properties of posts describing personal experiences. The means and standard deviations between groups for several of the content-based features with the highest Information Gain across all datasets are shown in the table below.

Feature	Personal Tweets	Impersonal Tweets
# of Possessive Pronouns	Mean: 0.09 SD: 0.06	Mean: 0.03 SD: 0.05
Average Word Length	Mean: 4.38 SD: 0.73	Mean: 5.12 SD: 1.00
# of Possession Modifiers	Mean: 0.63 SD: 0.74	Mean: 0.15 SD: 0.41
# of Conversation Words	Mean: 2.03 SD: 1.62	Mean: 0.88 SD: 1.07
Total Post Length	Mean: 16.91 SD: 6.22	Mean: 13.48 SD: 5.38
Parsing Tree Max Depth	Mean: 11.93 SD: 4.35	Mean: 9.21 SD: 3.60

Table (17). Means and standard deviations of some content-based features.

This table demonstrates several important properties of posts about personal experiences. First, these posts tend to be longer in terms of total length, but shorter in average word length, in part due to the fact that words referring to self ("I," "me," etc.) are shorter. Also, these posts are on average three times more likely to contain possessive pronouns and possession modifiers. This is probably due to the presence of statements about the personal experiences about friends and relatives, such as the snippet "my aunt's treatment" from one Personal tweet.

Personal posts are also more likely to contain conversational words, which is intuitive given that descriptions of personal experiences are likely to be informal personal accounts. Less intuitively, these posts also have deeper parsing trees, which is a rough approximation of syntactic complexity. One possible reason for this is that many Impersonal posts are actually Twitter posts containing news headlines, which tend to be shorter than full sentences and do not contain as much grammatical complexity. Also, Twitter posts containing news headlines

generally contain a link to the article in the tweet, which is likely the main reason why Impersonal posts are so much more likely to contain URLs. Finally, news headlines are generally written in the third person, which explains the lack of words referring to self, the lack of possession modifiers, and the lack of conversational words.

Chapter 7

Sentiment Analysis and Transfer of Learning Between Domains

7.1 Sentiment Analysis

Although the four datasets used in this experiment were tagged for experience rather than sentiment, it is possible to use unsupervised measures of sentiment in order to examine the relationship between affect and presence of personal experience in these Twitter posts. The nature of this relationship carries important implications for sentiment analysis researchers. If sentiment is strongly correlated with the presence of personal experience, this suggests that personal experience can be used as a useful feature in a sentiment analysis classifier. However, if this is the case, it creates a sampling problem for researchers who want to first filter posts by personal experience and then perform sentiment analysis, as the initial filtering for experience will exclude positive and negative posts in unequal proportions.

In this section, we will discuss two types of measures of sentiment: The SentiStrength classifier developed by Thelwall et al. (2010) and the number of positive-affect and negative-affect words in each tweet, using modified sentiment wordlists from [102,124]. These metrics were applied to the two largest datasets discussed in this thesis, the breast cancer dataset and the diabetes dataset.

The SentiStrength system is a supervised learning classification model designed for short texts such as tweets; it uses a classification scheme similar to the bag-of-words model, with some special rules for negation words and other grammatical constructs. The model has been trained on a variety of texts from different domains, which makes it possible, according to the developers of the model, to simply apply it to a new domain without retraining. The SentiStrength model

outputs two measures, a positive sentiment score on a scale of 1 through 5, and a negative sentiment score on a scale of -1 through -5.

We can also use the frequency of positive and negative words in each tweet as a simple measure of sentiment. As in Chapter 5, we use the wordlists from [102]. Using these wordlists, we can devise two methods of computing a sentiment score: The percentage of words in a tweet that are positive, and the percentage of words in a tweet that are negative. Together, these two methods give us four ways of measuring sentiment: SentiStrength score (+), SentiStrength score (-), positive word percentage, and negative word percentage. We can then examine the distributions of these metrics on both Personal and Impersonal posts in the breast cancer and diabetes datasets.

First, we can examine the SentiStrength score (+) in the breast cancer dataset:

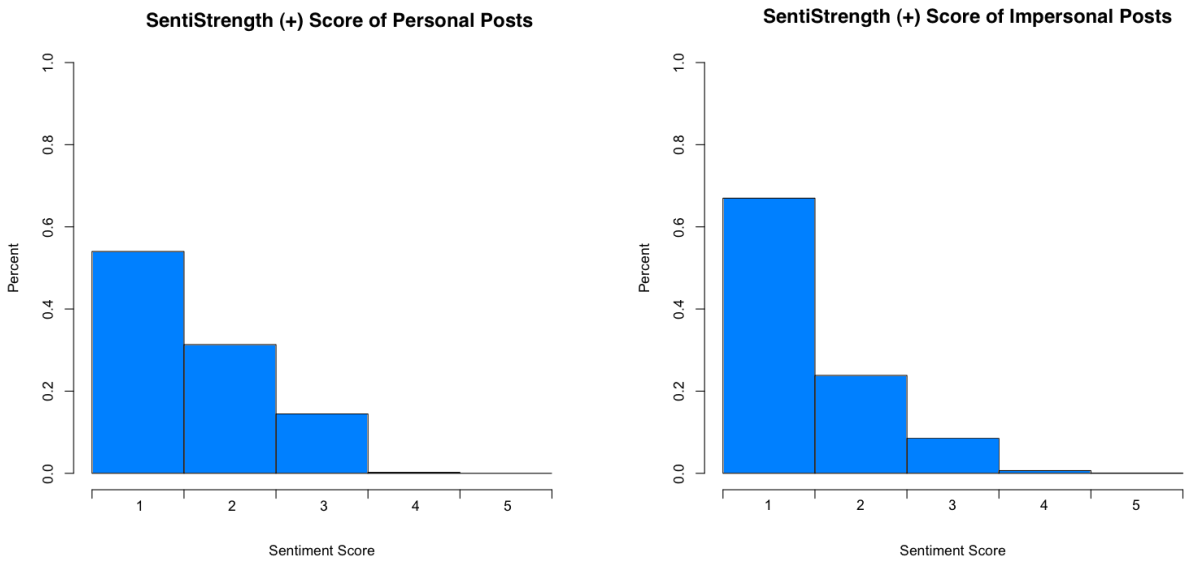

Figure (29). SentiStrength (+) scores for breast cancer dataset.

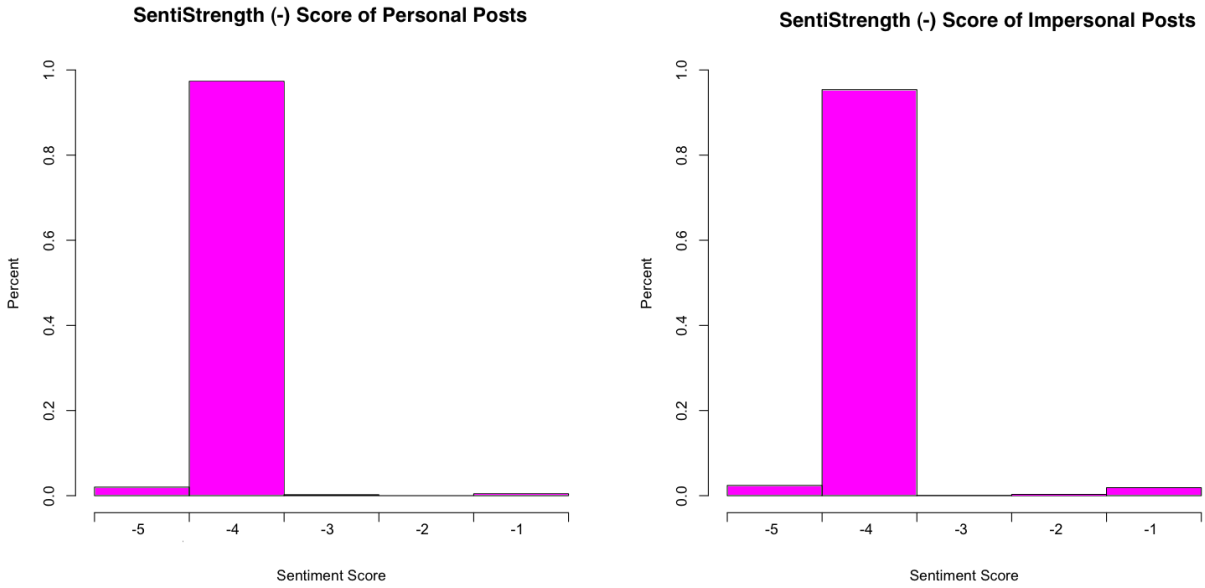

Figure (30). SentiStrength (-) scores for breast cancer dataset.

As the histograms above show, according to the SentiStrength score (+) metric most posts in the breast cancer dataset have a neutral sentiment, regardless of whether they are classified as Personal or Impersonal, with the sentiment strength increasing inversely proportional to the frequency. The SentiStrength score (-), however, reports a very different conclusion, with nearly all posts tagged as -4. Unfortunately, this is in fact due to the design of the SentiStrength system, which assigns a strong negative weight to the word "cancer," which appears in all of the posts in the breast cancer dataset. Due to the proprietary and closed-source nature of the software, this error cannot be corrected. Fortunately, this does not impact the word count scores, as shown in the figures below.

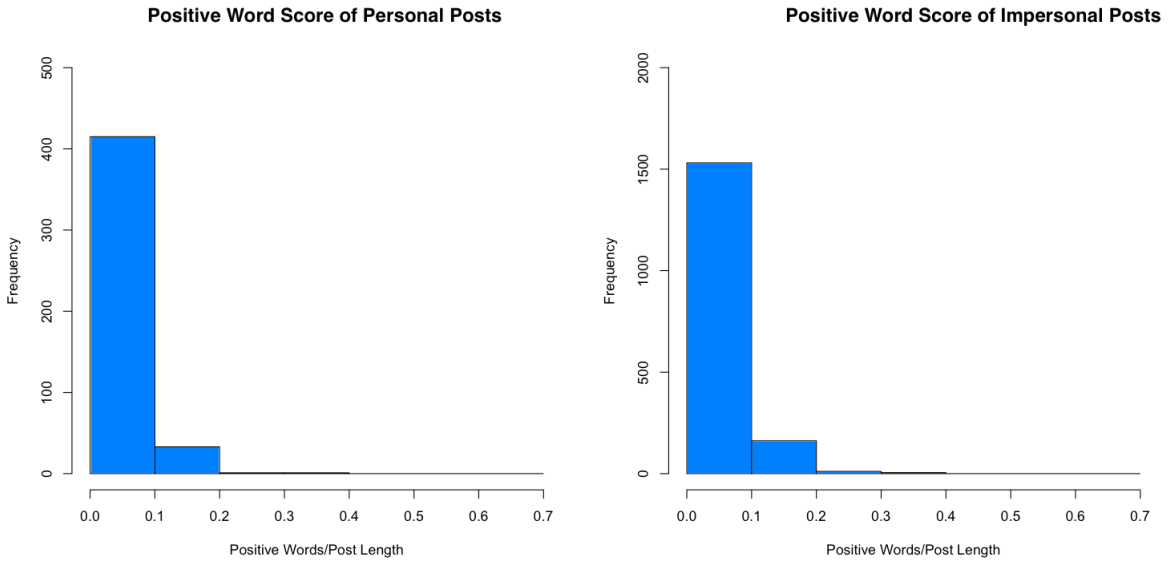

Figure (31). Positive word count scores for breast cancer dataset, using modified sentiment wordlists from [102].

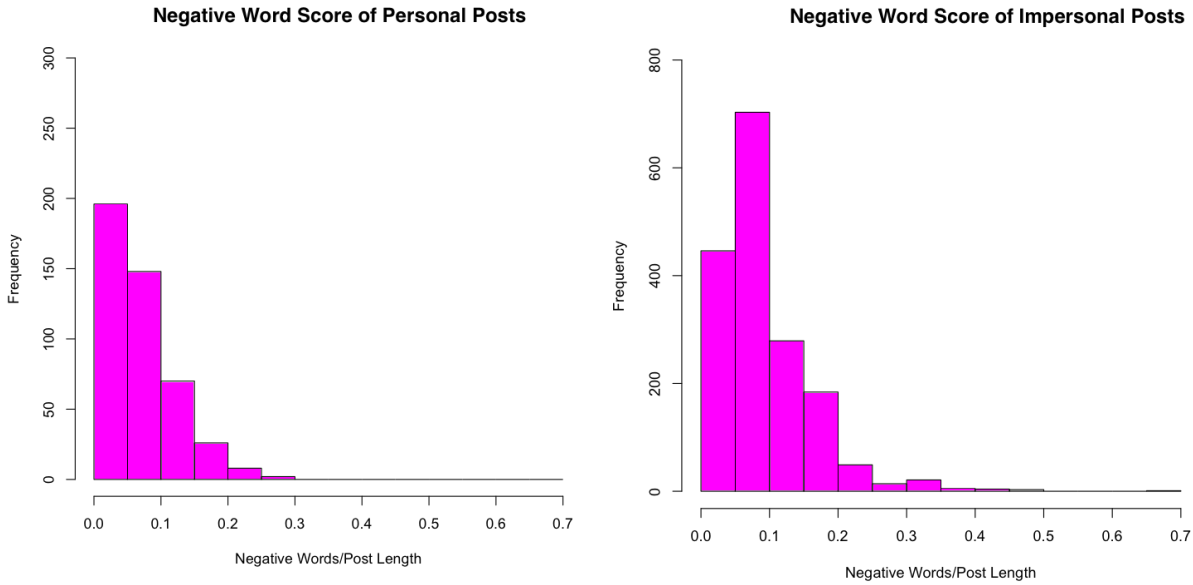

Figure (32). Negative word count scores for breast cancer dataset, using modified sentiment wordlists from [102].

These metrics confirm the SentiStrength score (+) results shown above, and suggest that most posts have neutral sentiment, regardless of class; only the negative word count metric appears to vary between classes. As shown in the histograms below, this lack of significant difference in sentiment between the two classes is also observed in the diabetes dataset.

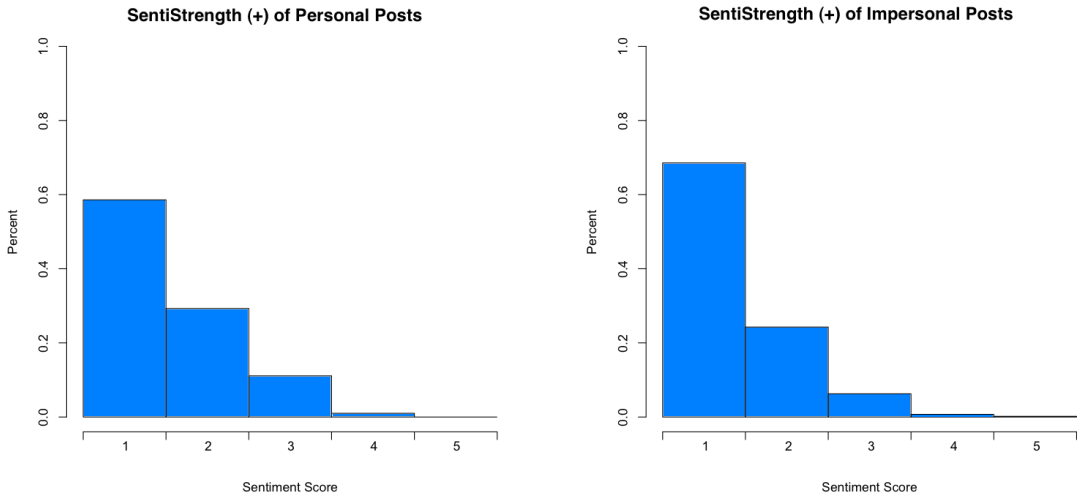

Figure (33). SentiStrength (+) scores for diabetes dataset.

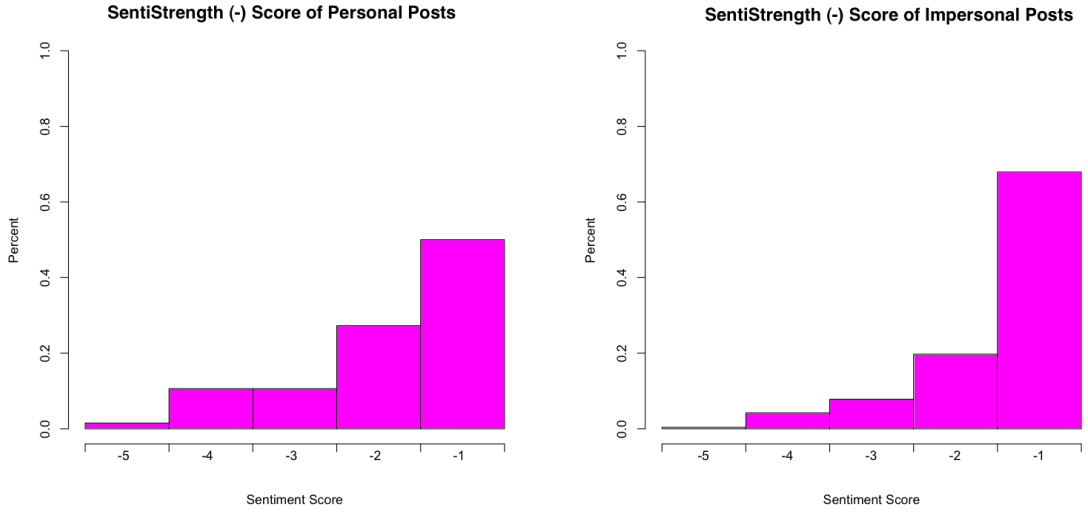

Figure (34). SentiStrength (-) scores for diabetes dataset.

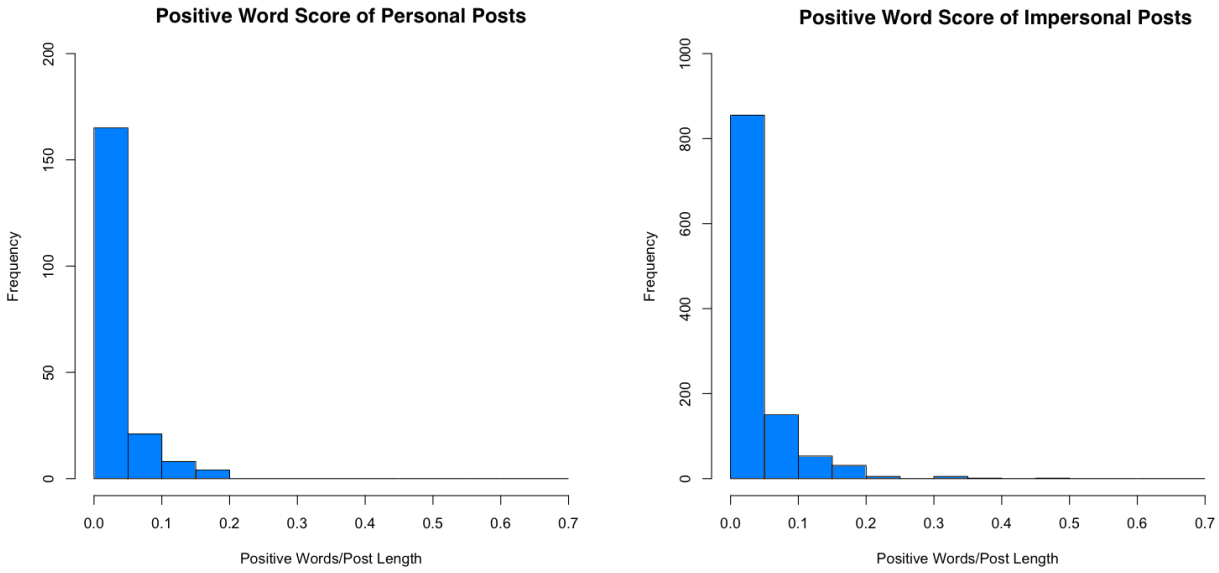

Figure (35). Positive word count scores for diabetes dataset, using modified sentiment wordlists from [102].

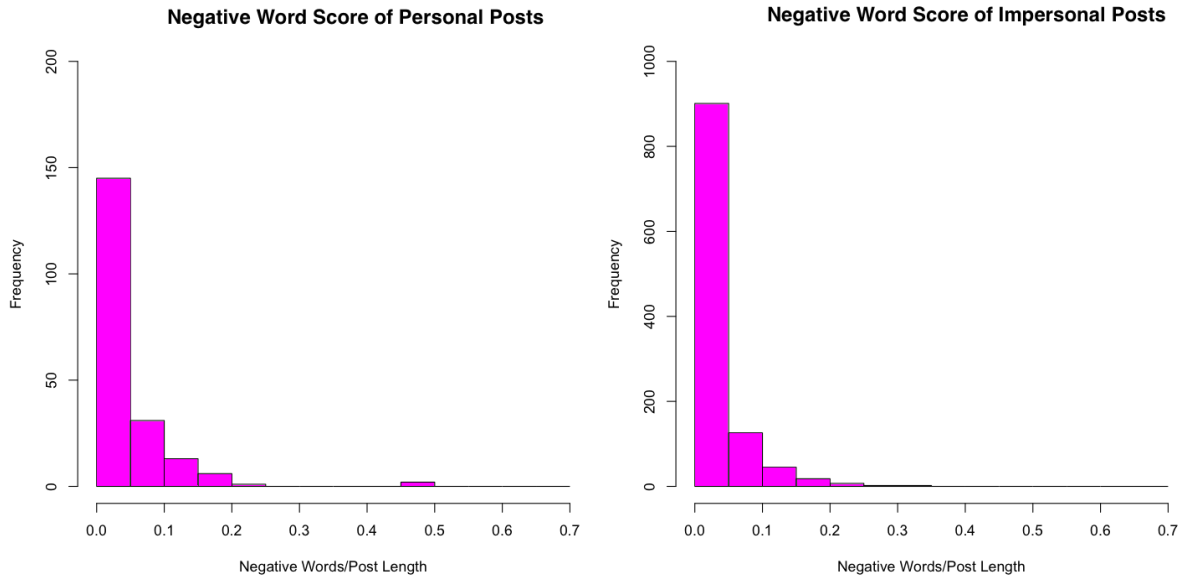

Figure (36). Negative word count scores for diabetes dataset, using modified sentiment wordlists from [102].

Again, as in the breast cancer dataset, only the negative word count appears to vary between the two classes.

In order to more rigorously determine whether there is a statistically significant relationship between these measures of sentiment and personal experience, we can use a Kolomogorov-Smirnov test, which a non-parametric mean difference test. A non-parametric test is necessary because the distributions of these four measures of sentiment do not follow a normal distribution, or indeed any regular probability distribution function. Using the Kolomogorov-Smirnov test, we can test the one-sided hypothesis "Impersonal posts have lower sentiment than Personal posts." The null hypothesis is that we cannot conclude that there is a statistically significant relationship between sentiment and personal experience. The p-values for this test on the breast cancer dataset are shown in the table below:

Metric	P-value
SentiStrength (+)	0.98
SentiStrength (-)	0.98
Positive Words	0.47
Negative Words	0.99

Table (18). Kolomogorov-Smirnov test on breast cancer dataset.

As shown in the table, we fail to reject the null hypothesis at the $p = 0.05$ significance level for all four metrics of sentiment, suggesting that there is not a significant relationship between these metrics and the presence of personal experience in the breast cancer dataset. Applying this same test to the diabetes dataset yields similar results:

Metric	P-value
SentiStrength (+)	0.99
SentiStrength (-)	0.99
Positive Words	0.30
Negative Words	0.01

Table (19). Kolomogorov-Smirnov test on diabetes dataset.

As in the breast cancer dataset, we fail to reject the null hypothesis for the SentiStrength (+), SentiStrength (-), and Positive Words metrics. However, we can reject the null hypothesis at the $p = 0.05$ significance level for the Negative Words metric. Nevertheless, given that the other three metrics show no significant difference, this result is not powerful enough to suggest a strong divergence in sentiment between the Personal and Impersonal classes.

Overall, this is a useful result for researchers who are interested in isolating social media posts about personal experiences and then performing sentiment analysis, as it demonstrates that researchers can use standard methods of feature extraction to perform the personal experience classification task without disproportionally filtering out positive or negative posts. It also suggest that there is a wide range of personal experiences discussed on Twitter by users who post about chronic diseases, ranging from strongly negative experiences, such as diagnosis of a chronic disease, to strongly positive experiences, such recovering from a chronic disease. This is useful for healthcare professionals and other researchers interested in mining health information from social media, as it shows that there are a wealth of different experiences that can be mined from Twitter.

7.2. Possibilities for Transfer of Learning

Due to the time-consuming nature of labeling messy social media data, many researchers have explored transfer of learning approaches to text mining problems. These approaches are advantageous because they do not require new labeled data for every new domain we wish to investigate, and they can reveal new insights about the similarity between particular domains. In this section, we explore whether it is feasible to transfer learning between domains for the personal experience classification task.

In order to determine whether the problem of personal experience prediction is amenable to a transfer of learning approach, we consider two specific research questions. First, can knowledge be transferred between cancer domains using the context-based and content-based features discussed above? Second, can knowledge be transferred between the cancer domain and the diabetes domain using these features? In this section, we attempt to answer these questions with a series of supervised learning experiments in which we transfer knowledge from one dataset to another. First, we transfer knowledge between the three cancer datasets, and then between the combined cancer datasets and the diabetes dataset.

The classification procedure used for this set of experiments is similar to the procedure used in the supervised learning experiments in the previous chapter: A classification model is trained on all of the instances from one dataset (the "source" dataset), and 5-fold cross-validation is used to select the parameters of each classifier on this source dataset. The list of classifiers and parameters is identical to those in the previous chapter. If feature selection is performed, it is also performed using 5-fold cross-validation on the source dataset. The resulting classifier is then used to classify all of the instances in another dataset (the "target" dataset).

In the first set of experiments, classifier parameters were optimized to minimize the error rate on the source dataset:

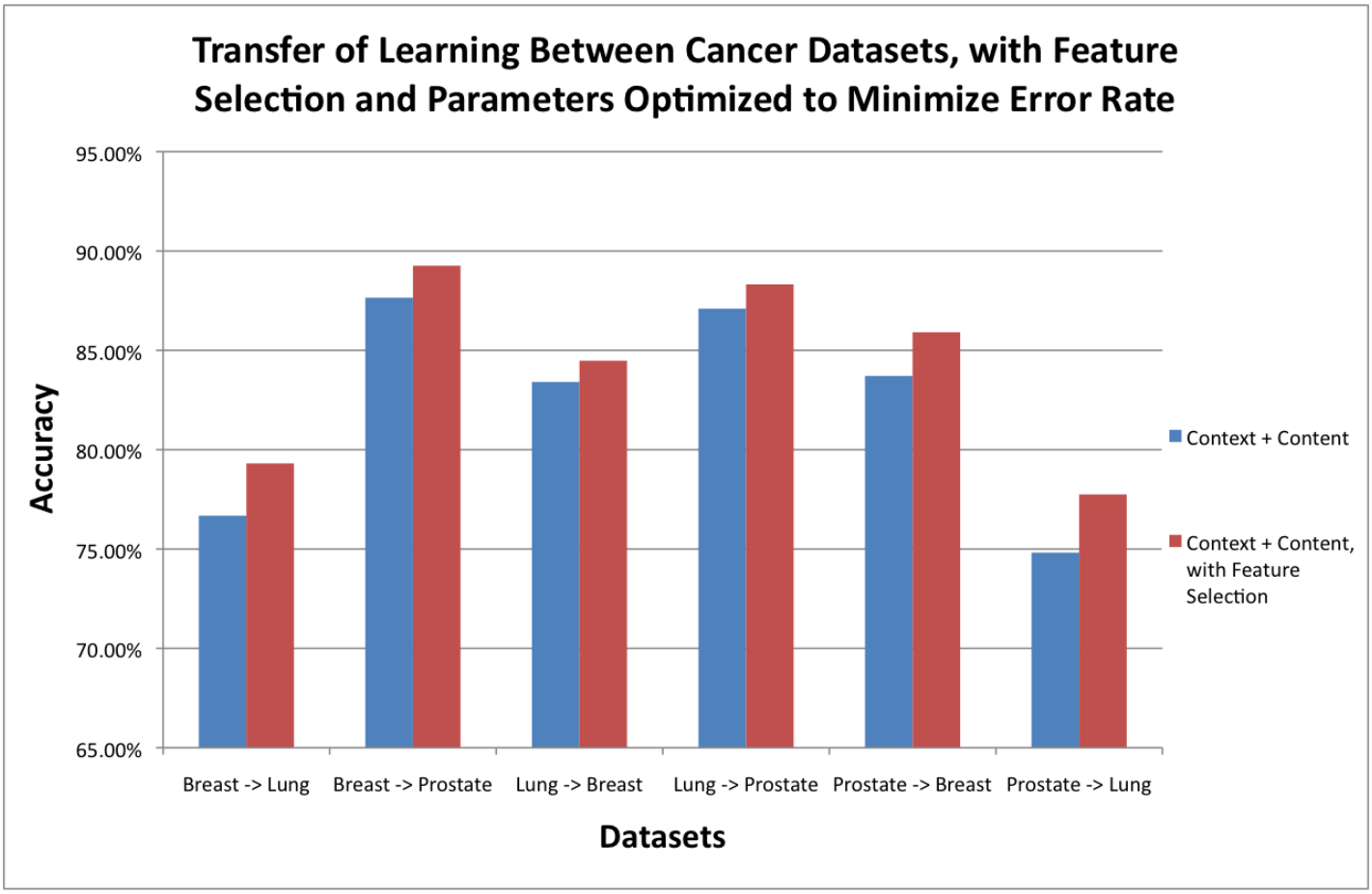

Figure (37). Classification accuracies for transfer learning experiments on cancer datasets, using Content + Context features, with parameters optimized to minimize error rate on source dataset.

As shown in the figure above, these accuracies are comparable to the supervised learning classifiers in the previous chapter. In the six transfer of learning experiments shown above, knowledge can be transferred between the breast cancer dataset and prostate cancer dataset with only 1-2% reduction in accuracy compared to the classifiers shown in Chapter 6. However, transferring knowledge into the lung cancer dataset is less effective. Across all six sets of classifiers, feature selection on the set of Context + Content features improves performance.

Next, the three cancer datasets were combined into a single dataset, and the transfer of learning procedure was applied to this combined cancer dataset and the diabetes dataset. The results of these experiments are shown below:

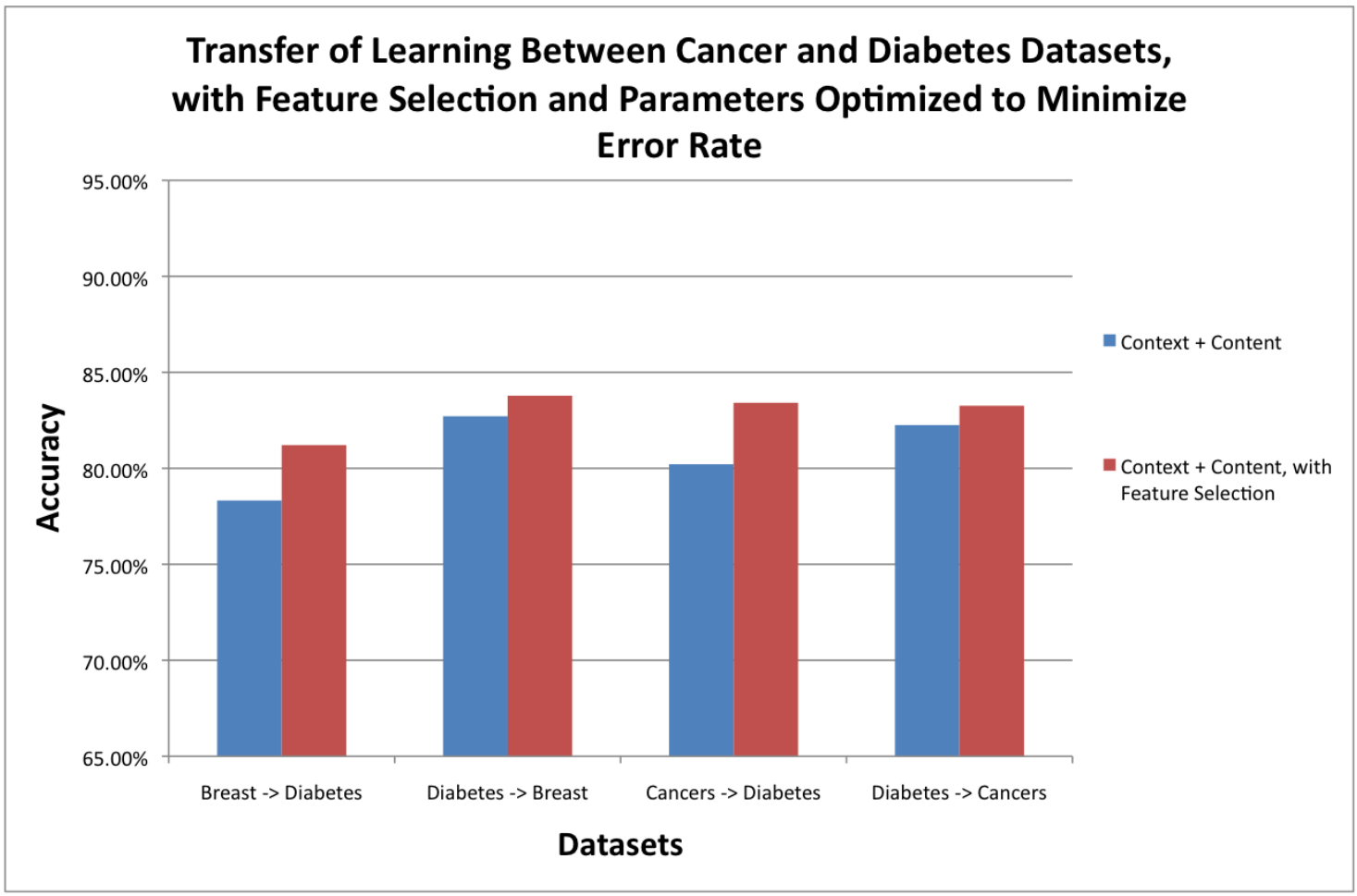

Figure (38). Classification accuracies for transfer learning experiments on cancer and diabetes datasets, using Content + Context features, with parameters optimized to minimize error rate on source dataset.

Compared to the three cancer datasets, transferring learning between the cancer domain and diabetes domain results in a significant decrease in accuracy. Interestingly, even with this drop in accuracy, feature selection was still uniformly effective at increasing performance. This is likely due to the fact that there are some features that are not particularly useful in either domain, and removing them increases performance regardless of the specific classification task.

However, as discussed in the previous chapter, accuracy is not a suitable metric to assess classifier performance on this task, as all four datasets are highly unbalanced. Therefore, these

experiments were repeated, optimizing classifier parameters to maximize f-measure rather than minimizing error rate.

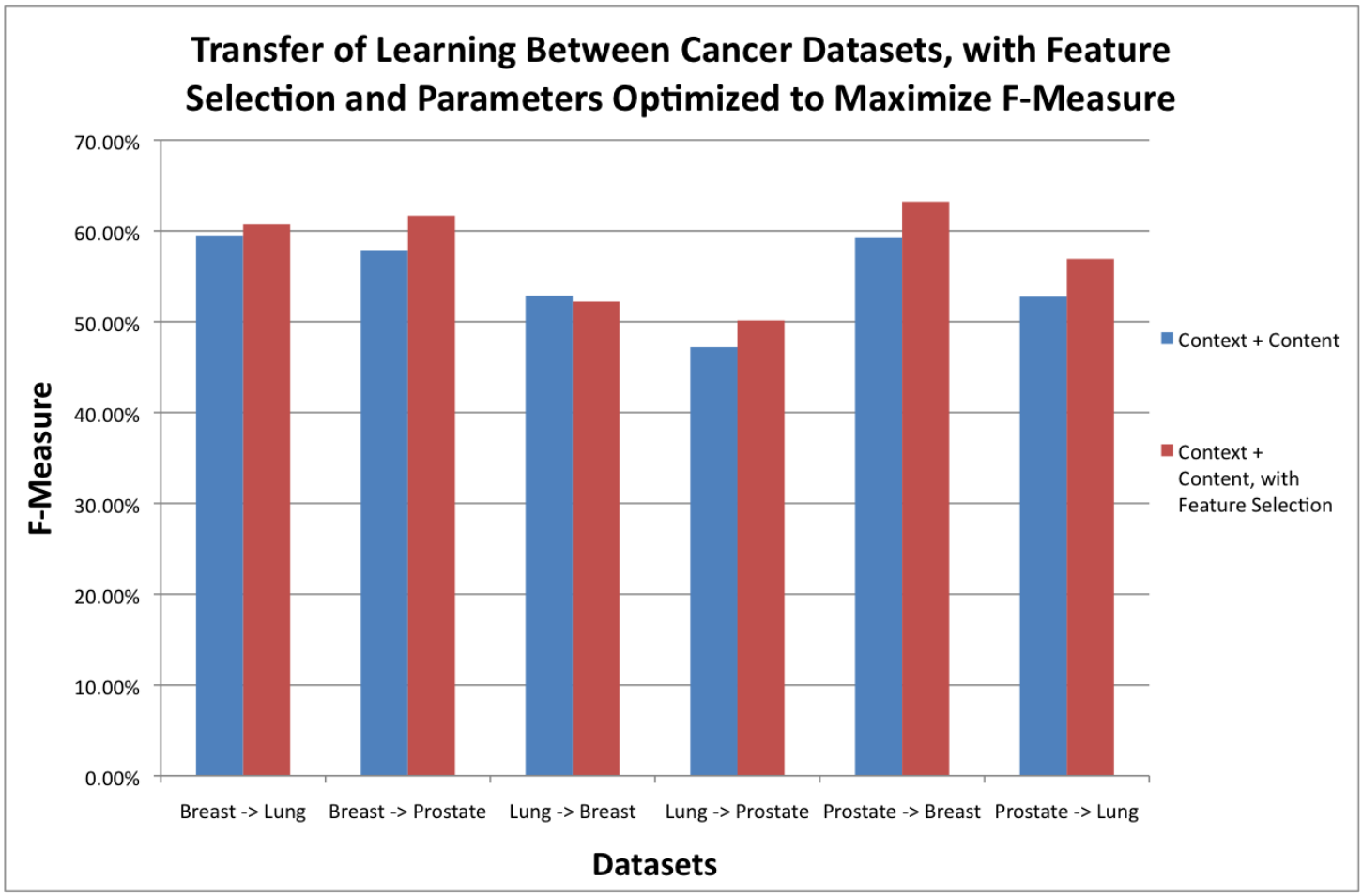

Figure (39). Classification f-measures for transfer learning experiments on cancer datasets, using Content + Context features, with parameters optimized to maximize f-measure on source dataset.

As the figure above shows, when examining f-measure rather than accuracy, we observe that some of these classifiers are not as powerful when forced to train in one cancer domain and test in another. Average f-measure on the breast cancer dataset decreased by 3 percentage points and average f-measure on the prostate cancer dataset decreased by 4 percentage points compared to the supervised learning classifiers with these same featuresets in the previous chapter.

However, average f-measure for the lung cancer dataset actually remained the same as in the previous chapter, despite the fact that these classifiers were not trained in the lung cancer domain. This suggests that that three cancer datasets do share a great deal of domain-specific information.

As above, this procedure was repeated for the combined cancer dataset and diabetes datasets:

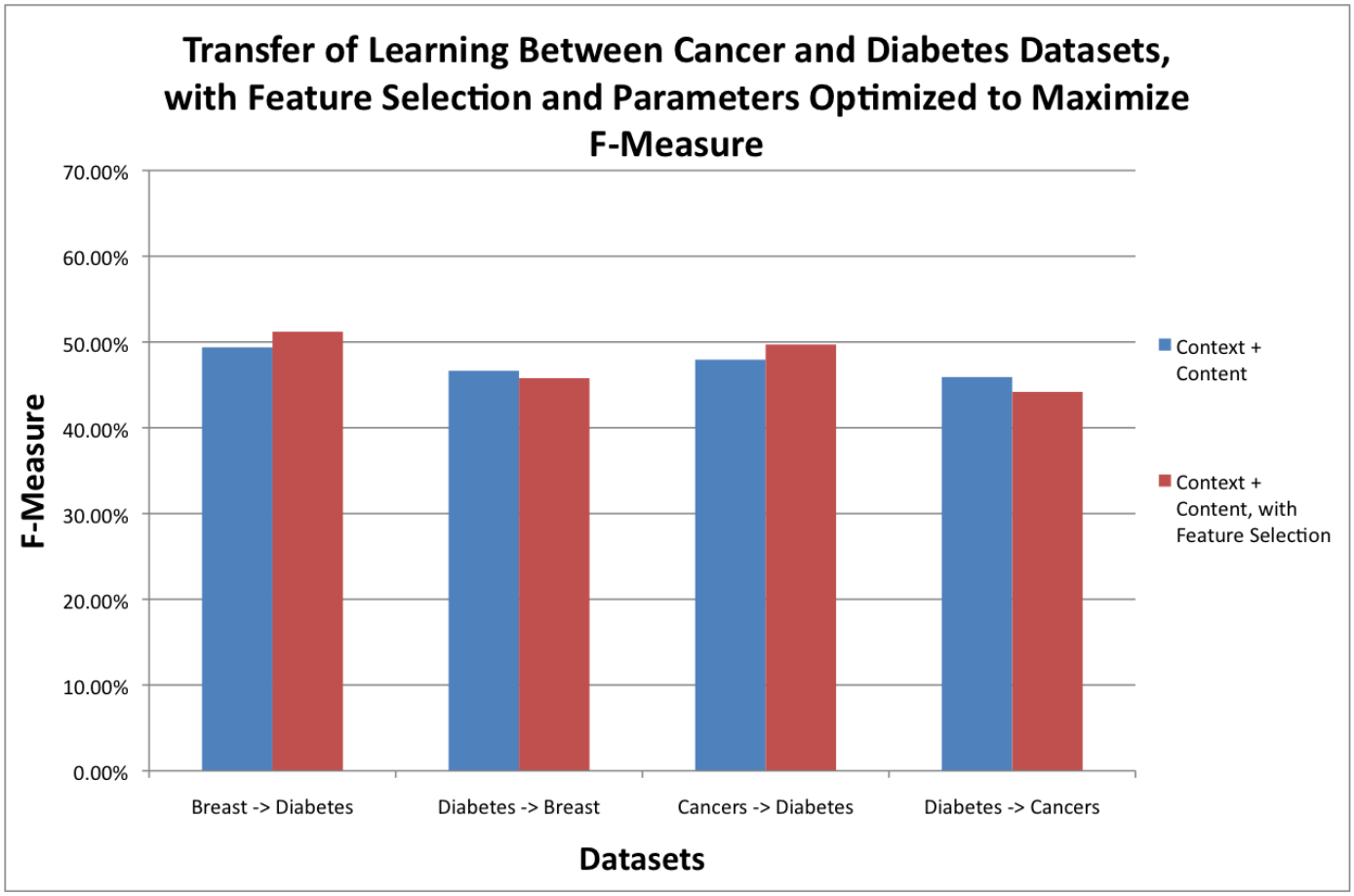

Figure (40). Classification f-measures for transfer learning experiments on cancer and diabetes datasets, using Content + Context features, with parameters optimized to maximize f-measure on source dataset.

This set of experiments demonstrates that knowledge is not as easily transferred between two different types of chronic diseases at it is between types of cancer—average f-measure on the breast cancer dataset dropped by nearly 20 percentage points, and average f-measure on the

combined cancer dataset decreased by 15 percentage points. Interestingly, however, average f-measure on the diabetes dataset increased by 10 percentage points compared to the standard supervised learning approach. This somewhat surprising result has two possible explanations: First, it is possible that the large size of the combined cancer dataset is giving the classifier more predictive power, which is why a classifier trained on this dataset performs well on the diabetes dataset but not vice versa. However, it is also possible that the combined cancer dataset is more general than the diabetes dataset, and so a classifier trained on this combined dataset can be applied to a more specific domain, the diabetes dataset.

Overall, these results suggest that this domain is indeed amenable to a transfer of learning approach. This is a positive finding for researchers who wish to study social media but do not have access to a large labeled dataset in the domain they wish to study, as it shows that labeled data from similar domains can also be used. These experiments also show that the Context + Content features discussed in this thesis are generalizable to multiple domains, which is a very useful property when studying a multifaceted domain such as public health.

Chapter 8

Conclusion

In this thesis, we reflected on the use of supervised learning methods to classify Twitter posts about cancer based on their relevance to personal experiences. To do so, we discussed several novel methods to improve supervised learning of short, messy texts related to chronic disease. First, we showed that a fuzzy string matching algorithm can be used to identity retweets in a dataset of Twitter posts, which is a common obstacle when training a supervised learning system. Second, we introduced new methods of feature extraction and feature engineering for short social media posts, including features based on both the context of a social media post and NLP-based features that reflect its content.

We demonstrated that these features perform better than the bag-of-words model on a dataset of 4,821 labeled tweets about cancer and diabetes. Despite the highly unbalanced nature of this dataset, Twitter posts can be classified with 85% accuracy and 65% f-measure using content and context-based features. We also observed that context-features are generally inferior to content-based features, due in part to the irregular distribution of these properties across the population of Twitter users. By contrast, content-based features are very effective at distinguishing between posts about personal experiences and posts containing other types of content. In particular, features related to the presence of words referring to oneself and the presence of URLs were very powerful and performed well across all domains.

Finally, we showed that these features capture knowledge of personal experiences in a way that can be easily transferred between domains, as classifiers trained on one dataset perform well when used to classify posts belonging to another dataset. This is a useful finding for researchers interested in mining short text data about to disease-related personal experiences, as it

suggests that domain-specific labeled data is not necessary to accurately identify texts related to personal experiences. This finding also suggests that users communicate in similar ways across medical domains, even though the diseases being discussed are different.

This line of research has many applications in the medical field, and might lead to the development of newer, more patient-centric means of providing healthcare. The ability to accurately identify descriptions of personal experience can aid in both treatment and diagnosis, not only in the chronic disease domain but also for infectious diseases and mental illness.

In terms of treatment, automated personal experience mining systems have the potential to crowdsource drug testing by monitoring patients' experiences with particular medications on the Internet. Currently, online communities such as PatientsLikeMe [91] provide a platform for patients to voluntarily share their drug regimens and experiences, but an automated text mining program continuously monitoring public social media sites for descriptions of personal experiences could collect data from an even larger population of patients. These descriptions of personal experiences with medications could then be aggregated and searched for descriptions of Adverse Drug Reactions or other side effects. Such a system could function as a low-cost supplement to traditional controlled trials, or allow pharmaceutical companies to monitor the popularity and effectiveness of their products over time. However, this technology also presents new legal and ethical challenges, as there are privacy concerns associated with mining patient data, even that data is in the public domain. Additionally, information found on the Internet is often impossible to verify, which suggests that pharmaceutical manufacturers and healthcare providers should take a cautious approach when using this data to make real-world decisions.

Healthcare providers could also leverage an automated experience mining system for the purposes of diagnosis at both a global and local level. Personal experience mining is a powerful tool in the hands of public health surveillance researchers, who could use it to collect statistics about the prevalence of disease from social media rather than relying on more rudimentary

keyword-based searches. For example, as discussed previously, text mining is frequently used for influenza detection. Rather than simply scanning for the keywords “flu” or “influenza,” however, researchers could use a personal experience classifier to obtain a more accurate estimate of how many people are affected by a flu epidemic. As shown in this thesis, such an approach is also applicable to chronic health problems such as diabetes and cancer.

At a more local level, personal experience mining systems could be used to monitor a patient’s health over time by monitoring a record of their experiences mined from their social media profiles or from personal journal entries. When combined with other forms of text mining, such as sentiment analysis, personal experience mining systems could identify individuals who report having a long history of negative or stressful experiences, which might be an indicator of anxiety or some other mental disorder. This could also be applied to chronic diseases such as cancer, in which an automated system identifies landmark events over the course of a patient’s treatment (such as surgery or chemotherapy) and tracks how these affect various measures of sentiment such as self-reported happiness or sentiment score extracted from social media posts.

However, the ability to monitor patients’ experiences at individual level will also force us to rethink how we understand patient confidentiality. For example, consider an automated text mining system that crawls social media sites for individuals who describe personal experiences that are indicators of depression and suicidal tendencies, and then notifies a suicide prevention organization. This challenges our traditional notion of patient privacy, even if the information used is entirely in the public domain. Future legislators and healthcare professionals will need to decide how such systems can be used, and whether individual monitoring through the web is ethically allowable. Individuals may also need to be more vigilant about what information they post publically.

Outside of diagnosis and treatment, personal experience mining in the healthcare domain has the potential to create new research opportunities. For example, experience mining can also

be used as a tool to study the growth and development of online health communities, much in the same way that sentiment analysis is currently used to measure community engagement. Overall, as more and more health information is published to the Internet, the ability to automatically identity descriptions of health-related personal experiences is a powerful text analysis tool that may have noticeable impacts on how we perform public health surveillance and monitor patient progress.

Appendix A

Supervised Learning Experiment Results

Supervised learning classification performance for each of the 12 classifiers using the bag-of-words model + bigrams, with classifier parameters optimized to minimize error rate (with backtracking feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	85.55%	86.57%	58.52%
SVM (linear kernel)	85.22%	71.43%	57.32%
Logistic Model Trees	85.13%	85.02%	55.67%
Bagging	85.13%	84.67%	56.03%
Logistic Regression	85.04%	84.61%	55.52%
J48 Decision Tree	84.99%	74.24%	55.77%
RandomForest	84.72%	83.16%	58.15%
Bayesian Network	84.62%	84.45%	56.46%
K-Nearest Neighbors	84.53%	77.41%	55.92%
AdaBoost	84.44%	71.63%	55.20%
SVM (RBF kernel)	84.34%	65.06%	45.96%
Naïve Bayes	83.56%	84.13%	57.70%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	89.13%	79.04%	51.11%
Logistic Regression	88.82%	79.93%	52.32%
Logistic Model Trees	88.82%	78.10%	54.23%
SVM (linear kernel)	88.51%	69.17%	51.92%
J48 Decision Tree	88.35%	67.04%	38.06%
K-Nearest Neighbors	88.05%	68.11%	51.08%
Bayesian Network	88.04%	73.52%	49.01%
SVM (RBF kernel)	87.73%	61.37%	36.17%
Bagging	87.73%	79.39%	36.95%
AdaBoost	87.42%	71.79%	45.93%
RandomForest	87.27%	76.45%	45.92%
Naïve Bayes	86.34%	79.54%	52.34%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	82.98%	67.18%	50.73%
SVM (linear kernel)	82.97%	71.11%	57.69%
Logistic Regression	82.83%	77.32%	58.86%
Logistic Model Trees	82.69%	78.67%	57.28%
LogitBoost	81.98%	78.58%	56.20%
Bayesian Network	81.69%	76.65%	55.11%
RandomForest	81.69%	77.58%	54.35%
J48 Decision Tree	80.97%	66.26%	48.72%
AdaBoost	80.69%	69.77%	51.92%
Bagging	80.12%	76.48%	51.54%
Naïve Bayes	79.11%	78.27%	57.73%
K-Nearest Neighbors	78.11%	71.24%	51.08%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Bagging	85.68%	80.81%	29.94%
LogitBoost	85.53%	78.25%	37.80%
J48 Decision Tree	85.30%	65.79%	32.66%
SVM (RBF kernel)	85.22%	55.25%	19.67%
Bayesian Network	85.14%	73.06%	39.80%
RandomForest	85.07%	76.96%	37.18%
Logistic Model Trees	84.76%	79.34%	37.20%
Logistic Regression	84.76%	79.87%	38.72%
AdaBoost	84.53%	70.66%	21.22%
SVM (linear kernel)	84.45%	60.78%	34.45%
K-Nearest Neighbors	82.83%	67.83%	37.50%
Naïve Bayes	81.52%	79.70%	46.38%

Average performance for each dataset:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	84.77%	79.37%	55.68%
Prostate Cancer	88.02%	73.62%	47.09%
Lung Cancer	81.32%	74.09%	54.27%
Diabetes	84.56%	72.36%	34.38%

Supervised learning classification performance for each of the 12 classifiers using the bag-of-words model + bigrams, with classifier parameters optimized to maximize f-measure (with backtracking feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	85.59%	86.58%	58.93%
SVM (linear kernel)	85.36%	71.60%	57.63%
Logistic Model Trees	85.13%	85.02%	55.67%
Bagging	85.13%	84.67%	56.03%
Logistic Regression	85.04%	84.61%	55.52%
J48 Decision Tree	84.81%	75.84%	55.86%
Bayesian Network	84.62%	84.45%	56.46%
RandomForest	84.53%	83.08%	58.35%
K-Nearest Neighbors	84.53%	77.41%	55.92%
AdaBoost	84.44%	71.63%	55.20%
SVM (RBF kernel)	84.25%	65.33%	46.50%
Naïve Bayes	83.56%	84.13%	57.70%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	89.60%	79.58%	54.39%
Logistic Regression	88.82%	79.93%	52.32%
Logistic Model Trees	88.82%	78.10%	54.23%
SVM (linear kernel)	88.51%	69.17%	51.92%
K-Nearest Neighbors	88.05%	68.11%	51.08%
Bayesian Network	88.04%	73.52%	49.01%
J48 Decision Tree	88.04%	67.77%	41.71%
Bagging	87.58%	78.22%	36.49%
SVM (RBF kernel)	87.42%	61.19%	35.58%
RandomForest	87.27%	76.45%	45.92%
AdaBoost	86.95%	70.97%	45.56%
Naïve Bayes	86.34%	79.54%	52.34%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	82.98%	67.18%	50.73%
Logistic Regression	82.83%	77.32%	58.86%
Logistic Model Trees	82.69%	78.67%	57.28%
SVM (linear kernel)	82.69%	70.53%	56.72%
Bayesian Network	81.69%	76.65%	55.11%
LogitBoost	81.54%	78.22%	56.74%
AdaBoost	80.69%	69.77%	51.92%
Bagging	80.55%	76.91%	52.19%
RandomForest	80.40%	76.35%	53.24%
J48 Decision Tree	80.12%	65.30%	49.62%
Naïve Bayes	79.11%	78.27%	57.73%
K-Nearest Neighbors	78.11%	71.24%	51.08%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	85.53%	78.70%	38.59%
J48 Decision Tree	85.22%	63.70%	39.21%
SVM (RBF kernel)	85.22%	55.46%	20.37%
Bayesian Network	85.14%	73.06%	39.80%
Bagging	85.07%	79.61%	25.07%
RandomForest	84.99%	77.29%	37.44%
Logistic Model Trees	84.76%	79.34%	37.20%
Logistic Regression	84.76%	79.87%	38.72%
SVM (linear kernel)	84.45%	61.60%	35.93%
AdaBoost	84.22%	71.89%	23.43%
K-Nearest Neighbors	82.83%	67.83%	37.50%
Naïve Bayes	81.52%	79.70%	46.38%

Average performance for each dataset:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	84.75%	79.53%	55.81%
Prostate Cancer	87.95%	73.55%	47.55%
Lung Cancer	81.12%	73.87%	54.27%
Diabetes	84.48%	72.34%	34.97%

Supervised learning classification performance for each of the 12 classifiers using the context-based features, with classifier parameters optimized to minimize error rate (with no feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Regression	79.53%	75.83%	24.34%
Bagging	79.43%	78.95%	32.60%
J48 Decision Tree	79.39%	68.64%	43.25%
Logistic Model Trees	79.25%	77.40%	32.68%
SVM (linear kernel)	79.16%	50.00%	0.00%
SVM (RBF kernel)	79.16%	50.00%	0.00%
LogitBoost	79.16%	79.73%	30.43%
RandomForest	78.46%	78.57%	35.48%
AdaBoost	77.91%	73.65%	15.33%
Bayesian Network	77.53%	79.03%	50.76%
K-Nearest Neighbors	75.08%	62.04%	39.69%
Naïve Bayes	39.46%	75.79%	39.67%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	85.09%	50.00%	0.00%
SVM (RBF kernel)	85.09%	50.00%	0.00%
Bayesian Network	84.78%	81.12%	46.81%
Logistic Regression	84.78%	77.68%	28.36%
Logistic Model Trees	84.32%	70.53%	9.75%
Bagging	84.01%	82.66%	14.42%
RandomForest	84.00%	80.68%	31.17%
AdaBoost	83.85%	76.04%	17.22%
J48 Decision Tree	83.39%	64.89%	30.06%
LogitBoost	83.38%	83.23%	21.41%
K-Nearest Neighbors	79.66%	59.18%	31.11%
Naïve Bayes	68.50%	77.24%	39.88%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	75.54%	50.00%	0.00%
SVM (RBF kernel)	75.54%	50.00%	0.00%
AdaBoost	75.54%	60.36%	0.00%
Logistic Model Trees	75.54%	50.00%	0.00%
LogitBoost	75.11%	67.35%	2.20%
Bagging	74.96%	65.58%	1.03%
J48 Decision Tree	74.39%	50.97%	9.60%
Logistic Regression	73.82%	64.61%	1.00%
RandomForest	71.96%	67.40%	16.63%
Bayesian Network	67.39%	66.36%	34.53%
K-Nearest Neighbors	65.67%	54.36%	30.67%
Naïve Bayes	48.65%	64.06%	43.83%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	84.76%	50.00%	0.00%
SVM (RBF kernel	84.76%	50.00%	0.00%
AdaBoost	84.76%	64.25%	0.00%
Bagging	84.76%	76.70%	2.83%
J48 Decision Tree	84.76%	50.00%	0.00%
Logistic Model Trees	84.76%	50.00%	0.00%
LogitBoost	84.53%	74.78%	0.91%
Logistic Regression	84.45%	69.59%	0.00%
RandomForest	84.37%	73.90%	24.21%
Bayesian Network	80.99%	70.69%	17.44%
K-Nearest Neighbors	77.75%	56.48%	25.42%
Naïve Bayes	32.25%	66.55%	28.81%

Average performance for each dataset:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	75.29%	70.80%	28.69%
Prostate Cancer	82.57%	71.10%	22.52%
Lung Cancer	71.17%	59.25%	11.62%
Diabetes	79.41%	62.75%	8.30%

Supervised learning classification performance for each of the 12 classifiers using the context-based features, with classifier parameters optimized to maximize f-measure (with no feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	79.53%	79.59%	38.89%
Logistic Regression	79.53%	75.83%	24.34%
Bagging	79.39%	78.68%	36.20%
Logistic Model Trees	79.25%	77.40%	32.68%
J48 Decision Tree	79.25%	67.76%	45.41%
SVM (linear kernel)	79.16%	50.00%	0.00%
SVM (RBF kernel)	79.16%	50.00%	0.00%
RandomForest	78.79%	75.80%	41.48%
AdaBoost	77.91%	77.95%	15.33%
Bayesian Network	77.53%	79.03%	50.76%
K-Nearest Neighbors	75.08%	62.04%	39.69%
Naïve Bayes	39.46%	75.79%	39.67%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	85.09%	50.00%	0.00%
Bayesian Network	84.78%	81.12%	46.81%
Logistic Regression	84.78%	77.68%	28.36%
AdaBoost	84.63%	79.84%	32.87%
Logistic Model Trees	84.32%	70.53%	9.75%
SVM (linear kernel)	84.32%	53.22%	8.37%
Bagging	83.85%	82.95%	17.77%
J48 Decision Tree	83.70%	61.93%	35.26%
RandomForest	83.69%	79.01%	33.43%
LogitBoost	82.61%	80.91%	24.13%
K-Nearest Neighbors	79.66%	59.18%	31.11%
Naïve Bayes	68.50%	77.24%	39.88%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	75.54%	50.00%	0.00%
SVM (RBF kernel)	75.54%	50.00%	0.00%
AdaBoost	75.54%	61.47%	0.00%
Logistic Model Trees	75.54%	50.00%	0.00%
Bagging	74.53%	64.79%	10.59%
Logistic Regression	73.82%	64.61%	1.00%
LogitBoost	71.53%	65.45%	18.86%
RandomForest	71.24%	64.67%	23.45%
J48 Decision Tree	69.67%	55.95%	30.61%
Bayesian Network	67.39%	66.36%	34.53%
K-Nearest Neighbors	65.67%	54.36%	30.67%
Naïve Bayes	48.65%	64.06%	43.83%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	84.76%	50.00%	0.00%
SVM (RBF kernel)	84.76%	50.00%	0.00%
AdaBoost	84.76%	69.08%	0.00%
Logistic Model Trees	84.76%	50.00%	0.00%
Logistic Regression	84.45%	69.59%	0.00%
Bagging	84.14%	76.63%	4.47%
LogitBoost	83.91%	75.49%	18.65%
RandomForest	83.60%	71.87%	26.47%
Bayesian Network	80.99%	70.69%	17.44%
J48 Decision Tree	80.29%	59.80%	23.85%
K-Nearest Neighbors	77.75%	56.48%	25.42%
Naïve Bayes	32.25%	66.55%	28.81%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	75.34%	70.82%	30.37%
Prostate Cancer	82.49%	71.13%	25.64%
Lung Cancer	70.39%	59.31%	16.13%
Diabetes	78.87%	63.85%	12.09%

Supervised learning classification performance for each of the 12 classifiers using the content-based features, with classifier parameters optimized to minimize error rate (with no feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	88.14%	79.25%	69.27%
SVM (RBF kernel)	88.10%	77.01%	66.97%
Logistic Model Trees	88.00%	92.79%	68.35%
Logistic Regression	88.00%	92.60%	69.52%
LogitBoost	87.96%	92.67%	69.11%
Bagging	87.73%	92.28%	69.87%
AdaBoost	87.63%	92.01%	70.79%
RandomForest	87.22%	91.24%	65.16%
J48 Decision Tree	85.69%	72.56%	63.09%
Bayesian Network	82.08%	89.20%	64.08%
K-Nearest Neighbors	81.47%	70.66%	53.67%
Naïve Bayes	80.55%	86.76%	60.76%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Bagging	90.37%	91.85%	62.24%
Logistic Model Trees	89.44%	90.73%	59.15%
SVM (RBF kernel)	89.28%	66.58%	48.45%
LogitBoost	89.13%	89.98%	54.58%
AdaBoost	88.98%	90.20%	63.06%
RandomForest	88.36%	89.81%	46.80%
SVM (linear kernel)	88.20%	68.99%	49.98%
J48 Decision Tree	87.73%	75.33%	57.51%
Logistic Regression	87.42%	87.34%	58.03%
K-Nearest Neighbors	84.94%	66.79%	43.34%
Bayesian Network	84.32%	90.48%	57.71%
Naïve Bayes	82.61%	88.39%	54.96%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	83.83%	86.86%	59.13%
RandomForest	82.98%	84.32%	57.96%
Logistic Regression	82.69%	84.35%	60.96%
AdaBoost	82.69%	83.81%	55.53%
Logistic Model Trees	82.69%	84.40%	58.32%
SVM (RBF kernel)	82.40%	67.58%	51.56%
Bagging	82.12%	85.25%	52.38%
SVM (linear kernel)	81.97%	67.29%	50.72%
J48 Decision Tree	80.54%	71.68%	55.29%
Naïve Bayes	77.97%	83.86%	60.88%
Bayesian Network	77.68%	84.55%	61.19%
K-Nearest Neighbors	75.82%	65.97%	47.51%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	87.68%	88.15%	47.12%
SVM (linear kernel)	86.91%	67.18%	46.90%
Logistic Regression	86.91%	85.22%	49.51%
Logistic Model Trees	86.76%	87.52%	44.47%
Bagging	86.68%	85.49%	36.03%
RandomForest	86.07%	83.94%	33.20%
SVM (RBF kernel)	85.99%	55.30%	19.29%
AdaBoost	85.53%	84.30%	37.19%
J48 Decision Tree	84.29%	68.39%	39.33%
K-Nearest Neighbors	81.83%	61.76%	35.54%
Bayesian Network	78.98%	83.34%	46.94%
Naïve Bayes	75.98%	81.14%	48.03%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	86.05%	85.75%	65.89%
Prostate Cancer	87.56%	83.04%	54.65%
Lung Cancer	81.12%	79.16%	55.95%
Diabetes	84.47%	77.65%	40.30%

Supervised learning classification performance for each of the 12 classifiers using the content-based features, with classifier parameters optimized to maximize f-measure (with no feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel)	88.42%	80.16%	70.39%
SVM (RBF kernel)	88.05%	76.98%	66.89%
Logistic Model Trees	88.00%	92.79%	68.35%
Logistic Regression	88.00%	92.60%	69.52%
LogitBoost	87.96%	92.67%	69.11%
AdaBoost	87.63%	92.01%	70.79%
Bagging	87.59%	92.06%	69.92%
RandomForest	87.54%	91.19%	66.64%
J48 Decision Tree	85.73%	72.23%	63.25%
Bayesian Network	82.08%	89.20%	64.08%
K-Nearest Neighbors	81.47%	70.66%	53.67%
Naïve Bayes	80.55%	86.76%	60.76%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Bagging	90.53%	91.03%	62.55%
Logistic Model Trees	89.44%	90.73%	59.15%
SVM (RBF kernel)	89.28%	66.58%	48.27%
SVM (linear kernel)	89.14%	74.30%	59.19%
AdaBoost	88.98%	90.20%	63.06%
LogitBoost	88.97%	89.68%	60.55%
J48 Decision Tree	88.36%	75.97%	59.73%
RandomForest	88.36%	89.81%	46.80%
Logistic Regression	87.42%	87.34%	58.03%
K-Nearest Neighbors	84.94%	66.79%	43.34%
Bayesian Network	84.32%	90.48%	57.71%
Naïve Bayes	82.61%	88.39%	54.96%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	83.83%	86.08%	61.01%
SVM (RBF kernel)	82.83%	69.25%	54.75%
RandomForest	82.69%	84.06%	56.86%
Logistic Regression	82.69%	84.35%	60.96%
AdaBoost	82.69%	85.03%	55.07%
Logistic Model Trees	82.69%	84.40%	58.32%
Bagging	81.83%	85.34%	53.41%
SVM (linear kernel)	81.54%	70.97%	56.90%
J48 Decision Tree	80.54%	70.33%	55.54%
Naïve Bayes	77.97%	83.86%	60.88%
Bayesian Network	77.68%	84.55%	61.19%
K-Nearest Neighbors	75.82%	65.97%	47.51%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	87.60%	87.13%	50.03%
SVM (linear kernel)	87.22%	68.61%	49.77%
Logistic Regression	86.91%	85.22%	49.51%
Logistic Model Trees	86.76%	87.52%	44.47%
Bagging	86.45%	85.69%	38.45%
SVM (RBF kernel)	86.14%	55.80%	20.84%
RandomForest	86.14%	79.74%	39.24%
AdaBoost	85.91%	86.10%	43.98%
J48 Decision Tree	83.14%	66.16%	42.27%
K-Nearest Neighbors	81.83%	61.76%	35.54%
Bayesian Network	78.98%	83.34%	46.94%
Naïve Bayes	75.98%	81.14%	48.03%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	86.09%	85.78%	66.12%
Prostate Cancer	87.70%	83.44%	56.11%
Lung Cancer	81.07%	79.52%	56.87%
Diabetes	84.42%	77.35%	42.42%

Supervised learning classification performance for each of the 12 classifiers using the content-based and context-based features, with classifier parameters optimized to minimize error rate (with no feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Model Trees	88.23%	92.91%	68.79%
LogitBoost	88.14%	93.52%	70.16%
RandomForest	88.00%	91.42%	67.34%
Bagging	87.96%	92.40%	69.81%
Logistic	87.91%	92.66%	68.90%
SVM (RBF kernel)	87.86%	77.03%	66.42%
SVM (linear kernel)	87.77%	78.93%	68.24%
AdaBoost	87.35%	92.19%	69.96%
J48 Decision Tree	86.75%	74.15%	65.94%
Bayesian Network	82.82%	90.67%	65.86%
K-Nearest Neighbors	81.52%	69.61%	52.42%
Naïve Bayes	78.42%	86.89%	60.91%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Bagging	90.53%	91.56%	61.07%
Logistic Model Trees	89.76%	92.09%	61.35%
AdaBoost	89.75%	91.46%	63.52%
LogitBoost	89.44%	91.86%	55.44%
SVM (linear kernel)	88.98%	69.14%	52.59%
SVM (RBF kernel)	88.97%	68.70%	51.71%
J48 Decision Tree	88.67%	78.67%	58.47%
RandomForest	88.35%	90.53%	46.24%
Logistic Regression	87.12%	84.17%	60.06%
Bayesian Network	85.24%	91.64%	59.68%
K-Nearest Neighbors	84.78%	67.07%	44.63%
Naïve Bayes	80.59%	85.72%	52.38%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	84.12%	70.62%	57.21%
AdaBoost	83.12%	86.28%	61.09%
SVM (linear kernel)	83.12%	68.65%	53.17%
Bagging	82.83%	85.37%	54.13%
LogitBoost	82.69%	87.35%	54.06%
Logistic Model Trees	82.69%	86.65%	58.50%
J48 Decision Tree	82.40%	71.90%	60.58%
RandomForest	80.97%	83.59%	51.81%
Logistic Regression	80.55%	82.99%	57.68%
Bayesian Network	77.98%	85.14%	61.43%
Naïve Bayes	75.68%	83.64%	59.10%
K-Nearest Neighbors	75.25%	64.20%	45.63%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Model Trees	87.22%	88.11%	47.89%
Bagging	87.22%	86.01%	39.87%
Logistic Regression	87.22%	85.18%	52.34%
LogitBoost	87.14%	87.43%	44.39%
SVM (linear kernel)	87.07%	68.83%	50.05%
RandomForest	86.99%	83.38%	38.01%
SVM (RBF kernel)	86.68%	59.74%	32.39%
AdaBoost	85.60%	83.48%	38.91%
J48 Decision Tree	84.91%	63.01%	39.70%
K-Nearest Neighbors	80.75%	59.11%	30.51%
Bayesian Network	79.83%	84.49%	49.43%
Naïve Bayes	64.28%	81.02%	42.02%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	86.06%	86.03%	66.23%
Prostate Cancer	87.68%	83.55%	55.60%
Lung Cancer	80.95%	79.70%	56.20%
Diabetes	83.74%	77.48%	42.13%

Supervised learning classification performance for each of the 12 classifiers using the content-based and context-based features, with classifier parameters optimized to minimize error rate (with feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Model Trees	87.96%	92.31%	68.18%
Bayesian Network	87.91%	92.84%	72.01%
SVM (linear kernel)	87.82%	77.85%	67.14%
Logistic Regression	87.82%	92.44%	67.51%
LogitBoost	87.77%	92.66%	68.81%
Bagging	87.26%	92.01%	67.93%
AdaBoost	87.17%	91.69%	69.45%
SMO (RBF kernel)	86.61%	73.15%	60.54%
Random Forest	86.57%	88.69%	65.71%
J48 Decision Tree	86.34%	79.24%	64.69%
Naïve Bayes	85.83%	90.08%	67.42%
K-Nearest Neighbors	83.93%	75.39%	58.98%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Bagging	90.84%	91.70%	63.30%
Logistic Model Trees	89.59%	91.43%	60.79%
SVM (linear kernel)	89.28%	73.35%	58.17%
Logistic Regression	89.12%	90.79%	58.41%
LogitBoost	88.81%	90.57%	58.69%
AdaBoost	88.66%	90.55%	58.05%
RandomForest	88.50%	89.39%	56.60%
SVM (RBF kernel)	88.35%	64.42%	43.03%
J48 Decision Tree	88.04%	76.29%	56.08%
K-Nearest Neighbor	87.73%	75.74%	57.58%
Bayesian Network	86.18%	91.04%	59.85%
Naïve Bayes	84.78%	88.56%	55.93%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Regression	84.12%	86.87%	60.92%
Logistic Model Trees	83.97%	86.71%	60.98%
LogitBoost	83.40%	86.07%	59.15%
Bagging	83.26%	85.35%	55.92%
SVM (RBF kernel)	82.98%	68.08%	52.27%
SVM (linear kernel)	82.54%	69.24%	54.34%
AdaBoost	82.40%	83.82%	58.10%
J48 Decision Tree	82.26%	78.15%	56.17%
Naïve Bayes	82.12%	86.10%	63.32%
K-Nearest Neighbors	80.83%	75.08%	56.71%
Bayesian Network	79.54%	86.13%	61.01%
RandomForest	78.97%	80.08%	51.34%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	87.22%	87.83%	46.67%
Logistic Regression	86.37%	86.07%	41.71%
SVM (linear kernel)	86.07%	60.99%	34.86%
AdaBoost	85.99%	83.91%	45.22%
Bagging	85.91%	86.52%	33.41%
RandomForest	85.91%	82.99%	41.50%
Logistic Model Trees	85.76%	85.76%	39.04%
J48 Decision Tree	85.30%	68.99%	36.43%
SVM (RBF kernel)	84.76%	50.00%	0.00%
Bayesian Network	83.07%	86.10%	52.23%
K-Nearest Neighbors	82.45%	65.97%	41.82%
Naïve Bayes	78.22%	84.05%	49.13%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	86.92%	86.53%	66.53%
Prostate Cancer	88.32%	84.49%	57.21%
Lung Cancer	82.20%	80.97%	57.52%
Diabetes	84.75%	77.43%	38.50%

Supervised learning classification performance for each of the 12 classifiers using the content-based and context-based features, with classifier parameters optimized to maximize f-measure (with no feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	88.56%	93.16%	71.07%
Logistic Model Trees	88.28%	92.87%	69.38%
Logistic Regression	88.05%	91.76%	69.77%
SVM (RBF kernel)	87.96%	77.11%	66.77%
SVM (linear kernel)	87.96%	79.87%	69.45%
RandomForest	87.45%	91.42%	65.47%
Bagging	87.26%	92.26%	69.23%
AdaBoost	87.22%	92.03%	70.07%
J48 Decision Tree	86.71%	75.46%	66.61%
Bayesian Network	83.05%	90.61%	66.49%
K-Nearest Neighbors	81.43%	70.37%	53.49%
Naïve Bayes	78.74%	86.19%	61.31%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	90.22%	90.87%	64.08%
SVM (RBF kernel)	90.07%	70.77%	56.35%
AdaBoost	89.90%	90.66%	63.34%
Logistic Model Trees	89.29%	92.20%	59.79%
Bagging	88.82%	90.59%	54.00%
SVM (linear kernel)	88.81%	76.22%	60.58%
RandomForest	88.67%	90.50%	54.56%
J48 Decision Tree	87.89%	78.48%	58.61%
Logistic Regression	86.95%	83.85%	57.80%
Bayesian Network	85.09%	91.57%	59.99%
K-Nearest Neighbors	83.85%	66.53%	43.31%
Naïve Bayes	81.83%	87.04%	54.81%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	83.54%	70.75%	57.29%
Logistic Model Trees	82.69%	86.45%	59.43%
RandomForest	82.11%	83.47%	57.43%
AdaBoost	81.68%	86.13%	57.71%
LogitBoost	81.40%	84.91%	56.88%
Bagging	81.39%	85.29%	53.85%
SVM (linear kernel)	80.97%	72.28%	57.99%
J48 Decision Tree	79.68%	67.26%	56.10%
Logistic Regression	79.54%	80.95%	56.42%
Bayesian Network	77.11%	84.18%	60.05%
Naïve Bayesian	76.25%	81.62%	59.60%
K-Nearest Neighbors	74.25%	63.05%	43.40%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	87.92%	87.63%	50.46%
SVM (linear kernel)	87.22%	71.60%	53.74%
Logistic Model Trees	87.14%	87.88%	45.92%
SVM (RBF kernel)	86.84%	60.53%	33.95%
Logistic Regression	86.38%	85.15%	50.51%
Bagging	86.07%	85.46%	37.24%
RandomForest	85.68%	80.01%	35.26%
AdaBoost	83.99%	83.63%	41.65%
J48 Decision Tree	83.37%	60.40%	40.35%
K-Nearest Neighbors	80.45%	57.78%	28.08%
Bayesian Network	80.44%	84.19%	50.46%
Naïve Bayes	65.36%	81.35%	42.74%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	86.05%	86.09%	66.59%
Prostate Cancer	87.62%	84.11%	57.27%
Lung Cancer	80.05%	78.86%	56.35%
Diabetes	83.40%	77.13%	42.53%

Supervised learning classification performance for each of the 12 classifiers using the content-based and context-based features, with classifier parameters optimized to maximize f-measure (with feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (linear kernel	87.86%	77.75%	67.28%
Bayesian Network	87.73%	92.57%	72.01%
Logistic Regression	87.68%	92.25%	66.78%
LogitBoost	87.68%	92.64%	68.89%
Logistic Model Tree	87.26%	91.84%	65.33%
AdaBoost	87.26%	92.01%	69.61%
Bagging	86.89%	91.80%	68.00%
SVM (RBF kernel)	86.71%	73.10%	60.65%
RandomForest	86.20%	89.16%	66.02%
J48 Decision Tree	85.78%	84.40%	63.65%
Naïve Bayes	85.32%	90.04%	65.83%
K-Nearest Neighbors	84.44%	79.84%	62.04%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Bagging	90.53%	91.87%	62.80%
RandomForest	90.06%	90.18%	59.97%
Logistic Regression	89.91%	91.94%	60.23%
AdaBoost	89.90%	90.80%	64.17%
Logistic Model Trees	89.75%	92.05%	59.54%
LogitBoost	89.59%	91.46%	58.87%
SVM (linear kernel)	88.97%	72.39%	55.67%
SVM (RBF kernel)	88.82%	66.47%	46.51%
J48 Decision Tree	87.58%	74.11%	57.66%
Bayesian Network	87.42%	91.95%	62.06%
K-Nearest Neighbors	87.42%	76.41%	57.43%
Naïve Bayes	87.26%	89.25%	59.11%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Model Trees	84.27%	87.87%	62.04%
Logistic Regression	84.12%	87.53%	62.26%
SVM (RBF kernel)	83.98%	69.30%	55.29%
SVM (linear kernel)	83.84%	70.75%	57.63%
Naïve Bayes	83.12%	86.53%	65.86%
Bagging	83.12%	86.66%	59.12%
AdaBoost	82.26%	86.61%	57.81%
LogitBoost	82.12%	85.15%	59.30%
J48 Decision Tree	82.12%	73.18%	57.26%
Bayesian Network	80.69%	86.13%	62.55%
RandomForest	80.54%	78.51%	58.44%
K-Nearest Neighbors	78.83%	74.21%	55.61%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Regression	86.45%	87.18%	44.19%
Logistic Model Trees	86.37%	87.40%	43.65%
LogitBoost	85.91%	87.12%	46.29%
RandomForest	85.91%	82.57%	43.39%
SVM (linear kernel)	85.84%	62.36%	36.49%
Bagging	85.45%	86.86%	37.25%
AdaBoost	84.91%	83.63%	36.19%
SVM (RBF kernel)	84.60%	50.65%	3.09%
J48 Decision Tree	83.91%	69.01%	45.16%
K-Nearest Neighbors	82.45%	66.18%	42.24%
Bayesian Network	82.45%	85.47%	51.59%
Naïve Bayes	76.99%	83.44%	48.64%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	86.73%	87.28%	66.34%
Prostate Cancer	88.93%	84.91%	58.67%
Lung Cancer	82.42%	81.04%	59.43%
Diabetes	84.27%	77.66%	39.85%

Supervised learning classification performance for each of the 12 classifiers using the bag-of-words model, bigrams, content-based and context-based features, with classifier parameters optimized to minimize error rate (with feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
RandomForest	88.84%	92.45%	70.98%
Logistic Model Trees	88.24%	92.31%	69.18%
Logistic Regression	88.10%	91.42%	68.75%
LogitBoost	88.00%	93.31%	69.40%
SVM (linear kernel)	87.96%	78.15%	67.96%
Bagging	87.91%	92.88%	70.06%
Bayesian Network	87.63%	93.35%	72.99%
SVM (RBF kernel)	87.17%	74.95%	63.66%
AdaBoost	87.08%	92.66%	69.25%
Naïve Bayes	86.85%	91.18%	70.50%
J48 Decision Tree	86.75%	79.65%	66.36%
K-Nearest Neighbors	84.85%	76.66%	63.05%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Model Trees	90.68%	91.64%	66.54%
Bagging	90.68%	91.78%	61.99%
SVM (linear kernel)	90.22%	72.35%	59.03%
Logistic Regression	90.22%	89.42%	64.27%
RandomForest	89.75%	90.71%	56.75%
SVM (RBF kernel)	89.60%	68.10%	51.55%
AdaBoost	89.44%	91.08%	65.59%
LogitBoost	88.51%	91.17%	56.91%
J48 Decision Tree	87.57%	69.31%	51.08%
Bayesian Network	87.42%	92.18%	62.93%
K-Nearest Neighbors	87.12%	73.68%	56.33%
Naïve Bayes	84.63%	88.86%	55.91%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	83.69%	70.01%	55.68%
RandomForest	83.55%	83.48%	61.17%
SVM (linear kernel)	83.40%	72.39%	59.53%
LogitBoost	83.12%	84.41%	59.68%
Bagging	82.69%	86.52%	54.83%
Logistic Model Trees	82.69%	83.65%	58.23%
AdaBoost	82.55%	84.54%	55.39%
J48 Decision Tree	82.26%	74.52%	57.17%
Logistic Regression	82.12%	81.57%	57.12%
Bayesian Network	81.69%	87.35%	65.22%
Naïve Bayes	81.26%	85.23%	64.27%
K-Nearest Neighbors	77.54%	69.94%	54.14%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	86.91%	86.79%	45.82%
AdaBoost	86.45%	85.61%	46.45%
SVM (RBF kernel)	86.14%	58.71%	29.60%
RandomForest	86.07%	83.44%	37.37%
Logistic Model Trees	85.99%	84.56%	43.47%
Bagging	85.76%	85.48%	34.38%
SVM (linear kernel)	85.37%	61.75%	36.47%
J48 Decision Tree	85.37%	66.67%	38.37%
Logistic Regression	85.30%	83.06%	41.94%
Bayesian Network	83.22%	85.99%	53.59%
K-Nearest Neighbors	82.91%	63.63%	39.21%
Naïve Bayes	75.90%	84.57%	50.59%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	87.45%	87.41%	68.51%
Prostate Cancer	88.82%	84.19%	59.07%
Lung Cancer	82.21%	80.30%	58.54%
Diabetes	84.62%	77.52%	41.44%

Supervised learning classification performance for each of the 12 classifiers using the bag-of-words model, bigrams, content-based and context-based features, with classifier parameters optimized to maximize f-measure (with feature selection).

Breast Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
RandomForest	88.61%	92.36%	70.71%
Logistic Model Trees	88.24%	92.31%	69.18%
Logistic Regression	88.10%	91.42%	68.75%
LogitBoost	88.05%	93.39%	69.77%
SVM (linear kernel)	87.96%	78.40%	68.21%
Bagging	87.91%	92.88%	70.06%
Bayesian Network	87.63%	93.35%	72.99%
SVM (RBF kernel)	87.12%	74.92%	63.58%
AdaBoost	87.08%	92.66%	69.25%
Naïve Bayes	86.85%	91.18%	70.50%
J48 Decision Tree	86.75%	79.13%	66.29%
K-Nearest Neighbors	84.85%	76.66%	63.05%

Prostate Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
Logistic Model Trees	90.68%	91.64%	66.54%
Logistic Regression	90.22%	89.42%	64.27%
SVM (linear kernel)	90.06%	72.69%	59.21%
RandomForest	89.91%	88.60%	57.14%
SVM (RBF kernel)	89.75%	68.60%	52.48%
Bagging	89.59%	91.13%	57.87%
AdaBoost	89.44%	91.08%	65.59%
LogitBoost	88.97%	90.71%	60.03%
J48 Decision Tree	87.57%	69.06%	51.64%
Bayesian Network	87.42%	92.18%	62.93%
K-Nearest Neighbors	87.12%	73.68%	56.33%
Naïve Bayes	84.63%	88.86%	55.91%

Lung Cancer			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
SVM (RBF kernel)	83.69%	70.01%	55.68%
RandomForest	83.55%	83.48%	61.17%
Bagging	83.40%	86.80%	58.20%
LogitBoost	82.83%	84.24%	59.98%
SVM (linear kernel)	82.83%	71.82%	58.50%
Logistic Model Trees	82.69%	83.65%	58.23%
AdaBoost	82.55%	84.54%	55.39%
Logistic Regression	82.12%	81.57%	57.12%
Bayesian Network	81.69%	87.35%	65.22%
Naïve Bayes	81.26%	85.23%	64.27%
J48 Decision Tree	80.83%	70.74%	54.11%
K-Nearest Neighbors	77.54%	69.94%	54.14%

Diabetes			
Classifier	**Accuracy**	**ROC Area**	**F-Measure**
LogitBoost	86.76%	85.69%	47.58%
AdaBoost	86.45%	85.61%	46.45%
SVM (RBF kernel)	86.07%	58.67%	29.51%
Bagging	86.06%	85.47%	37.40%
Logistic Model Trees	85.99%	84.56%	43.47%
RandomForest	85.68%	79.94%	38.45%
SVM (linear kernel)	85.37%	61.96%	36.94%
Logistic Regression	85.30%	83.06%	41.94%
J48 Decision Tree	83.76%	63.23%	36.79%
Bayesian Network	83.22%	85.99%	53.59%
K-Nearest Neighbors	82.91%	63.63%	39.21%
Naïve Bayes	75.90%	84.57%	50.59%

Average performance for all datasets:

Dataset	**Average Accuracy**	**Average ROC Area**	**Average F-Measure**
Breast Cancer	87.43%	87.39%	68.53%
Prostate Cancer	88.78%	83.97%	59.16%
Lung Cancer	82.08%	79.95%	58.50%
Diabetes	84.46%	76.86%	41.83%

Appendix B

Transfer of Learning Experiment Results

Transfer of learning experiments using Content + Context features, with classifier parameters optimized to minimize error rate.

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Lung	76.68%	81.27%	59.48%
Breast → Prostate	87.65%	85.80%	58.07%
Lung → Breast	83.41%	82.39%	51.25%
Lung → Prostate	87.10%	82.39%	46.75%
Prostate → Breast	83.71%	82.43%	58.39%
Prostate → Lung	74.82%	76.58%	52.59%

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Diabetes	78.32%	80.28%	49.18%
Diabetes → Breast	82.71%	81.44%	45.83%
Cancers → Diabetes	80.21%	79.87%	47.94%
Diabetes → Cancers	82.25%	80.89%	45.16%

Transfer of learning experiments using Content + Context features, with classifier parameters optimized to minimize error rate (with feature selection).

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Lung	79.31%	82.42%	60.85%
Breast → Prostate	89.26%	86.91%	61.67%
Lung → Breast	84.48%	83.32%	51.91%
Lung → Prostate	88.32%	83.89%	49.93%
Prostate → Breast	85.91%	85.39%	63.07%
Prostate → Lung	77.75%	79.57%	56.55%

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Diabetes	81.21%	82.23%	51.15%
Diabetes → Breast	83.78%	83.34%	45.83%
Cancers → Diabetes	83.41%	81.79%	50.22%
Diabetes → Cancers	83.26%	82.03%	44.32%

Transfer of learning experiments using Content + Context features, with classifier parameters optimized to maximize f-measure.

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Lung	76.61%	81.21%	59.40%
Breast → Prostate	87.59%	85.85%	57.88%
Lung → Breast	83.50%	82.45%	52.83%
Lung → Prostate	86.98%	82.52%	47.19%
Prostate → Breast	83.72%	82.14%	59.22%
Prostate → Lung	74.22%	76.03%	52.75%

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Diabetes	78.37%	80.34%	49.38%
Diabetes → Breast	82.64%	79.97%	46.65%
Cancers → Diabetes	80.19%	79.87%	47.94%
Diabetes → Cancers	82.13%	79.32%	45.91%

Transfer of learning experiments using Content + Context features, with classifier parameters optimized to maximize f-measure (with feature selection).

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Lung	79.26%	82.39%	60.71%
Breast → Prostate	89.26%	86.91%	61.67%
Lung → Breast	84.52%	83.37%	52.21%
Lung → Prostate	88.36%	83.91%	50.14%
Prostate → Breast	85.91%	85.37%	63.21%
Prostate → Lung	77.83%	79.57%	56.91%

Dataset	Average Accuracy	Average ROC Area	Average F-Measure
Breast → Diabetes	81.23%	82.23%	51.20%
Diabetes → Breast	83.63%	82.16%	45.78%
Cancers → Diabetes	83.10%	81.82%	49.70%
Diabetes → Cancers	83.06%	80.98%	44.19%

BIBLIOGRAPHY

[1] Abe, S., Inui, K., Hara, K., Morita, H., Sao, C., Euchar, M., Sumita, A., Murakami, K., Matsuyoshi, S. 2011. Mining personal experiences and opinions from Web documents. Web Intelligence and Agent Systems: An International Journal, pp. 1–13. 2011.

[2] Agarwal, A. Boyi, X. Vovsha, I. Rambow, O., Passonneau, R. 2011. Sentiment Analysis of Twitter Data. In Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38, Portland, Oregon, 23 June 2011.

[3] Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T. 2012. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In Proceedings of the 23rd ACM conference on Hypertext and social media (HT '12). ACM, New York, NY, USA, pp. 187-196.

[4] Alemi, F., Torri, M., Clementz, L., Aron, D.C. 2012. Feasibility of Real-Time Satisfaction Surveys Through Automated Analysis of Patients' Unstructured Comments and Sentiments. Quality Management in Health Care, Volume 21, Issue 1, pp. 9-19.

[5] Barbosa, L. and Feng, J. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In Proceedings of COLING 2010: Poster Volume, pages 36–44, Beijing, August 2010.

[6] Bennett, G.G. and Glasgow, R.E. 2009. The delivery of public health interventions via the Internet: actualizing their potential. Annual Review of Public Health 2009;30:273-292.

[7] Bhattacharya, S., Train, H., Srinivasa, P. 2012. Discovering Health Beliefs in Twitter. In Proceedings of the Association for the Advancement of Artificial Intelligence, AAAI Technical Report FS-12-05, pp. 2-7. 2012.

[8] Black, P. E. 2004. Ratcliff/Obershelp pattern recognition. Dictionary of Algorithms and Data Structures, U.S. National Institute of Standards and Technology. http://www.nist.gov/dads/HTML/ratcliffObershelp.html

[9] Blitzer, J., Dredze, M., Pereira, F. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 440–447, Prague, Czech Republic, June 2007.

[10] Bobicev, V., Sokolova, M., Jafer, Y., Schramm, D. 2012. Learning Sentiments from Tweets with Personal Health Information. In Proceedings of Canadian AI 2012, LNAI 7310, pp. 37–48, 2012.

[11] Bollen, J., Mao, H., and Zeng, X. 2011. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), March 2011, pp. 1-8.

[12] Bollen, J., Pepe, A., and Mao, H. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Proceedings of the 5th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media (pp. 450- 453). Menlo Park, CA: The AAAI Press.

[13] Brownstein, C.A., Brownstein, J.S., Williams, D.S., Wicks, P., and Heywood, J.A. 2009. The power of social networking in medicine. Nature Biotechnology vol. 27, pp. 888-890, 2009.

[14] Brubaker, J.R., Kivran-Swaine, F., Taber, L., Hayes, G.R. 2011. Grief-Stricken in a Crowd: The Language of Bereavement and Distress in Social Media. In Proceedings of ICWSM, 2012. AAAI Press.

[15] Cambria, E., Benson, T., Eckl, C., and Hussain, A. 2012. Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Systems with Applications, vol 39, issue 12, September 2012, pp. 10533-10543.

[16] Cambria, E., Hussain, A. et al. 2010. Sentic Computing for Patient Centered Applications. In Proceedings of ICSP 2010, pp. 1279-1282.

[17] Chee, B., Berlin, R., and Schatz, B. 2009. Measuring population health using personal health messages. 2009 AMIA Annual Symposium Proceedings, pp. 92-96.

[18] Chee, B., Berlin, R., and Schatz, B. 2011. Predicting adverse drug events from personal health messages. In Proceedings of AMIA, 2011.

[19] Chen, H., Compton, S., Hsiao, O. 2013. DiabeticLink: A Health Big Data System for Patient Empowerment and Personalized Healthcare. In Proceedings of ICSH 2013, Lecture Notes in Computer Science Volume 8040,, pp. 71–83, 2013.

[20] Chen, Z. and Koh, P.W. 2011. Analyzing Patient Interactions within Cancer Support Groups. Stanford University.

[21] Cheng, A. and Zhulyn, O. 2012. A System For Multilingual Sentiment Learning On Large Data Sets. In Proceedings of COLING 2012: Technical Papers, pages 577–592, COLING 2012, Mumbai, December 2012.

[22] Cherry, C., Mohammad, S.M., and de Bruijn, B. 2012, Binary classifiers and latent sequence models for emotion detection in suicide notes. Biomedical Informatics Insights, vol. 5, no. Suppl 1, pp. 147-154, 2012.

[23] Chetviorkin, I. and Loukachevitch, N. 2012. DomEx: Extraction of Sentiment Lexicons for Domains and Meta-Domains. In Proceedings of COLING 2012: Demonstration Papers, pages 77–86, COLING 2012, Mumbai, December 2012.

[24] Choudhury, M.D., Counts, S, Horvitz, E. 2013. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth. In Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work, 2013.

[25] Choudhury, M.D., Counts, S, Horvitz, E. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3267-3276.

[26] Collins, M. 2002. Machine Learning Methods in Natural Language Processing. MIT CSAIL.

[27] Crouch, S., an Khosla, R. 2012. Sentiment Analysis of Speech Prosody for Dialogue Adaptation in a Diet Suggestion Program. SIGHIT Rec. Volume 2, Issue 1 (March 2012), pp. 8.

[28] Davison, K.P., Pennebaker, J.W., and Dickerson, S.S. 2000. Who talks? The social psychology of illness support groups. American Psychologist, Vol 55(2), Feb 2000, 205-217.

[29] Denecke, K., Taytsarau, M., Palpanas, T., Brosowski, M. 2009. Topic-related Sentiment Analysis for Discovering Contradicting Opinions in Weblogs. Technical Report, University of Trento (2009).

[30] Desai, T., Shariff, A., Shariff, A., Kats, M., Fang, X., et al. 2012. Tweeting the Meeting: An In-Depth Analysis of Twitter Activity at Kidney Week 2011. PLoS ONE, vol. 7, issue 7, e40253.

[31] Desmet, B., and Hoste, V. 2012. Emotion detection in suicide notes. Expert Systems with Applications, 40 (2013) pp. 6351–6358.

[32] Dini, L., Mazzini, G. 2002. Opinion classification through information extraction. In Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields, pp. 299-310.

[33] Foster, J., Cetinoglu, O., Wagner, J., et al. 2011. #hardtoparse: POS Tagging and Parsing the Twitterverse. Analyzing Microtext: Papers from the 2011 AAAI Workshop (WS-11-05), pp. 20-25.

[34] Fox, S. 2012. Pew Internet: Health. Pew Internet & American Life Project, February 20, 2012, http://pewinternet.org/Commentary/2011/November/Pew-Internet-Health.aspx, accessed on March 8, 2013.

[35] Gillingham, G., Conway, M.A., Chapman, W.W., Casale, M.B., and Pettigrew, K.B. 2012. #wheezing: A Content Analysis of Asthma-Related Tweets. In Proceedings of ISDS, 2012.

[36] Goeuriot, L., Na, J., Kyaing, W., Foo, S., Khoo, C., Theng, Y., Chang, Y. 2011. Textual and Informational Characteristics of Health-Related Social Media Content: A Study of Drug Review Forums. In Proceedings of the Asia Pacific Conference Library & Information Education & Practice, 2011, pp. 548-557.

[37] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A, Donaldson, L. 2013. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Qual Saf. 2013 March, 22(3), pp. 251-255.

[38] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson L. 2012. Machine learning and sentiment analysis of unstructured free-text information about patient experience online. The Lancet, vol. 380, pp. S10.

[39] Greene, J.A., Choudhry, N.K., Kilabuk, E., Shrank, W.H. 2010. Online Social Networking by Patients with Diabetes: A Qualitative Evaluation of Communication with Facebook. Journal of General Internal Medicine (2010), vol. 26, no. 3, pp. 287–292.

[40] Handelman, L.D. and Lester, D. 2007. The content of suicide notes from attempters and completers. Crisis 28(2), pp. 102–104, 2007.

[41] Hastie, T., Tibshirani, R., Frieman, J. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2001.

[42] He, Y., Lin, C., Alani, H. 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 19-24 Jun 2011, Portland, Oregon, USA.

[43] Heaivilin, A., Gerbert, B., Page, J.E., Gibbs, J.L. 2011. Public Health Surveillance of Dental Pain via Twitter. Journal of Dental Research (2011), vol. 90, pp. 1047-1051.

[44] Huang, Y., Goh, T., and Liew, C.L. 2007. Hunting Suicide Notes in Web 2.0 – Preliminary Findings. Ninth IEEE International Symposium on Multimedia, 2007, pp. 517-521.

[45] Inui, K., Abe, S.,Hara, K., Morita, H., Sao, C., Eguchi, M., Sumida, A., Murakami, K., and Matsuyoshi, S. 2008. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In Proceedings of 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 314–321, 2008.

[46] Jamison-Powell, S., Linehan, C., Daley, L., Garbett, A., and Lawson, S. 2012. I can't get no sleep: discussing #insomnia on twitter. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI '12). ACM, New York, NY, USA, 1501-1510.

[47] Jialin Pan, S., Yang, Q. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, vol.22, no.10, pp. 1345-1359.

[48] Jiang, Y., Liao, Q.V., Cheng, Q., Berlin, R.B., Schatz, B.R. 2012 Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. In Proceedings of the American Medical Informatics Association Annual Symposium, pp. 417–426.

[49] Jijkoun, V., de Rijke, M., Weerkamp, W., Ackermans, P., Geleijnse, G. 2010. Mining User Experiences from Online Forums: An Exploration. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 17–18, Los Angeles, California, June 2010.

[50] Jones, N. and Bennell, C. 2007. The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. Archives- of Suicide- Research:- Official- Journal- of- the- International- Academy- for Suicide Research. 2007;11(2):219.

[51] Kaiser, C. and Bodendorf, F. 2012. Mining Patient Experiences on Web 2.0 A Case Study in the Pharmaceutical Industry. In Proceedings of the 2012 Service Research and Innovation Institute Global Conference, pp. 139-145. 2012.

[52] Kessler, W. and Schutze, H. 2012. Classification of Inconsistent Sentiment Words Using Syntactic Constructions. In Proceedings of COLING 2012: Posters, pages 569–578, COLING 2012, Mumbai, December 2012.

[53] Kim, H., Castellanosb, M.G., Hsub, M., Zhaia, C., Dayalb, U., Ghoshb, R. 2013. Ranking Explanatory Sentences for Opinion Summarization. In Proceedings of SIGIR 2013, pp. 1069-1072, July 28–August 1, 2013, Dublin, Ireland.

[54] Kucuktunc, O., Weber, I., Cambazoglu, B.B., Ferhatosmanoglu, H. 2012. A Large-Scale Sentiment Analysis for Yahoo! Answers. In Proceedings of WSDM 2012, pp. 633-642, February 8–12, 2012, Seattle, Washington, USA.

[55] Lamb, A., Paul, M.J., and Dredeze, M. 2012. Investigating Twitter as a Source for Studying Behavioral Responses to Epidemics. AAAI Fall Symposium Series, North America, oct. 2012

[56] Lansdall-Welfare, T., Lampos, V., and Cristianini, N. 2012. Nowcasting the mood of the nation. Significance, vol 9, Issue 4, pp. 26–28, August 2012.

[57] Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., Gonzalez, G. 2010. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from

User Posts to Health-Related Social Networks. In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pp. 117–125, Uppsala, Sweden, 15 July 2010.

[58] Lewis, M.M., Welliver, M.D., Leach, S. 2012. The Doctor of Nursing Practice: A Sentiment Analysis and Credential Correlation. International Journal of Advanced Nursing Studies, vol 1 (2012), no. 3, pp. 43-57.

[59] Li, T., Chau, M., Wong, P., and Yip, P. 2012. A Hybrid System for Online Detection of Emotional Distress. In Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics, pp. 73-80.

[60] Liu, B. 2010. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition (editors: N. Indurkhya and F. J. Damerau), 2010.

[61] Liu, B. and Zhang, L. 2012. A Survey of Opinion Mining and Sentiment Analysis. Mining Text Data, pp. 415-463. 2012. New York.

[62] Liu, K., Li, W., Guo, M. 2012. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 1678-1684, 2012.

[63] Liu, S., Agam, G., Grossman, D. 2012. Generalized Sentiment-Bearing Expression Features for Sentiment Analysis. In Proceedings of COLING 2012: Posters, pages 733–744, COLING 2012, Mumbai, December 2012.

[64] Lo, A.S., Esser, M.J., Gordon, K.E. 2010. YouTube: A gauge of public perception and awareness surrounding epilepsy. Epilepsy & Behavior 17 (2010), pp. 541–545.

[65] Luckycx, K., Vaassen, F., Peersman, C., Daelemans, W. 2012. Fine-Grained Emotion Detection in Suicide Notes: A Thresholding Approach to Multi-Label Classification. Biomed Inform Insights. 2012; 5(Suppl 1): pp. 61–69.

[66] McCart, J.A., Finch, D.K., Jarman, J. et al. 2012. Using Ensemble Models to Classify the Sentiment Expressed in Suicide Notes. Biomed Inform Insights. 2012; 5(Suppl 1): pp. 77–85.

[67] McNeil, K., Brna, P.M., Gordon, P.E. 2012. Epilepsy in the Twitter era: A need to re-tweet the way we think about seizures. Epilepsy & Behavior 23 (2012), pp. 127–130.

[68] Mejova, Y. Srinivasan, P., Boynton, B. 2013. GOP Primary Season on Twitter: "Popular" Political Sentiment in Social Media. In Proceedings of WSDM 2013, pp. 517-525, February 4–8, 2013, Rome, Italy.

[69] Melton, G.B. and Hripcsak, G. 2005. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. Journal of the American Medical Informatics Association Volume 12, No. 4, pp. 448-457.

[70] Miao, Q., Zhang, S., Meng, Y., Fu, Y., Yu, H. 2012. Healthy or Harmful? Polarity Analysis Applied to Biomedical Entity Relationships. In Proceedings of the 12th Pacific Rim international conference on Trends in Artificial Intelligence (PRICAI'12) pp. 777–782, 2012.

[71] Mishne, G. and De Rijke, M. 2006. Capturing global mood levels using blog posts. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.

[72] Mohtarami, M., Amiri, H., Lan, M., Tran, T.P., Tan, C.L. 2012. Sense Sentiment Similarity: An Analysis. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 1706-1712, 2012.

[73] Mohtarami, M., Lan, M., Tan, C.L. 2013. From Semantic to Emotional Space in Probabilistic Sense Sentiment Analysis. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pp. 711-717, 2013.

[74] Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R. 2013. Spotting Opinion Spammers using Behavioral Footprints. In Proceedings of KDD 2013, pp. 632-640, August 11–14, 2013, Chicago, Illinois, USA.

[75] Mukherjee, S., Bhattacharyya, P. 2012. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In Proceedings of COLING 2012: Technical Papers, pp. 1847–1864, COLING 2012, Mumbai, December 2012.

[76] Murff, H.J., FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. Journal of the American Medical Association, Aug 24, vol. 306, no. 8, pp. 848-855.

[77] Murphy, K. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press, August 2012.

[78] Myaeng, S.H., Jung, Y., Jeong, Y. 2012. Experiential Knowledge Mining. Foundations and Trends in Web Science Vol. 4, No. 1 (2012), pp. 1–102.

[79] Myslin, M., Zhu, S., Chapman, W., and Conway, M. 2013. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. J Med Internet Res. 2013 August; 15(8): e174.

[80] Na, J., Kyaing, W., Khoo, C., Foo, S., Chang, Y., Leng, T. 2012. Sentiment Classification of Drug Reviews Using a Rule-Based Linguistic Approach. In ICADL, volume 7634 of Lecture Notes in Computer Science, page 189-198, 2012.

[81] Neustein, A. 2007. Sequence Package Analysis: A New Natural Language Understanding Method for Intelligent Mining of Recordings of Doctor-Patient Interviews and Health-Related Blogs. In Proceedings of the International Conference on Information Technology (ITNG '07). IEEE Computer Society, Washington, DC, USA, 431-438.

[82] Niu, Y. and Hirst, G. 2004. Analysis of semantic classes in medical text for question answering. In ACL 2004 Workshop on QA in Restricted Domains.

[83] Niu, Y., Zhu, X., Li, J., Hirst, G. 2005. Analysis of polarity information in medical text. Proceedings of the AMIA Annual Symposium. pp. 570–574, 2005.

[84] Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, Y., Portier, K., Greer, G. 2013. In Proceedings of the 2013 International Conference on Social Intelligence and Technology (SOCIETY 2013), pp. 109-113.

[85] Ogilvie D., Stone P., Shneidman E. 1966. Some characteristics of genuine versus simulated suicide notes. National Institute of Mental Health Bulletin of Suicidology. 1966, pp. 27–32.

[86] Pak, A. and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of LREC 2010 (pp. 1320–1326). Paris: European Language Resource Association.

[87] Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, vol. 2, no. 1-2(2008), pp. 1–135, 2008.

[88] Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

[89] Park, K.C., Jeong Y., Myaeng, S.H. 2010. Detecting Experiences from Weblogs. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1464–1472, Uppsala, Sweden, 11-16 July 2010.

[90] Parker, J., Wei, Y., Yayes, A., Freider, O., Goharian, N. 2013. A Framework for Detecting Public Health Trends with Twitter. In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, 2013.

[91] PatientsLikeMe. http://patientslikeme.com

[92] Paul, M.J. and Dredze, M. 2011. You are what you tweet: Analyzing Twitter for public health. Artificial Intelligence (2011), pp. 265-272.

[93] Paul, M.J., Wallace, B.C., and Dredze, M. 2013. What Affects Patient (Dis)satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. In AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI), 2013.

[94] Pestian, J., Matykiewicz, P., Cohen, B., Grupp-Phelan, J., Richey, L.A., Meyers, G., Canter, C. M., Sorter, M.T. 2013. Suicidal Thought Markers: A Controlled Trial Examining the Language of Suicidal Adolescents. Preprint.

[95] Pestian, J., Matykiewicz, P., Grupp-Phelan, J., Lavanie, S.A., Combs, J., and Kowatch, R. 2008. Using Natural Language Processing to classify suicide notes. AMIA Annu Symp Proceedings. 2008:1091.

[96] Pestian, J., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K.B., Hurdle, J. and Brew, C. 2012. Sentiment analysis of suicide notes: a shared task. Biomedical Informatics Insights, vol. 5, no. Suppl 1, pp. 3-16, 2012.

[97] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., and Leenaars, A. 2011. Suicide note classification using natural language processing: a content analysis. Biomedical Informatics Insights. 2011;3, pp. 19–28.

[98] Pieterse, V., and Black, P. Algorithms and Theory of Computation Handbook, CRC Press LLC, 1999, "Levenshtein distance", in Dictionary of Algorithms and Data Structures [online], 22 August 2013. http://www.nist.gov/dads/HTML/Levenshtein.html

[99] Ponomareva, N. and Thalwell, M. 2012. Biographies or Blenders: Which Resource Is Best for Cross-Domain Sentiment Analysis? In Proceedings of CICLing 2012, Part I, Lecture Notes in Computer Science 7181, pp. 488–499, 2012.

[100] Prabhu, V. 2012. Using Social Media to Gauge Reaction to the United States Preventive Services Task Force's (USPSTF) Report Twitter as an Investigative Tool. The Journal of Urology (0022-5347), 187 (4), p. e170.

[101] Prochaska, J.J., Penchmann, C., Kim, R., Leonhardt, J.M. 2011. Twitter=quitter? An analysis of Twitter quit smoking social networks. Tobacco Control, 2011.

[102] Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, D., Yen, J., Greer, G.E., and Portier. K. Get Online Support, Feel Better-Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. Proceedings of the Third IEEE International Conference on Social Computing, Boston, Massachusetts, USA, 2011.

[103] Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. 2012. Tracking "gross community happiness" from tweets. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12). ACM, New York, NY, USA, 965-968.

[104] Quercia, D., Seaghdha, D., and Crowcroft, J. 2012. Talk of the City: Our Tweets, Our Community Happiness. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, pp. 555-558.

[105] Read, J., Velldal, E. and Øvrelid, L. 2012. Topic Classification for Suicidology. JCSE, vol. 6, no. 2, pp.143-150, June 20.

[106] Ryu, J., Jung, Y., Kim, K., Myaeng, S.H. 2010. Automatic Extraction of Human Activity Knowledge from Method-Describing Web Articles. In Proceedings of the 1st Workshop on Automated Knowledge Base Construction, pp. 16-23, Grenoble, France, 2010.

[107] Saif, H., He, Y., Alani, H. 2012. Semantic Sentiment Analysis of Twitter. In Proceedings of ISWC 2012, Part I, Lecture Notes in Computer Science 7649, pp. 508–524, 2012.

[108] Salathé, M. and Bonhoeffer, S. 2008. The effect of opinion clustering on disease outbreaks. J R Soc Interface 5: 1505–1508.

[109] Salathé, M. and Khandelwal, S. 2011. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. PLoS Comput Biol 7(10): e1002199.

[110] Salathé, M., Bengtsson, L., Bodnar, T.J., Brewer, D.D., Brownstein, J.S., et al. 2012. Digital Epidemiology. PLoS Comput Biol 8(7): e1002616.

[111] Salathé, M., Vu, D., Khandelwal, S., and Hunter, D. 2012. The Dynamics of Health Behavior Sentiments on a Large Online Social Network. Preprint, arXiv:1207.7274.

[112] Sarker, A., Mollá, D. and Paris, C. 2011. Outcome Polarity Identification of Medical Papers. Proceedings of the 2011 Australasian Language Technology Workshop (ALTA 2011), Canberra, Australia.

[113] Sauer, C. and Roth-Berghofer, T. 2012. Solution Mining for Specific Contextualised Problems: Towards an Approach for Experience Mining. In Proceedings of International World Wide Web Conference (WWW 2012), pp. 729-738, 2012.

[114] Scanfeld, D., Scanfeld, V., and Larson, E.L. 2010. Dissemination of health information through social networks: Twitter and antibiotics, American Journal of Infection Control, Volume 38, Issue 3, April 2010, Pages 182-188, ISSN 0196-6553.

[115] Signorini, A., Segre, A.M., and Polgreen, P.M. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS ONE 6(5): e19467.

[116] Silva, I.S., Gomide, J., Veloso, A., Meira, W., Ferreira, R. 2011. Effective Sentiment Stream Analysis with Self-Augmenting Training and Demand-Driven Projection. In Proceedings of SIGIR 2011, pp. 583-592, July 24–28, 2011, Beijing, China.

[117] Smith, P. and Lee, M. 2012. Cross-discourse Development of Supervised Sentiment Analysis in the Clinical Domain. In Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 79–83.

[118] Smyth, B., Champin, P., Briggs, P., Coyle, M. 2009. The case-based experience web. Expert Systems with Applications, Volume 40, Issue 2, 1 February 2013, pp. 500–507.

[119] Steele, R. and Min, K. Health System Zeitgeist: How Tweets Can Provide Real-time Insight into the Health System. Proceedings of IEEE 10th International Conference on Industrial Informatics, Beijing, 2012.
[120] Swaminathan, A., Sharma, R., Yang, H. 2010. Opinion Mining for Biomedical Text Data: Feature Space Design and Feature Selection. In the Ninth International Workshop on Data Mining in Bioinformatics, BIOKDD 2010.
[121] Tan, S., Cheng, X., Wang, Y., Xu, H. 2009. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. In Proceedings of ECIR 2009, Lecture Notes in Computer Science 5478, pp. 337–349, 2009.
[122] Tan, S., Wu, G., Tang, H., Cheng, X. 2007. A Novel Scheme for Domain-transfer Problem in the context of Sentiment Analysis. In Proceedings of CIKM 2007, pp. 979-982, November 6-8, 2007, Lisboa, Portugal.
[123] The World Health Organization. 2014. "Health topics: Public health surveillance." http://www.who.int/topics/public_health_surveillance/en/
[124] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), pp. 2544-2558.
[125] Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
[126] Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the Association for Computational Linguistics. pp. 417–424.
[127] Twitter Well-Being Tracker. http://wellbeingtracker.meyouhealth.com
[128] Twitter, Form S-1. http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm
[129] Twitter. http://twitter.com
[130] Wang, W., Chen, L., Tan, M., Wang, S., and Sheth, A.P. 2012. Discovering Fine-grained Sentiment in Suicide Notes. Biomed Inform Insights. 2012; 5(Suppl 1): pp. 137–145.
[131] WeFeelFine. http://wefeelfine.org
[132] Wicentowsky, R. and Sydes, M.R. 2012. Emotion Detection in Suicide Notes using Maximum Entropy Classification. Biomedical Informatics Insights 2012:5 (Suppl. 1) pp. 51–60.
[133] Wu, Q. and Tan, S. 2011. A two-stage framework for cross-domain sentiment classification. In Proceedings of Expert Systems with Applications, 2011, pp 14269–14275.
[134] Wu, Q., Tan, S., Zhai, H., Zhang, G., Duan, M., Cheng, X. 2009. SentiRank: Cross-Domain Graph Ranking for Sentiment Classification. In Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, pp. 309-314, 2009.
[135] Xia, L., Gentile, A.L., Munroe, J., Irea, J. 2009. Improving Patient Opinion Mining through Multi-step Classification. In proceeding of: Text, Speech and Dialogue, 12th

International Conference, TSD 2009, Pilsen, Czech Republic, September 13-17, 2009.

[136] Xia, R. and Zong, C. 2011. A POS-based Ensemble Model for Cross-domain Sentiment Classification. In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 614–622, Chiang Mai, Thailand, November 8-13, 2011.

[137] Xu, X., Tan, S., Liu, Y., Cheng, X., Lin, Z., Guo, J. 2012. Find Me Opinion Sources in Blogosphere: A Unified Framework for Opinionated Blog Feed Retrieval. In Proceedings of WSDM 2012, February 8-12, 2012, Seattle, Washington, USA.

[138] Xu, Y., Wang, Y., Liu, J., Tu, Z., Sun, J.T., Tsujii, J., and Chang E. Suicide note sentiment classification: a supervised approach augmented by web data. Biomedical Informatics Insights, vol. 5, no. Suppl 1, pp. 31-41, 2012.

[139] Yalamanchi, D. 2011. Sideffective: System to Mine Patient Reviews: Sentiment Analysis. (Master's thesis).

[140] Yang, C., Yang, H., Jiang, L., and Zhan M. 2012. Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media. In Proceedings of the 2012 international workshop on smart health and wellbeing, pp. 33-40.

[141] Yang, H., Willis, A., de Roeck, A., and Nuseibeh, B. 2012. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes. Biomedical Informatics Insights 2012:5 (Suppl. 1) pp. 17–30.

[142] Yates, A. and Goharian, N. 2013. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In Proceedings of the 35th European conference on Advances in Information Retrieval, pp. 816–819, 2013.

[143] Yoon, S. and Bakken, S. 2012. Methods of Knowledge Discovery in Tweets. In Proceedings of the 11th International Congress on Nursing Informatics (NI 2012), pp. 463-467.

[144] Yoshida, Y., Hirao, T., Iwata, T., Nagata, M., Matsumoto, Y. 2011. Transfer Learning for Multiple-Domain Sentiment Analysis — Identifying Domain Dependent/Independent Word Polarity. In Proceedings of Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 1286-1291, 2011.

[145] Yu, B. 2011. The emotional world of health online communities. Proceedings of iConference 2011, February 8-11, 2011. Seattle, WA.

[146] Zhai, Z., Liu, B., Xu, H., Jia, P. 2011. Clustering Product Features for Opinion Mining. In Proceedings of WSDM 2011, pp. 347-354, February 9–12, 2011, Hong Kong, China.

[147] Zhang, R., Pakhmomov, S., et al. 2012. Automated Assessment of Medical Training Evaluation Text. In Proceedings of the AMIA, 2012, pp. 1459–1468.

[148] Zhao, K., Greer, G., Qiu, B., Mitra, P., Portier, K., Yen, J. 2014. Finding influential users of an online health community: a new metric based on sentiment influence. Journal of the American Medical Informatics Association, January 2014.

[149] Zhao, K., Qiu, B., Caragea, C., Wu, D., Mitra, P., Yen, J., Greer, G.E., and Portier, K. 2011. Identifying Leaders in an Online Cancer Survivor Community. Proceedings of

the 21st Annual Workshop on Information Technologies and Systems (WIST 2011), Shanghai, China, December 3-4, 2011.

[150] Ziebland, S., Chapple, A., Dumelow, C., Evans, J., Prinjha, S., and Rozmovits, L. 2004. How the internet affects patients' experience of cancer: a qualitative study. BMJ 2004;328:564.

ACADEMIC VITA

William Murphy

Education

- Masters of Science in Information Science and Technology, May 2014
- Bachelor of Science in Information Science and Technology, May 2014

Publications

Michelle Newman, John Yen, Prasenjit Mitra, William Murphy, Nicholas Jacobson and Hanjoo Kim. (2013). "Supervised Machine Learning Classification of Journals Entries About Emotional Personal Experiences." *In Proceedings of American Medical Informatics Association Symposium, 2013.*