Identification and Characterization of Diseases on Social Web

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover zur
Erlangung des Grades
Doktor der Naturwissenschaften
Dr.rer.nat
genehmigte Dissertation

von

M.Sc. **Mustafa Sofean**

geboren am 09.12.1980
in Taiz - Yemen

1. Referent: Prof. Dr. Matthew Smith

2. Referentin: Prof. Dr.-Ing. Gabriele von Voigt

Tag der Promotion: 03. Mai 2017

Abstract

Social network sites such as Twitter, MySpace and Facebook are web-based services that allow individuals to instantly communicate with large networks of friends, acquaintances and colleagues, and share activities, interests, or real-life connections. Millions of users share their experiences, thoughts, and behaviors online through these sites. They also share their symptoms and diseases in real-time. Twitter is a popular social website, which can be a source and distributor of public health information via short 140-character microblogs that are informally known as tweets and characterized by a real-time nature. The work in this dissertation considers those microblogs as a source of information. Moreover, it presents techniques for leveraging the wealth of available social web documents to track, identify, and characterize different types of diseases based on names of normal diseases, viruses, bacteria causing diseases, and symptoms, as well as people's behavior toward health-related topics are presented. By automatically tracking, identifying, and analyzing disease-related messages, this work can ultimately offer substantial improvements in disease surveillance, and forecast future disease levels such as influenza rates. To understand the nature of diseases that exist on Twitter, this dissertation first characterizes a large set of Twitter messages(tweets) that are associated to public health. Specifically, this work analyzes tweets to find out what the users are posting about their health and behaviors, to understand the nature of public health related data on Twitter, and to develop techniques that perform tasks to achieve the goal and meet challenges. The huge amount of irrelevant data at hand in those microblogs requires sophisticated filtering methods to identify relevant postings. Therefore, annotation guidelines are developed to create

a dataset and build a classifier to distinguish between relevant and irrelevant disease-monitoring. An established set of case definition guidelines within the medical intelligence community is exploited for defining disease-reporting messages. This work incorporates salient information and creates annotation guidelines to determine the feature set that produces the best sparse-text classifier to identify relevant tweets. Because of the immense volume of Twitter messages, an annotation tool is developed to enable the annotator to collect the training dataset easily. Then, two classification algorithms are evaluated on the labeling dataset with several feature sets. To track and identify disease-related messages on Twitter, a novel real-time filtering system that uses data mining techniques is developed to crawl, index, extract, cleanse, and classify postings in real-time overtime. The real-time filtering system tracks status updates of the population in real-time and has the ability to remove spams (e.g., shot-related tweets, vaccination-related tweets, and Bieber-fever-related tweets). Furthermore, the system is using the state-of-the-art text classification to filter pure-disease-related tweets. Specifically, the system aims to select high-quality relevant messages that reflect useful information about diseases and outbreaks. The run-time performance of the filtering system is evaluated and the accuracy performance of the classifier model is tested. To understand the meaning of disease-related messages, first the challenges faced when using the traditionally named entity recognition tools are presented. Then, the semantic of collecting tweets is discovered by developing a a novel disease entity recognition framework to determine the syntactic structure of postings, extract medical entities and locations, and recognize human disease-related events. Additionally, the disease entity recognition system is evaluated with selected tweets for different diseases. In order to provide a proof of the techniques in this work that are used to track, identify, and analyze disease-reporting messages, all these techniques were combined into one system. Then the system's output that is represented by the personal human disease-related events is compared to

national health statistics. This work presents a correlation between information on diseases, that the system produced from Twitter data and influenza like illness (ILI) values that were reported by Centers for Disease Control and Prevention. The system that is developed in this work can predict values of disease rates before being reported by official resources. Moreover, Twitter is used as an example for a high-traffic social network for measuring sentiment analysis on smoking. Specifically, two state-of-the-art models are applied on several types of features (e.g., unigram and bigram) to identify sentiments of users for many different themes of smoking (e.g., cigarette, marijuana, and shisha), and the daily behavior of smokers is presented by analyzing their tweets. Overall, the work presented in this dissertation provides a real-time essential methodology for identifying social web messages that are medical cases, health behaviors, or outbreak events toward improved disease surveillance.

Keywords: Public Health Surveillance, Social Networks, Text Mining

Zusammenfassung

Social Network Sites wie Twitter, MySpace und Facebook sind webbasierte Dienste, die es Einzelpersonen ermöglichen, sofort mit großen Netzwerken von Freunden, Bekannten und Kollegen zu kommunizieren und Aktivitäten, Interessen oder Real-Life-Verbindungen zu teilen. Millionen von Nutzern teilen ihre Erfahrungen, Gedanken und Verhaltensweisen online über diese Seiten. Sie teilen auch ihre Symptome und Krankheiten in Echtzeit. Twitter ist eine beliebte soziale Website, die eine Quelle und einen Verteiler der öffentlichen Gesundheit darstellt und Informationen über kurze 140-Zeichen Microblogs, die informell als Tweets bekannt und durch eine Echtzeit-Natur gekennzeichnet sind, überträgt. Die Arbeit in dieser Dissertation betrachtet diese Mikroblogs als Informationsquelle. Darüber hinaus stellt sie Techniken zur Nutzung der Fülle von verfügbaren sozialen Web-Dokumenten vor, wodurch verschiedene Krankheitensarten auf der Grundlage von Namen verfolgt, identifiziert und charakterisiert werden können. Normale Krankheiten und Symptome verursacht durch Viren und Bakterien sowie das Verhalten der Menschen gegenüber gesundheitsbezogenen Themen werden präsentiert. Durch die automatische Verfolgung, Identifizierung und Analyse von krankheitsbezogenen Tweets kann diese Arbeit letztlich erhebliche Verbesserungen bei der Krankheitsüberwachung bieten und zukünftige Krankheitsgrade wie Influenza-Raten prognostizieren. Um die Natur von Krankheiten zu verstehen, die auf Twitter existieren, charakterisiert diese Dissertation zuerst einen großen Satz von Twitter-Nachrichten (Tweets), die mit der öffentlichen Gesundheit verbunden sind. Speziell analysiert diese Arbeit Tweets, um herauszufinden, was die Benutzer über ihre Gesundheit und Verhaltensweisen veröffentlichen, um die Art der gesundheitsbezogenen Daten auf Twitter zu verstehen und Techniken zu entwickeln, die Aufgaben erfüllen, um das Ziel zu erreichen und Herausforderungen zu meistern. Die riesige Menge an irrelevanten Daten in diesen Microblogs erfordert anspruchsvolle Filtermethoden, um relevante Buchungen zu identifizieren. Daher werden Annotationsrichtlinien entwickelt, um

einen Datensatz und einen Klassifikator zu erstellen, wodurch zwischen der relevanten und irrelevanten Krankheitsüberwachung unterschieden werden kann. Ein festgelegter Satz von Falldefinitionsrichtlinien innerhalb der medizinischen Intelligenzgemeinschaft wird für die Definition von krankheitsberichtenden Nachrichten ausgenutzt. Diese Arbeit enthält markante Informationen und erstellt Annotationsrichtlinien, um den Feature-Set zu bestimmen, der den besten Sparse-Text-Klassifikator erzeugt, um relevante Tweets zu identifizieren. Wegen des immensen Volumens von Twitter-Nachrichten wird ein Annotations-Tool entwickelt, das es dem Annotator ermöglicht, den Trainings-Datensatz leicht zu sammeln. Dann werden zwei Klassifizierungsalgorithmen auf dem Etikettierungsdatensatz mit mehreren Merkmalsätzen ausgewertet. Zur Verfolgung und Identifizierung von krankheitsbezogenen Nachrichten auf Twitter wird ein neuartiges Echtzeit-Filtersystem entwickelt, das Data-Mining-Techniken nutzt, um Crawling, Index, Extrahieren, Reinigen und Klassifizieren von Postings in Echtzeit-Überstunden zu entwickeln. Das Echtzeit-Filtersystem verfolgt Statusaktualisierungen der Population in Echtzeit und hat die Fähigkeit, Spams zu entfernen (z.B. shot-related tweets, vaccination-related tweets und Bieber-fever-related tweets). Darüber hinaus verwendet das System die hochmoderne Textklassifikation, um rein krankheitsbezogene Tweets zu filtern. Speziell zielt das System darauf ab, qualitativ hochwertige relevante Meldungen auszuwählen, die nützliche Informationen über Krankheiten und Ausbrüche wiederspiegeln. Die Laufzeitleistung des Filtersystems wird ausgewertet und die Genauigkeit des Klassifizierungsmodells getestet. Um die Bedeutung von Krankheits-related tweets zu verstehen, werden zunächst die Herausforderungen, mit denen die traditionellen named entity recognitions verwendet werden, vorgestellt. Dann wird die Semantik des Sammelns von Tweets entdeckt, indem ein neuartiges named entity recognition framework entwickelt wird, um die syntaktische Struktur von Postings zu bestimmen, medizinische Einheiten und Orte zu extrahieren und menschliche krankheitsbezogene Ereignisse zu erkennen. Zusätzlich wird

das Krankheitserkennungssystem mit ausgewählten Tweets für verschiedene Krankheiten ausgewertet. Um einen Beweis für die Techniken in dieser Arbeit zu liefern, die verwendet werden, um Krankheitsberichterstattungsnachrichten zu verfolgen, zu identifizieren und zu analysieren, werden alle diese Techniken zu einem System zusammengefasst. Dann wird die Produktion des Systems, die durch die persönlichen menschlichen krankheitsbezogenen Ereignisse repräsentiert wird, mit der nationalen Gesundheitsstatistik verglichen. Diese Arbeit stellt eine Korrelation zwischen Informationen über Krankheiten dar, die das System aus Twitter-Daten und Influenza like illness (ILI) Werten produziert, welche von den Centers for Disease Control and Prevention (CDC) wurden. Das System, das in dieser Arbeit entwickelt wird, kann Werte von Krankheitsraten vorhersagen, bevor sie von offiziellen Ressourcen gemeldet werden. Darüber hinaus wird Twitter als Beispiel für ein verkehrsfreies soziales Netzwerk zur Messung der Stimmungsanalyse beim Rauchen verwendet. Speziell werden zwei hochmoderne Modelle auf verschiedene Arten von Merkmalen (z.B. Unigram und Bigram) angewendet, um Gefühle von Benutzern für viele verschiedene Themen des Rauchens (z.B. Zigarette, Marihuana und Shisha) und die Tageszeiten zu identifizieren. Das Verhalten der Raucher wird durch die Analyse ihrer Tweets präsentiert. Insgesamt bietet die in dieser Dissertation präsentierte Arbeit eine Echtzeit-Methode zur Identifizierung von sozialen Web-Nachrichten, die medizinische Fälle, Gesundheitsverhalten oder Ausbruch-Ereignisse zur verbesserten Krankheitsüberwachung darstellen.

Schlagworte: Krankheitsüberwachung, Soziale Netzwerke, Text Mining

Contents

Contents vii

List of Tables xi

List of Figures xiii

1 Introduction 1

2 Characterization of Health Surveillance 7

2.1 Traditional Public Health Surveillance 9

2.2 Public Health Surveillance in Social Web 13

2.2.1 The Nature of Tweets 13

2.2.2 Tweets of Public Health 15

2.2.3 Social Health in Literature 16

2.2.3.1 Google Flu Trends 17

2.2.3.2 Flu Surveillance via Twitter 17

2.2.3.3 Dengue Surveillance via Twitter 18

2.2.4 Analysis of Public Health Data on Twitter 19

2.3 Diseases Surveillance System 27

3 Annotation and Classification 31

3.1 Micro-Messages Case-Driven Annotation 32

3.1.1 Positive Disease-Reporting Micro-Message 34

3.1.2 Negative Disease-Reporting Micro-Message 34

3.1.3 Annotation Tool 35

3.2 Data Collection and Processing 37

3.3 Experiments . . . 38
3.3.1 Annotation Study . . . 39
3.3.2 Methods . . . 40
3.3.3 Experimental Goals and Setting . . . 43
3.3.4 Results . . . 44
3.3.4.1 Guideline Exploitation and Annotated Data . . . 44
3.3.4.2 Medical Case-Driven Feature Analysis . . . 45
3.3.4.3 Micro-Message Classification . . . 46
3.4 Discussion . . . 46
3.5 Conclusions . . . 48

4 Streaming Scenario **51**
4.1 The Real-time Collecting and Filtering System . . . 52
4.1.1 Crawler . . . 52
4.1.2 Indexer . . . 53
4.1.3 Extractor and Scanner . . . 54
4.1.4 Collector . . . 56
4.1.5 Temporary Repository . . . 57
4.1.6 Cleansing Process . . . 57
4.1.7 Classifier . . . 58
4.1.8 Central Repository . . . 58
4.1.9 Stream Handler . . . 58
4.2 Filtering System Scenario . . . 59
4.3 Implementation . . . 60
4.4 Evaluation . . . 61
4.4.1 Classifier Performance . . . 61
4.4.2 Run Time Performance . . . 62
4.4.2.1 Latency . . . 62
4.4.2.2 Throughput . . . 63
4.5 Conclusions . . . 65

5 Entity Extraction and Event Recognition **66**
5.1 Motivation . . . 68

5.1.1 Syntactic Processing (Parsing) . . . 68
5.1.1.1 Simple Rule Language (SRL) . . . 68
5.1.2 Event Extraction . . . 70
5.1.3 Mining Tweets in Literature . . . 71
5.1.4 Traditional NER Tools on Tweets . . . 71
5.2 Event Extraction on Tweets . . . 73
5.2.1 Medical-Entity Extraction on Tweets . . . 74
5.2.1.1 Disease-Reporting Postings . . . 74
5.2.2 Location-Entity Extraction on Tweets . . . 76
5.2.2.1 Location of the User . . . 76
5.2.2.2 Location in the Tweet Text . . . 77
5.2.3 The Scenario of NER . . . 78
5.3 Experiments . . . 79
5.3.1 Data Selection . . . 80
5.3.2 Evaluation Metrics . . . 81
5.3.3 Methods . . . 81
5.3.3.1 Support Vector Machines with N-gram . . . 81
5.3.3.2 Parsing Rules . . . 81
5.3.3.3 Identifying Location . . . 83
5.3.4 Experimental Results . . . 84
5.3.4.1 Using SVMs Algorithm with N-gram Features . . . 84
5.3.4.2 Using SRL rules . . . 85
5.3.5 Summarization . . . 86
5.4 Discussion . . . 86
5.4.1 Is it Influenza or a Stomach Flu . . . 87
5.4.2 Event Summarization . . . 88
5.5 Conclusions . . . 89

6 Performance Evaluation of Whole System 91
6.1 Classifier Performance . . . 92
6.2 Run-time Performance . . . 95
6.3 Comparison to Gold Standard Data . . . 95
6.3.1 Gold Standard and Twitter Data . . . 96

6.3.2 Evaluation measures . . . 96
6.3.3 Results . . . 98
6.4 Forecasting of Disease Rates . . . 99
6.5 Alerting . . . 100
6.6 Conclusions . . . 104

7 Sentiment Analysis on Smoking 105
7.1 Tweets on Smoking . . . 106
7.2 Methods . . . 107
7.2.1 Data Collection . . . 107
7.2.2 Smoke-related Tweet Annotation . . . 108
7.2.2.1 Positive Smoke-related Tweet . . . 108
7.2.2.2 Negative Smoke-related Tweet . . . 109
7.2.2.3 Neutral Smoke-related Tweet . . . 109
7.2.3 Classifier Model . . . 109
7.3 Evaluation and Results . . . 109
7.4 Sentiment Analysis on Smoking . . . 111
7.5 Discussion . . . 111
7.5.1 Tracking Public Interest with Twitter Data . . . 112
7.5.2 Tracking smoke-related Tweets in spaces . . . 113
7.6 Conclusion . . . 114

8 Related work 115
8.1 Disease Identification in Textual News . . . 115
8.2 Disease Identification on the Social Web . . . 116
8.3 Named Entity Recognition on Twitter . . . 119
8.4 Location Entity Recognition on Twitter . . . 120
8.5 Behavior Analysis on Twitter . . . 120

9 Conclusions and Future Work 122

Bibliography 124

List of Tables

2.1 Content features for a public health trend. 26

3.1 Examples of useful words for labeling a tweet as positive 35
3.2 Examples of useful words for labeling a tweet as negative 36
3.3 Number of positive, negative, and unknown labeled tweets out of 200 39
3.4 Annotated data showing the type of medical annotation and the corresponding number of positive and negative labeled tweets. . . 49
3.5 Frequency of named entity types and word classes in positive and negative labeled texts and POS tags. 50
3.6 Performance results of the classification experiments (Accuracy (Acc), Recall (Re), precision (pre), F-Measure (F-M), Medical Condition (MC), Medical Treatments (MT), Nouns (N), and Verbs (V) listed in the annotation guidelines). 50

4.1 Some of the medical patterns detected by the real-time architecture. 55
4.2 Some new data that was added as training data. 57
4.3 The processing time for a single tweet. 63

5.1 Experimental Annotated Dataset 80
5.2 Evaluation rates for experiment dataset using SVMs. 84
5.3 Evaluation rates for experiment dataset using SRL. 86
5.4 Evaluation rates for Entity Categories. 86
5.5 Disease statistics for 700,000 tweets posted in April 2012. 89

6.1 Statistical Results of the Correlation 98

7.1 Experimental Annotated Dataset 107
7.2 Evaluation Results (Accuracy (Acc), Recall (Re), precision (pre), and F-Measure (F-M)) 110

List of Figures

2.1 Taxonomy of public health showing epidemiologic surveillance as one part of a broader set of surveillance activities. Adapted from Rolka and O'Connor [51] 8
2.2 Screenshot of HealthMap map from March 5, 2013. 11
2.3 Screenshot of BioCaster from March 5, 2013. 12
2.4 Screenshot of Twitter user interface from March 5, 2013. 14
2.5 Excerpt about Google Flu from a New York Times article on November 12, 2008 17
2.6 Public health-related Messages Observed in April-2012. 20
2.7 Volume of URL-related messages as percentage of the volume of observed public health messages(April 2012). 21
2.8 Volume of personal-related messages as percentage of the volume of observed public health messages (April 2012). 22
2.9 Volume of vaccination-related messages as percentage of the volume of observed public health messages (April 2012). 23
2.10 Volume of retweet-related messages as percentage of the volume of observed public health messages (April 2012). 24
2.11 Volume of spam-related messages as percentage of the volume of observed public health messages (April 2012). 25
2.12 Diseases trends as the percentage of the volume of observed public health messages (April 2012). 27
2.13 Symptoms trend as percentage of volume of observed public health messages (April 2012). 28
2.14 Overview of diseases surveillance system 29

3.1 Annotation tool GUI with search functionality and facilities like file and language selection. 37
3.2 Annotation view of the Annotation tool. 37
3.3 Separating hyperplane in SVM. 41

4.1 The real-time collecting and filtering system. 53
4.2 Filtering System. 59
4.3 Number of tweets per day. 63
4.4 Number of disease related and unrelated tweets per day. 64

5.1 Name Entity Recognition Framework for Disease-related Postings. 78

6.1 Model Quality Assessment 92
6.2 Number of Twitter messages per week for 17 weeks (week 1 starts on December 5th, 2011, week 17 ends on March 31, 2012) 97
6.3 Daily Flags from Alerting Methods. 103

7.1 Smoke-term-Related Tweet Volume 108
7.2 Diurnal smoke-related Messages 110
7.3 Smoke-term-Related Tweet Volume 112
7.4 Number of Unique Users (smokers) for July and August 2012. . . 114

Chapter 1

Introduction

Public health surveillance (PHS) is the ongoing, systematic collection, analysis, interpretation, and dissemination of health-related data that is needed for the planning, implementation, and evaluation of public health practice [19]. The data from PHS can be used to provide early warning information to those responsible for preventing diseases, epidemics (outbreaks), and other health hazard like smoking. PHS such as epidemic intelligence (EI) is being used by public health authorities to gather information related to disease activity, early warning, and infectious disease outbreak [38] [35]. There has been a growing interest in using internet-based health surveillance systems, through which official reports can be systematically gathered from formal and informal sources [45]. Surveillance systems, such as MediSys [38], the Global Public Health Intelligence Network [45], and BioCaster, [44] gather data from global media sources, such as news wires and web sites, to identify information about disease outbreaks. The improvement of these systems has been established by Google's Flu Trends research [62] that estimates the activity of influenza around the world in near real-time by aggregating user search data related to the flu. In fact, Google has the ability to show that certain search terms are highly correlated with reports from the Centers for Disease Control and Prevention Influenza-Like Illness (ILI)[19]. The problem in this dissertation is similar to the PHS [9] [38] that uses news documents and official reports(e.g., ProMED-mail[1]) to diagnose case confirmation, in order

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

to give an early warning of a possible outbreak and to monitor an outbreak's magnitude, geography, rate of change, and life cycle. However, this dissertation uses the social web documents instead of traditional sources that focus on identifying diseases and other outbreaks over news documents. Indeed, the news articles adhere to correct grammar and syntax and provide complete description of events. Thus, most disease outbreak detection techniques, including natural language processing (NLP) tools such as named-entity recognition and extraction, and part-of-speech tagging, are working well with news documents [63] [65]. Conversely, social web documents such as the micro-messages of Twitter usually have short and noisy properties including slang and often unclear language. Dealing with such data has posed a tremendous challenge to current surveillance systems [67]. Therefore, this dissertation is interested in disease surveillance by using social web, which represents a data source for internet-based surveillance, since online social network sites (SNS) have become a global phenomenon. Some comprise large communities and are increasingly drawing a larger population into the online world. Twitter is a social network site that allows hundreds of millions of users to communicate with each other in real-time. Huge amounts of Twitter messages, informally known as "tweets", are sent every minute. People want to inform friends, colleagues, and others about personal feelings, latest thoughts, and things they have done. The relevance of Twitter for monitoring and surveillance purposes increases. For instance, Twitter was used to monitor the U.S. presidential debate in 2008 [15], as well as the impact of earthquake effects [52]. Consequently, tweets could be useful to learn about current public health status updates, because people tend to tweet when they are feeling ill, recognize specific symptoms, or have been diagnosed with a disease. Previous works [14], [11], and [5] have studied the possibility of early disease outbreak detection using Twitter. While these previous studies mainly focused on specific diseases (e.g., flu), this dissertation goes beyond single disease detection and monitors many medical conditions via detection of names of diseases, viruses, bacteria, and symptoms, as well as disease outbreaks. This dissertation specially uses Twitter as the source for public health data to monitor medical conditions, as Twitter is faster and cheaper than public health providers because the data are coming directly from the population instead of hospitals and research centers. In addition, Twitter

is publicly available for all people and provides massive data that are coming in real-time. Therefore, in addition to traditional techniques, this dissertation presents more advanced techniques to create a robust-real-time system for detection, tracking, and analysis of disease outbreak on Twitter. Because of the immense volume of tweets sent every day, it is extremely important to identify the messages relevant to disease-monitoring purposes and to understand what they mean. Comprehensive analysis techniques required for the automated analysis of texts (e.g., named entity recognition, linguistic parsing, and information extraction) need time to process. Thus, to reduce the processing time and noise introduced by irrelevant postings the dataset size needs to be reduced in advance, so that only relevant texts are processed. Existing approaches for text filtering in the medical domain that use keyword lists or plentiful text to identify relevant articles do not take into account the peculiarities of tweets. A challenge faced in using sparse and noisy microblog postings is that the available textual features are limited. Tweets contain a limited number of characters, but when the noisy characters are eliminated even fewer terms that have good discriminating power for medical intelligence are available. This challenge makes it difficult to build an adequate classifier. The choice of classifier features used for this type of text is therefore more crucial than for corpora that contain richer textual features. Moreover, new terms, abbreviations, and slang make it difficult to create complete keyword lists. To address this issue, an established set of case definition guidelines within the medical intelligence community are exploited for defining disease-reporting messages. The dissertation incorporates salient information and creates annotation guidelines to determine the feature set that produces the best sparse-text classifier to identify relevant tweets and evaluate the classifier on an annotated dataset by using 10-fold cross-validation. Recently, the big challenge of detection and tracking of diseases on Twitter is how to deal with the huge amount of incoming messages in real-time? and how to get more discriminative data? Collecting and filtering methods are important to detect public health data such as disease outbreaks or infections on Twitter in real-time. Thus, this dissertation provides techniques to track status updates of Twitter users by developing a filtering system to detect positive diseases-related postings on Twitter, as an example for a high-traffic social network. The real-time filtering system

uses Twitter application programming interface (API) to crawl tweets as they are posted. Data mining techniques are used to index, extract, and classify postings. Besides, this system provides a scalable and incremental solution to handle the high volume and remove the chatter of tweets. In addition, the system distinguishes between real-positive disease-related, shot-related, vaccination-related, and Bieber-fever-related postings which are postings about the pop star Justin Bieber. Then the run-time performance of the whole system will be evaluated with respect to latency and throughput. Identifying public health events on Twitter requires techniques to make the collected data understandable and valuable, which can be done by extracting entities and events. Information Extraction (IE) is one task of NLP that extracts useful structured information from unstructured text. The current IE techniques (e.g., OpenCalais[1] and Part-Of-Speech tagger[2]) are focusing on popular and formal text such as news articles, blogs, and clinical notes. However, their performance on text from Facebook status updates or Twitter tweets is poor [67], as the social data is noisy and contains lots of chatter. Furthermore, the text of social media is very short compared to formal domain text. For these reasons, building a named-entity recognition system for our own interest is important to extract entities and events of interest. Conversely, extraction of location entity is a major challenge because Twitter allows the users to specify their geographic location as meta-data. The location data is manually entered by the user or updated using a global positioning system (GPS) enabled device. The user can enter his/her location in the free-text form field. Therefore, the location information may be incorrect. For instance, the user may enter a fake location (e.g., heaven, in somewhere, or behind you) or may not enter any information. This work does not study the sparsity of the location, but it takes the location field information and get all geography related information (e.g., GPS data). To address the challenges, this dissertation presents an IE framework for this interest, that can analyze the semantic of the social post and extract medical and location entities from disease-related postings. Then different techniques of machine learning and NLP are evaluated on a dataset of tweets.

Overall, the previously described tasks are grouped in a real-time text mining

[1] http://www.opencalais.com/
[2] http://nlp.stanford.edu/software/tagger.shtml

system that uses a machine learning classifier to distinguish relevant postings from the irrelevant ones. It performs well on noisy and sparse data such as Twitter, whereas it extracts entities and events from relevant postings. The work in this dissertation focuses on the identification of disease-reporting tweets. Briefly, a disease-reporting tweet is one in which a person reports having symptoms or claims to have a certain disease. In this work, the terms such as tweet, micro-message, microblog, Twitter message, post, and posting are used as synonyms. The significance in this work is the use of the state-of-the-art text classification to filter postings into disease related or unrelated, track peoples' status updates in real-time, extract entities and events to make them understandable and valuable, provide a scalable and incremental solution to handle the massive volume, and identify events of disease-reporting tweets.

In summary, the contributions of this dissertation are as follows:

1. A qualitative study of PHS in social web (Chapter 2).
2. Building an annotated Twitter dataset to build a classifier that automatically distinguishes tweets that are relevant to disease surveillance from those that are irrelevant (Chapter 3).
3. Developing a filtering system that uses data mining techniques to crawl, index, extract, and classify postings in real-time overtime (Chapter 4).
4. Discovering the semantic of collecting data by developing a named entity recognition framework to determine the syntactic structure of postings, extract medical entities and locations, and recognize human disease-related events (Chapter 5).
5. Evaluating the whole system by evaluating each of its components: the performance of the microblog classifier, run-time performance of the real-time filtering system, and performance of the named entity recognition system. Importantly, the comparison of output of the system to national health statistics (Chapter 6).

Chapter 7 discusses several techniques for sentiment analysis on Twitter data related to smoking. Chapter 8 describes additional related work, and then the

conclusions and discuss directions for future work are presented in Chapter 9.

As the amount of social web data grows, more research will have to be conducted in order to identify robust ways to organize and filter the data. The work in this dissertation aims to provide scalable techniques for organizing social web documents associated with public health problems (e.g., symptoms, diseases, and outbreaks). With disease-posting collection, characterization, filtration, and events extraction techniques, this dissertation provides new opportunities for forecasting disease rates in real-time before being reported from official sources.

Chapter 2

Characterization of Public Health Surveillance

The public health surveillance is a part of biosurveillance that comprises human, animal, and agricultural surveillance. Biosurveillance is the collection and analysis of data related to biological agents, diseases, risk factors, and other health events. The aim is to improve the likelihood that a disease outbreak, whether man made or natural, is detected as early as possible so that the medical and public health communities can respond quickly. The term biosurveillance is defined by Homeland Security Presidential Directive (HSPD) [23] as "the process of active data gathering with appropriate analysis and interpretation of biosphere data that might relate to disease activity and threats to human or animal health, which include infectious, toxic, or metabolic threats, and regardless of intentional or natural origin, to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity".

One particular type of biosurveillance that is interested in this thesis is epidemiologic surveillance, which HSPD defines as "the process of actively gathering and analyzing data related to human health and disease in a population in order to obtain early warning of human health events, rapid characterization of human disease events, and overall situational awareness of disease activity in the human population."

Rolka and Connor [51] adopted a taxonomy (Figure 2.1) of public health

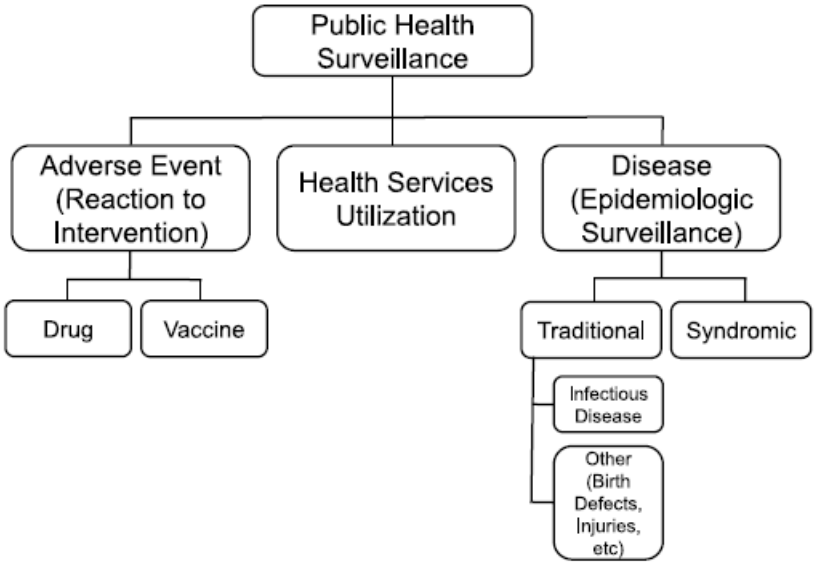

Figure 2.1: Taxonomy of public health showing epidemiologic surveillance as one part of a broader set of surveillance activities. Adapted from Rolka and O'Connor [51]

.

surveillance that encompasses the events related to drugs and vaccine, and how services of health are used. Moreover, the taxonomy includes the Palindromic surveillance that is a specific type of epidemiological surveillance and has been defined as "the ongoing, systematic collection, analysis, interpretation, and application of real-time (or near-real-time) indicators of diseases and outbreaks that allow their detection before public health authorities would otherwise note them" [57].

This chapter offers the following contributions:

- Describing the traditional public health surveillance that uses official sources to monitor public health issues; moreover, presenting a brief description of some existing biosurveillance systems (Section 2.1).
- Presenting in depth nature of public health data on social web. In addition, presenting several analyses on Twitter data to understand the type of public health data (Section 2.2).
- Presenting the overview of the public health surveillance system that is

suitable for large-scale social web data and uses significant techniques for tracking disease outbreak using Twitter (Section 2.3).

2.1 Traditional Public Health Surveillance

With growing awareness of the threat of emerging infections and bioterrorism, many research efforts are focusing on managing public health problems specially to detect disease outbreaks quickly; subsequently, many new types of disease surveillance systems have been developed. Some of those systems are more traditional, use medical data sources that are collected for other purposes (e.g., emergency room logs), and some use nonclinical data (e.g., pharmacy sales and school absenteeism) [1]. The improvement of those systems has been established for relying more and more on massive quantities of data from non-traditional sources such as internet news, electronic clinical reports, and ProMED-Mail reports that use the internet to broadcast information on outbreaks by e-mailing and posting case reports, including many gleaned from readers, along with expert commentary. The idea behind those systems is to provide early detection of disease outbreaks by using official sources from clinical reports to news streams. As such these systems may be seen as traditional public health surveillance systems. Conversely, the social web represents an alternative new source for health surveillance because the information is coming directly from the population in real-time. Some biosurveillance systems are described in the following:

BioSense The Centers for Disease Control and Prevention (CDC) developed a BioSense[1] program that tracks health problems in the United States. It provides a picture of the current situation of any health condition, anywhere and everywhere in the country. BioSense pulls together information on emergency department visits and hospitalizations from multiple sources such as civilian hospitals around the USA.

[1] http://www.cdc.gov/biosense/

EARS (Early Aberration Reporting System) EARS[1] was developed by the CDC for monitoring bioterrorism during large-scale events. For example, the EARS system was used in the aftermath of Hurricane Katrina to monitor communicable diseases in Louisiana [61]. Moreover, public health officials of various cities, counties, and states in the United States and abroad use EARS on syndromic data from emergency departments, data collected at shelters after disasters, reportable conditions, physician's office data, school and business absenteeism, and over-the-counter drug sales. EARS is convenient, easy to use, and available at no cost.

MedISys The medical information system (MedISys[2]) is a real-time alert system for medical- and health-related topics. MedISys gathers and analyzes information everyday from medical and news sites by using keyword patterns and pattern combinations.

HealthMap HealthMap[3] is a real-time surveillance system for monitoring disease outbreaks and emerging public health threats. HealthMap gathers information from different sources such as Google News, mailing lists, government web sites, and discussion forums. Then, it filters, and analyzes the information and plots it on the map, as shown in Figure 2.2.

BioCaster BioCaster[4] is an ontology-based text mining system developed by a multi-disciplinary team of experts [9] for detecting and tracking public health rumors from web news and has been running since 2006. The system continuously analyzes documents reported from over 1700 Rich Site Summary(RSS) feeds, classifies them for topical relevance, and plots them onto a Google map using geocoded information as shown in Figure 2.3. Currently, BioCaster is using new resources such as ProMED-mail, Google News, and the World Health Organization (WHO).

[1] http://www.bt.cdc.gov/surveillance/ears/
[2] http://medusa.jrc.it/medisys/homeedition/en/home.html
[3] http://healthmap.org/en/
[4] http://born.nii.ac.jp/

Figure 2.2: Screenshot of HealthMap map from March 5, 2013.

Figure 2.3: Screenshot of BioCaster from March 5, 2013.

The limitation of this description is that most of the traditional approaches are gathering information from official resources (e.g., news and ProMED-Mail) and do not consider social web data (e.g., tweets) as a source of public health surveillance.

2.2 Public Health Surveillance in Social Web

Social web sites, such as Twitter, allow people to share their personal feelings, in particular, information about their health. Twitter data has already been used as information for event detection from the Iran election or the reaction to the 2010 earthquake in Haiti [25]. Moreover, the status updates of Twitter have been used as evidence for the possibility of tracking epidemics and especially tracking the prevalence of ILI [37]. This dissertation uses public shared content of Twitter data as an information source for epidemiological surveillance by tracking human diseases, viruses, bacterium, outbreaks, and symptoms. This section aims to characterize the nature of tweets that helps to identify different public health information. Moreover, a better understanding of semantics and the type of public health tweets provides useful information for techniques that will be built based on those data. This section begins with introduction to the nature of tweets and then describes tweets related to public health. Subsequently, the public health tweets which are returned via search queries on Twitter content are analyzed in depth, and the overview of the whole system is described as well.

2.2.1 The Nature of Tweets

Twitter is a popular social networking site, that has hundreds of millions of registered users as of December 2012. In fact, the main function of Twitter is to only ask the question: What are you doing right now? The answer is limited to 140 characters. The current Twitter user interface is shown in Figure 2.4. Status updates on Twitter known as *tweet* or *microblog* can be sent via a web browser, SMS, e-mail, or third party application and are displayed on the users' profile.

Twitter messages can be considered first-hand information, because Twitter's technology breaks down communication barriers and allows actions in the real

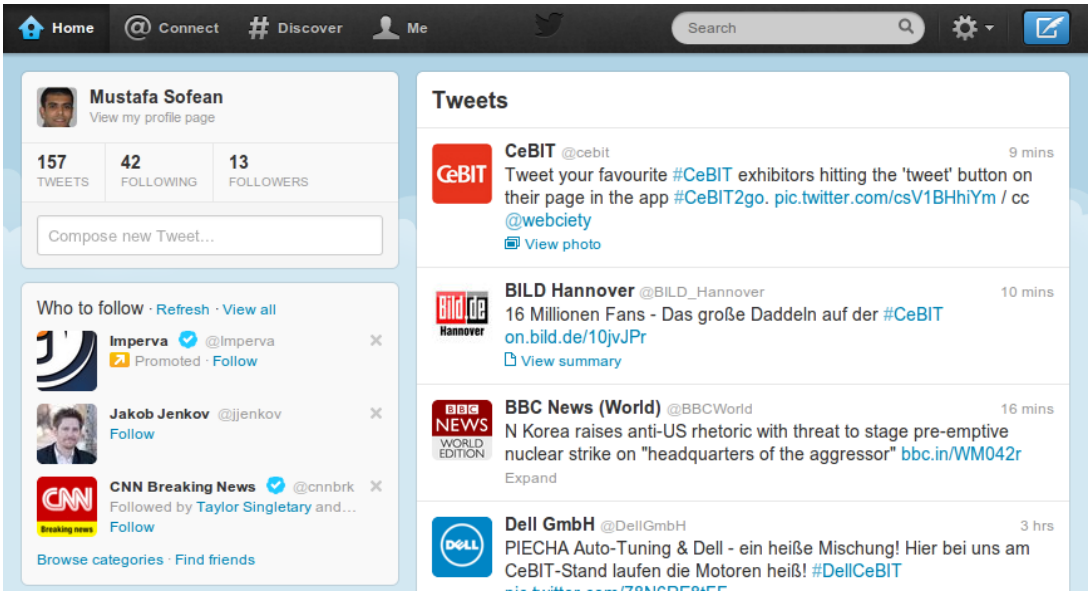

Figure 2.4: Screenshot of Twitter user interface from March 5, 2013.

world to be received in real-time, even before they are reported through official channels. It might thus provide additional information for disease monitoring and surveillance.

Moreover, Twitter allows a user to follow other users, because Twitter users can set their privacy settings so that their status updates are available only to the user's followers. However, by the default privacy settings the status updates are public. However, User A can follow User B without requiring approval from User B.

Twitter messages are text messages comprising up to 140 characters. It is common to assign hash tags to the tweets that denote the topic (e.g., #cancer). Conversely, retweet is a Twitter message that repeats some information previously tweeted by another user. A previous post can be retweeted using the "RT@username" text as a prefix to the original (or previous) post (e.g., *RT @rawstory: Second cholera outbreak affects 51 in Cuba: http://t.co/Bt9Uos6C*). Uniform Resource Locator (URL) in tweets are cited as references. Often, the URL is shortened using a URL shortening service such as TinyURL.com. Additionally, a mention is a message that includes other username in the text of the tweet (e.g., *have fever too @Chrisgbwe3*). Finally, a reply message is a tweet

from one user that is a response to another user's message and is identified by the fact that it starts with the replied-to username (e.g., *@headache I got the worst headache ever*).

2.2.2 Tweets of Public Health

In fact, Twitter data are being used in an increasing number of research studies to track events (e.g., earthquakes). On Twitter, people talk about matters in their life that are specially feelings about their symptoms and diseases. In addition, the official users of Twitter (e.g., governments, news, and health organizations) share information about disease outbreaks when they occure. Personal users can also share that information by tweeting or retweeting. Therefore, Twitter may appear to be an important tool for tracking disease outbreaks. Some example tweets are as follows:

1. *I'm very ill at the moment ... I have Measles,*

2. *H1N1 outbreak kills 2 in Mexico bit.ly/wWr3V3*

3. *Just got the HPV vaccine, better tighten up your b-holes everybody,*

4. *My girlfriend has got a massive case of diarrhea.*

Having these tweets as real-time information available to the public through Twitter can provide indications of potential disease epidemics faster than traditional epidemic surveillance activities. However, the work described in This dissertation considers Twitter with its real-time public health information firstly, as a source of information that is provided directly from a population, as in example 1. Secondly, as a *Distributor* that implies that Twitter distributes or links information from other media as in example 2. Both examples 1 and 2 show that disease-reporting messages can be found that are worth considering for monitoring purposes. The third example refers to a vaccination rather than a disease or

symptom. Examples 1, 2, and 4 are totally relevant to disease monitoring even though a symptom is mentioned.

In an examination of the nature of tweets with respect to the task of medical gathering, the observation is that the number of relevant and irrelevant tweets differs on the basis of the keyword. For example, 76% of tweets containing the keyword *headache* have been selected by the annotators as relevant or positive. Conversely, only 41% of tweets with the keyword *fever* have been marked relevant. These percentages might change every day depending on the relevance of keywords or depending on how many persons have certain symptoms or diseases and report about them.

Relevant tweets can be distinguished on the basis of their content regarding information about the health status of (1) the author, (2) a friend of the author, or (3) a prominent person. Furthermore, tweets also report about general health information or health education and official information or advice that deals with the prevention and management of health problems for international travelers [58].

A particular phenomenon of tweets is lingo. These are abbreviations of words that are normally known by all twitter users. Because there are only 140 characters in one tweet, it is necessary to reduce the word count to express as much as possible. Some examples for lingo are: *B4* for "before", *agn* for "again", *u* for "you", and *2* for "to". From a preliminary analysis, the conclusion is that this lingo is not used for content-bearing terms such as medical conditions; however the tweet may contain popular medical terms instead of the correct medical terms, for instance users write *stomach flu* instead of *gastroenteritis*. However, the techniques in chapter 5 have the ability to distinguish the popular medical terms.

2.2.3 Social Health in Literature

This section presents and discusses various efforts to study public health by using data of the social web. This work considers search engine queries as social data because they are provided by public users (Section 2.2.3.1), Influenza surveillance via Twitter (Section 2.2.3.2), and Dengue surveillance via Twitter (Section

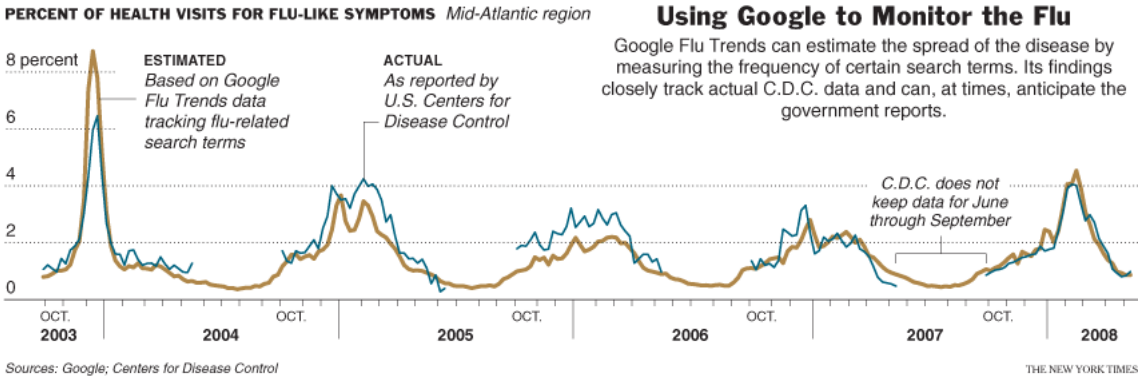

Figure 2.5: Excerpt about Google Flu from a New York Times article on November 12, 2008 .

2.2.3.3).

2.2.3.1 Google Flu Trends

Google[1] developed Google Flu Trends[2], which is designed to track "health seeking" behavior and estimate influenza-like illness (ILI) rates from internet searches. The idea is that sick people first attempt to self-treat before seeking medical attention, and often, the first step is a search engine query for information. During the 2007 and 2008 influenza season, Google Flu Trends estimates were highly correlated to CDC surveillance for ILI with a mean correlation coefficient over nine US Census Regions of 0.97 [22]. Figure 2.5 shows a strong temporal association among rates from each surveillance system from June 29, 2003 through May 31, 2008.

2.2.3.2 Flu Surveillance via Twitter

Seasonal influenza epidemics result in about three to five million cases of severe illness and about 250,000 to 500,000 deaths worldwide each year [32]. As people tweet on Twitter to complain about being sick with the flu and its symptoms, those tweets are used by experts and researchers to gather data about influenza. The key idea in most studies of influenza tracking on Twitter was to choose

[1]http://www.google.com
[2]http://www.google.org/flutrends/

keywords to filter and aggregate influenza-related messages, as well as the studies were able to forecast future influenza rates. Two similar studies were published to estimate national influenza rates from tweets. Both use linear regression to detect keywords that correlate with influenza rates and then combine these keywords to estimate national influenza rates.

First, Culotta [11] analyzed more than 500 million Twitter messages that were collected using Twitter's application programming interface (API) over the eight-month period from August 2009 to May 2010. Culotta used a small number of keywords (flu, cough, headache, and sore throat) to track rates of influenza-related messages on Twitter and tried to align frequency of tweets with data on influenza cases published by the Center of Disease Control and Preventative (CDC). Culotta actually aggregated keyword frequencies using separate predictor variables (e.g., multiple linear regression) and then outperformed aggregating keyword frequencies into a single predictor variable (e.g., simple linear regression). Subsequently, he used a simple bag-of-words classifier trained on 200 documents to filter messages and the final model achieved a strong correlation with CDC statistics over 5 weeks of validation data.

Secondly, Lampos and Cristianini [37] chose 73 keywords from 1,560 flu-related terms for ILI tracking in the UK. They trained and evaluated a large number dataset (28 million messages) and compared the dataset to the data from the Health Protection Agency lab.

2.2.3.3 Dengue Surveillance via Twitter

The potential of using search terms or Twitter data for the sake of dengue surveillance has been proved. Google Dengue Trends[1] uses aggregated Google search data to estimate current dengue activity around the world in near real-time. Because Google's model is built on the fraction of Google search volume for dengue-related queries, its dengue trends were able to adequately estimate true dengue activity according to official dengue case counts reported by national ministries of

[1] http://www.google.org/denguetrends/

health or the WHO for five selected countries (Bolivia, Brazil, India, Indonesia, and Singapore) and for majority of the seasons during the analyzed time frame. Moreover, Twitter has been used to track dengue fever in Brazil by software created by a collaboration between two Brazilian National Institutes of Science and Technology [29]. The team has analyzed how the Dengue outbreak is reflected on Twitter. They proposed a methodology on the basis of four dimensions: the amount of tweets, location of tweet, publishing time of tweet, and content of tweet that is the overall population perception/sentiment about the dengue epidemic. They collected two datasets Jan, 2009 to May, 2009 and December, 2010 to Apr, 2011; and collected all tweets that contained the word Dengue. They only considered the tweets that were published from Brazil and ignored tweets that contained invalid location. Their methodology that was tested on 2447 tweets showed that the personal experience tweets tightly correlated with the dengue outbreaks identified by the Brazilian Ministry of Health.

2.2.4 Analysis of Public Health Data on Twitter

In order to analyze the public health surveillance on Twitter, the work in this section tries to find out what Twitter users posted about their health and whether those posts could be a useful source of public health information. Understanding the nature of Twitter data will help to develop techniques that are required to meet the purpose. This section presents in depth the analyzing of the public health trend on Twitter and the analysis performed on the public health dataset collected from Twitter via Twitter API during the month of April 2012 (starting from the second day of April at 8 o'clock to the end). The data collected via monitoring tweets contains medical keywords such as disease, virus, bacteria, and symptom names. Figure 2.8 shows daily messages about the public health that could be relevant or irrelevant to disease outbreak.

The analysis is conducted to understand the characteristics of public health on Twitter and identify various types of trends (e.g., diseases, viruses, or spams) that take place in the dataset. In this analysis, $M_{(total)}$ refers to all messages in the dataset. For each trend, the associated number of tweets M_t is computed by matching each tweet with the features related to that trend, for example, the

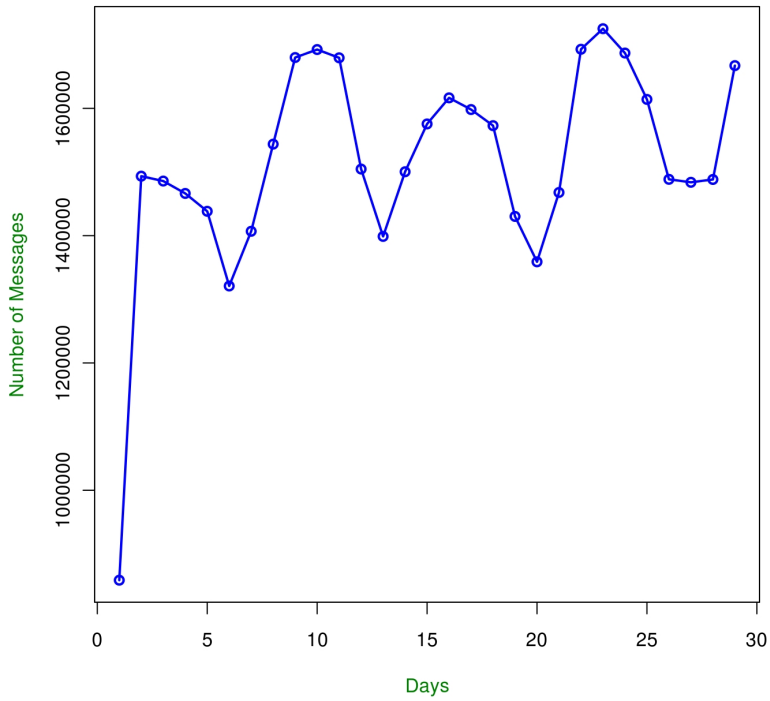

Figure 2.6: Public health-related Messages Observed in April-2012.

disease trend features are all disease names that appear in the dataset. Then the average number of messages with trend t is $M_{(t)}/M_{(total)}$. Through in depth analysis, the number of messages related to a single feature $M_{(f)}$ are calculated and are associated with a trend $M_{(t)}$ or $M_{(total)}$ (e.g., single disease, retweets, and hyperlinks). Because the volume of messages on Twitter varies overtime, usage statistics are expressed in terms of the fraction of the total tweets posted within the corresponding time interval (see Table 2.1). By using search terms, 43,936,873 messages (tweets) that are posted in April 2012 are analyzed to get the percentage of messages each day. As aforementioned before, Twitter distributes information from another media, because the message that contains a URL in its content is from other sources on the web. It was found that 13% of messages in the dataset contain URLs. Figure 2.7 shows the daily percentages of URL-related messages. Twitter is also considered as a source of public-health information. Therefore,

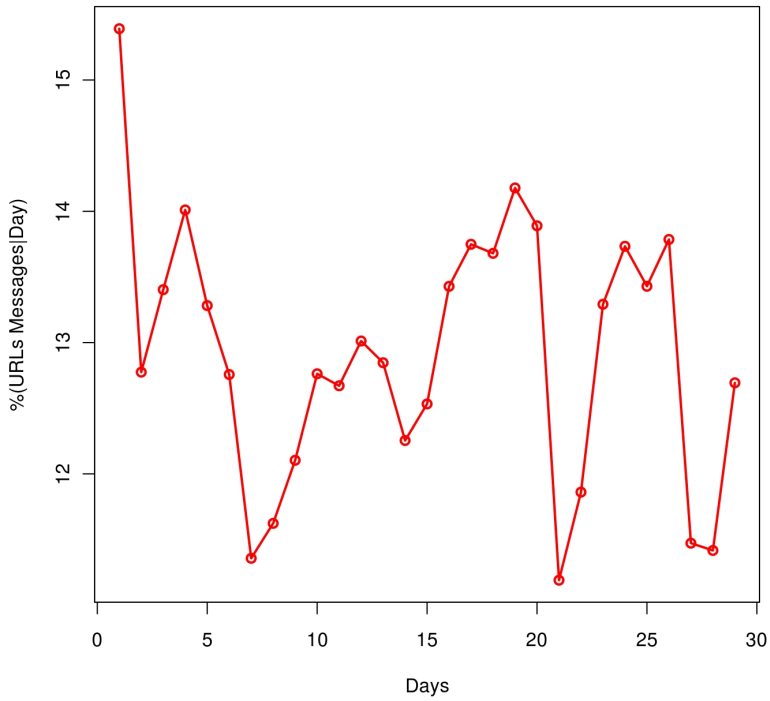

Figure 2.7: Volume of URL-related messages as percentage of the volume of observed public health messages(April 2012).

the personal messages are extracted by using search terms and phrases (e.g., I got, I have, you got, and down with). 59% of messages are personal messages. The daily percentages of personal-related messages are shown in Figure 2.8.

Figure 2.9 shows the percentage of vaccination-related messages mentioning vaccination terms that were selected according to the terms that are observed in the messages(e.g., vaccine, drug, shot, or immunization). Similarly, Figure 2.10 shows the percentage of retweeted messages that is 32% of dataset. In the current analysis of public health data on Twitter, an idea comes up with the term "spam". This was referred to any observed message that contains public health terms; however, they are actually not related to health. Examples of spam message are Bieber-fever-related messages, which are messages about the pop

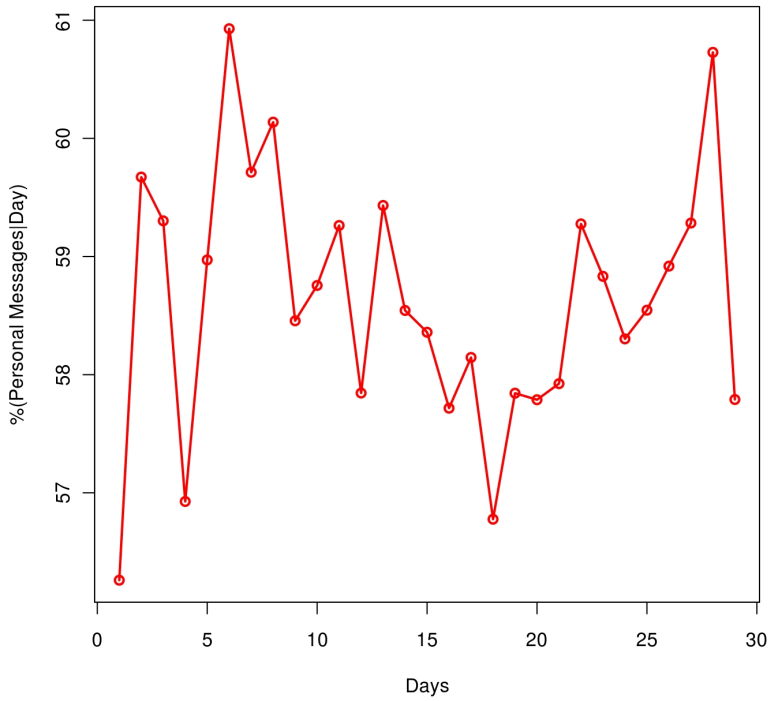

Figure 2.8: Volume of personal-related messages as percentage of the volume of observed public health messages (April 2012).

star Justin Bieber[1]; messages related to Typhoid Mary[2]; or messages related to beadles fever or beadles measles. Nevertheless, Six percent of the dataset is spam and the daily volume of spam is presented in Figure 2.11.

Moreover, the disease and symptom trends are identified by calculating the number of messages containing disease names (common names without viruses or bacteria) and messages containing symptom names including popular names(e.g., ache, and stomach flu). The results were that 14% of the dataset is disease trend, and 66% is symptom trend, as well as the daily percentages are shown in Figures 2.12 and 2.13, respectively.

The results of the quantitative analysis provide a strong indication that the characteristics of the messages associated with health can be used to monitor

[1]http://en.wikipedia.org/wiki/Justin_Bieber
[2]http://en.wikipedia.org/wiki/Typhoid_Mary/

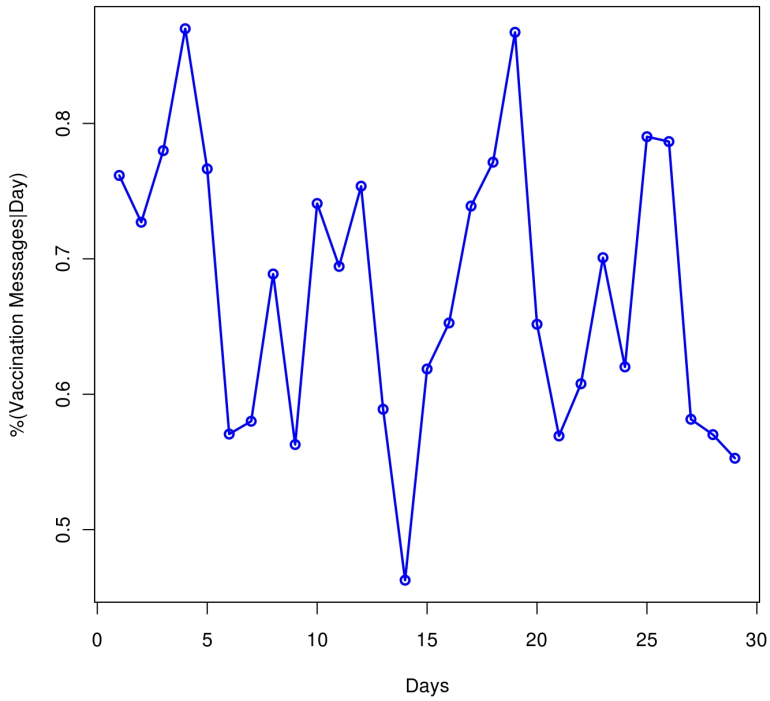

Figure 2.9: Volume of vaccination-related messages as percentage of the volume of observed public health messages (April 2012).

disease outbreaks. One of the major findings of this analysis is that the population on Twitter share information related to their personal feeling about diseases and symptoms in real-time, as has been shown previously; this information accounts to 59% of the public-health data. Twitter users share their feelings to inform their friends or get advice on handling health problems. Therefore, Twitter will be considered as a source of public-health data that is posted directly from the population. Moreover, the daily volume of public health data, especially messages referring to the terms, spike during the days when many unusual events take place; therefore, the volume of messages changes daily. According to Twitter API documentation[1], there will always be limits if filters for the Twitter streaming API are been used. The volume of returned tweets that is required to stream depends

[1]https://dev.twitter.com/docs

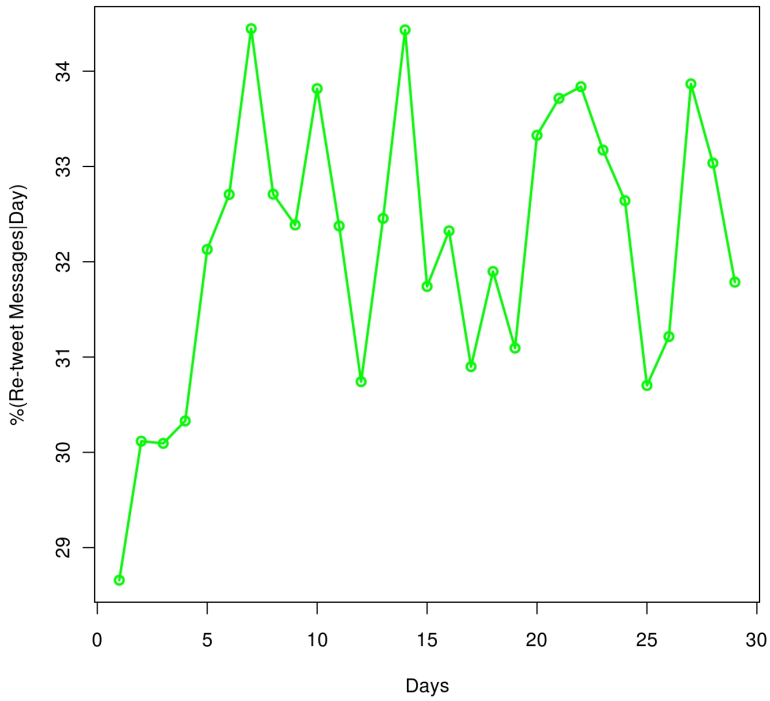

Figure 2.10: Volume of retweet-related messages as percentage of the volume of observed public health messages (April 2012).

on the Twitter API because if tweets for the term "flu" for one day are filtered, all the tweets about "flu" accounted to less than 1% of the total number of tweets on Twitter at the moment of streaming; thus all the tweets mentioning "flu" will be received. However, if a large quantity of people were tweeting about "flu" on that day and it rose above 1% of the total fire-hose volume[1], then there could be a portion of the tweets and rate limit message packets notifying how many tweets are missed. If the track terms "flu", "fever", "cancer", and "headache" were combined, the chance that the total volume of results get back would exceed 1% of the total fire-hose volume. Also the rate limit messages telling how many tweets were missed would be received, but not necessarily which terms they had matched with. To get the complete volume of Twitter data, there are some commercial

[1]The volume of all Twitter stream.

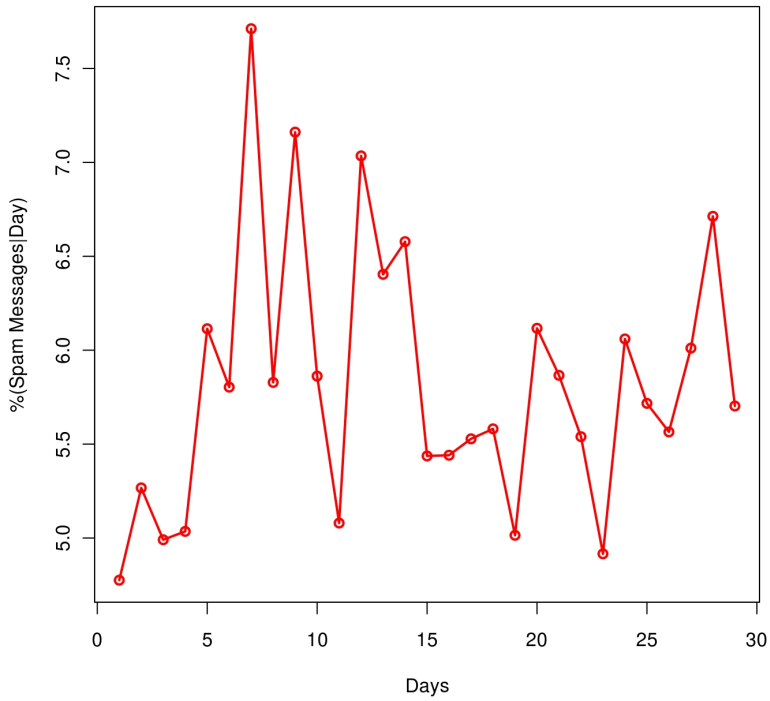

Figure 2.11: Volume of spam-related messages as percentage of the volume of observed public health messages (April 2012).

products (e.g., Gnip or Datasift) offering it; however, they are expensive.

From the analysis of public health information on Twitter, the nature of the data that might be posted directly from users, shared from another media, replied from other users (retweets) were understood. This data is totally different from popular and formal texts (e.g., news articles, blogs, and clinical reports) because of the volume of tweets, limited length, noises, and spams. Therefore, significant and sophisticated techniques are required for managing and analyzing that data.

In fact, this work focuses on tracking status updates of people and identifying if they are real disease reportings that include all disease messages, where people report about symptoms, claim that they have a certain disease, or indicate that there is an outbreak somewhere. Besides, not all Twitter data is related to the interest as was observed previously, huge amount of data are spam. Consequently,

Feature	**Description**
Rate of messages with Trend	Let $M_{(t,d)}$ be the number of messages in a day d containing a trend t and a subset of all messages in the day d. Then the rate of trend t is $M_{(t,d)}/M_{(total,d)}$.
Rate of messages with Feature	Let $M_{(f,d)} \subseteq M_{(total,d)}$ be the set of day messages in $M_{(total,d)}$ containing the feature f. Then, the rate of messages with feature f is $M_{(f,d)}/M_{(total,d)}$.
Rate of messages with URLs	Let $M_{(U,d)} \subseteq M_{(total,d)}$ be the set of day messages in $M_{(total,d)}$ with URLs. Then, the rate of messages with URLs is $M_{(U,d)}/M_{(total,d)}$.
Rate of messages with Vaccine	Let $M_{(V,d)} \subseteq M_{(total,d)}$. be the set of day messages in $M_{(total,d)}$ that are related to vaccine. Then the rate of messages containing vaccine is $M_{(V,d)}/M_{(total,d)}$.
Rate of messages with Spam	Let $M_{(S,d)} \subseteq M_{(total,d)}$ be the set of day messages in $M_{(total,d)}$ that contain medical condition but they are not related to public health (e.g., Bieber -Fever related-messages). Then the rate of message containing spam is $M_{(S,d)}/M_{(total,d)}$.
Rate of messages with retweets	Let $M_{(Re,d)} \subseteq M_{(total,d)}$ be the number day messages in $M_{(total,d)}$ that are "retweets" (e.g., messages started with RT@Username). Then, the rate of messages containing retweets is $M_{(Re,d)}/M_{(total,d)}$.
Rate of Personal messages	Let $M_{(P,d)} \subseteq M_{(total,d)}$ be the number Personal messages in $M_{(total,d)}$ that contain a personal feeling toward public health. Then the rate of personal messages is $M_{(P,d)}/M_{(total,d)}$.

Table 2.1: Content features for a public health trend.

in this dissertation, a significant system to meet all challenges for tracking human disease outbreaks via Twitter data is developed.

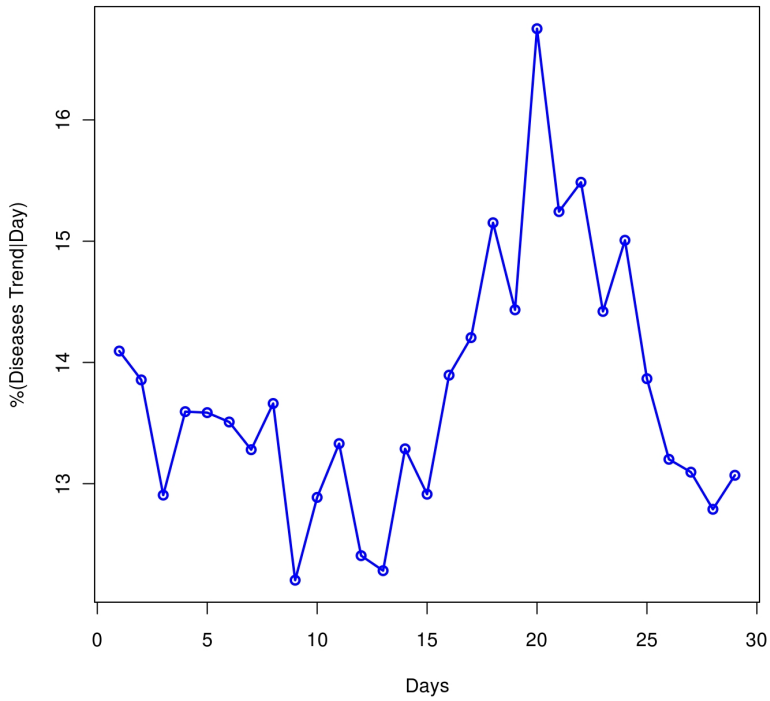

Figure 2.12: Diseases trends as the percentage of the volume of observed public health messages (April 2012).

2.3 Diseases Surveillance System

From the characteristics of the public health data (Section 2.2.2 and Section 2.2.4), Twitter data has limited length (140 characters) and may contain slang, noisy, and spams. However, Twitter data could be faster and cheaper to track diseases and outbreaks, because of the real-time nature and public availability. Therefore, developing a system to detect diseases and outbreaks on Twitter requires significant techniques to deal with the nature of Twitter data. Figure 2.14 shows the overview of all stages of the short-text mining system that is developed in this dissertation to track, identify, and analyze tweets of public health. The first stage is represented by filtering medical cases from tweets in real-time over a period of time. In fact, Twitter data is massive and contains a lot of chatter. Thus, preprocessing techniques take place to index, remove the noise or

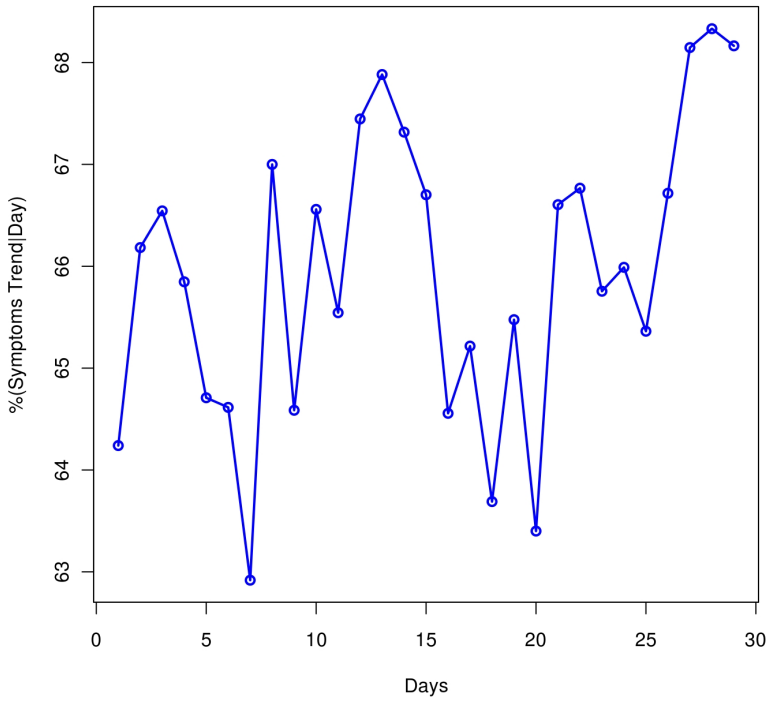

Figure 2.13: Symptoms trend as percentage of volume of observed public health messages (April 2012).

chatter, and scan each tweet to remove spam messages that contain a medical condition that are in fact not about public health (e.g., *"lol i'm infected with that damn Bieber-fever! !!:) ¡33 hehe but I love it"*), as well as remove drug and vaccination-related messages (e.g., *"Getting my last shot of hepatitis B vaccine today"*).

The classification process acts as a gate keeper to classify tweets of public health overtime into relevant or non-relevant data to the public health surveillance. Named Entity Recognition(NER) techniques performed for public health related tweets by recognizing the type of the tweet (e.g., personal, general, negation, or any other type of post), extract entities in each tweet such as symptom, normal diseases, viruses, bacteria, outbreaks, and time. In NER stage, the semantic analysis allows public health-related tweets to be more valuable and un-

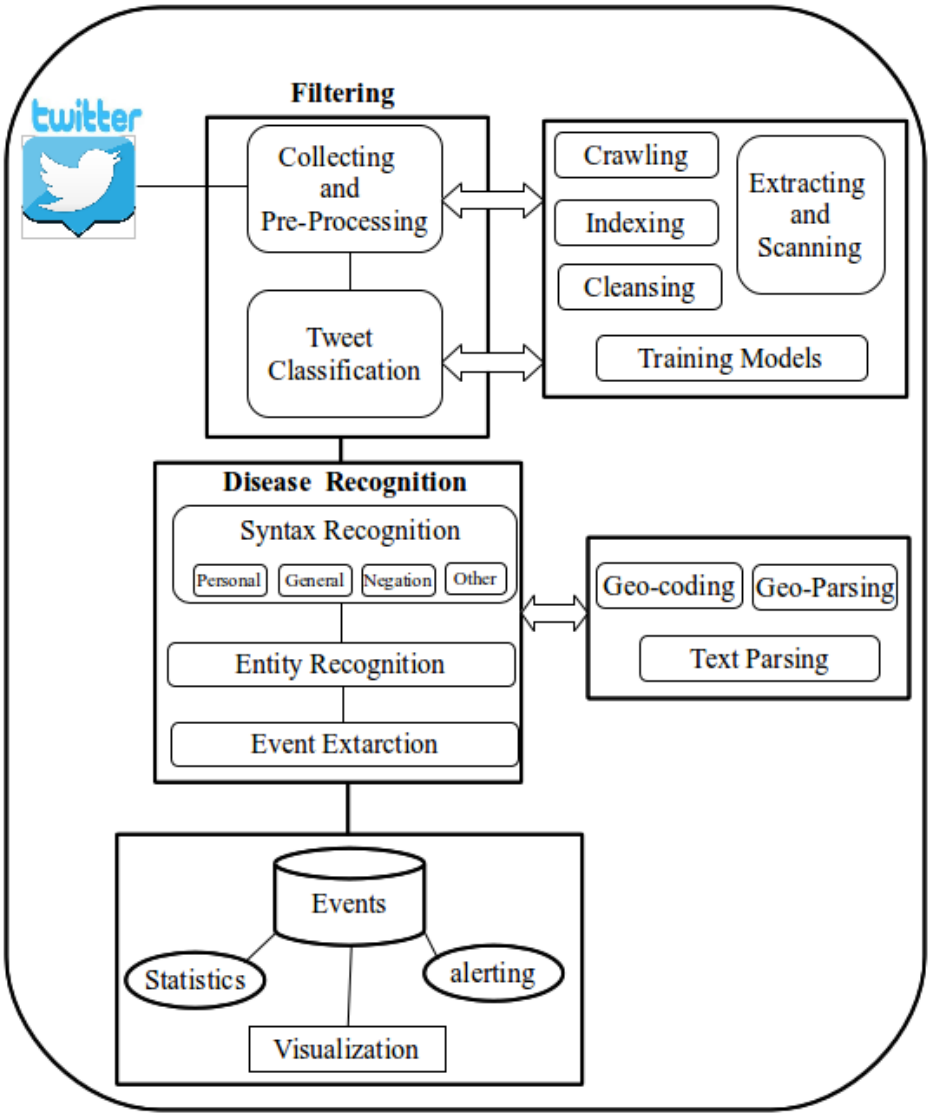

Figure 2.14: Overview of diseases surveillance system

derstandable by parsing the plain text. According to the nature of data of public health on Twitter, geoparsing and geocoding are the two techniques that are required to extract location entity from text or user's profile. All extracting entities are grouped as events and stored in a database. Statistics functions will be used to get the number of medical cases in a location and compare those cases with ground truth data (e.g., government data). Moreover, the alerting function is responsible for early alerting when something unusual might be happening (e.g.,

infection of disease in some location). Finally, the visualization process will be used to explore public health events and statistics.

Chapter 3

Medical Case: Annotation and Classification

The work in this chapter is my own except the annotation of data which has been done by me, Kerstin Denecke and Avare Stewart.

Because of the immense volume of Twitter messages(tweets) sent every day, it is extremely important to identify the messages that are relevant to disease monitoring purposes. Therefore, a machine learning classifier that distinguishes relevant Twitter postings (tweets) from irrelevant ones and performs well with noisy and sparse data is developed. This section focuses on the identification of tweets related to disease reporting. Briefly, a disease reporting tweet is one in which a person tweets about symptoms or claims that he/she has a certain disease. Existing approaches for text filtering in the medical domain, which use keyword lists or more plentiful text to identify relevant articles, do not take into account the distinctive characteristics of twitter messages. However, the characteristics of twitter messages are described with respect to linguistics and content (Section 2.2). Taking into account the characteristics of twitter messages, it was examined that how well established case definitions can be exploited to formulate annota-

tion guidelines that are used to define disease-reporting messages. Depending on annotation guidelines, a machine learning classifier is required to distinguish relevant disease-reporting postings from irrelevant ones. The approach is referred as medical case-driven annotation and classification.

In summary, the contributions presented in this chapter are as follows:

- An annotated Twitter dataset with disease- and non-disease reporting messages is built.
- The data collection and preprocessing, experimental setup, and the results of classification experiments are described. The feature set of collected tweets and their defectiveness on the classification results are also studied.

The section begins by outlining the annotation guidelines (Section 3.1), describing the data collection and processing, (Section 3.2) and then evaluating the techniques on an annotated dataset (Section 3.3).

The bulk of this chapter appeared in [53].

3.1 Micro-Messages Case-Driven Annotation

In a manual annotation process of tweets, we want to learn more about tweets and establish a dataset of microblogs for the purpose of relevance filtering. For this purpose, we labeled microblogs as disease-related (*positive*) or non-disease-related (*negative*). A tweet is considered *positive* when an individual is providing information about his own or someone else's health status, whereas *negative* does not fulfill these criteria. More details on the definition of *positive* and *negative* are described in the annotation guidelines (see Section 3.1.1 and 3.1.2). Furthermore, The label *unknown* is introduced for tweets that are not understandable or where the annotator cannot make a decision. The number of *undecidable* tweets was too small to have enough training material for an additional class. Therefore, these tweets are categorized as positive or negative class after discussions between the annotators. The classifier will consider a two-class classification problem. The main purposes of the annotation are to learn more about the nature of disease-reporting tweets and about tweets where a person has difficulties to decide from one of the two classes and to get a labeled dataset for training a classifier.

To ensure consistent labeling, annotation criteria have been specified and an annotation tool has been implemented. Furthermore, an annotation experiment is carried out where the inter-annotator agreement has been determined (Section 3.3.1).

In order to formulate criteria for classifying the noisy-sparse non-official data of Twitter, case definition guidelines are exploited. A *case definition* is a set of standard criteria for deciding whether a person has a particular disease or health related condition. These criteria can be clinical, laboratory, or epidemiological. This method was introduced by public health professionals to define who is included as a *case* in an outbreak investigation. In the medical intelligence domain, these guidelines are set forth by WHO for official institutions such as hospitals or health organizations to know when to notify officials about unusual or unexpected events[1]. The case definitions define a case in time, person, and place. For the annotation guidelines, we are mainly interested in three categories of cases:

1. *Putative Case*, e.g., three Chinese are suspected to have swine flu.
2. *Probable Case*, which is a suspect case with some evidence like X-rays.
3. *Confirmed Case*, which confirms someone is directly infected by an outbreak.

We can annotate the tweets into positive or negative based on these case definitions. In more detail, we consider any tweet that denotes that a person (a) has certain symptoms or diseases or (b) is infected by some disease, as a *case*. Any tweet that describes a case or an unusual event should be annotated as positive, otherwise negative. A tweet is defined as a case if the content of the tweet refers to an object infected by a disease or symptom. This object could be a person, animal, or plant. Specific terms referring to symptoms or diseases and also verbs used in a tweet, play an important role to determine whether the tweet is a case or mentions an unusual outbreak. We have collected some verbs and terms that were helpful for annotators to annotate tweets. These verbs and terms are presented in Tables 3.1 and 3.2. More annotation rule details on the two classes are presented in the following sections.

[1] http://apps.who.int/gb/ghs/pdf/IHR_IGWG2_ID4-en.pdf

3.1.1 Positive Disease-Reporting Micro-Message

Any tweet should be labeled as positive or case regardless whether it is a confirmed, putative, or probable case:

1. If it confirms that the user is infected with a disease or symptom, e.g., *I am sick now. I got influenza and I need medicine.*
2. If it confirms that another subject (e.g., animal and plant) has a disease or symptom e.g., *2 horses in Georgia contracted West Nile virus.*
3. If a test result is mentioned that confirms an infection, e.g., *Tyler is influenza positive!!!!*
4. If a suspicion is mentioned, e.g., *my son is suspected of having swine flu.*
5. If another outbreak or danger is described e.g., *six cases of Malaria have been reported from southern Greece, including a traveler.*

The result of analysis showed that there are some verbs and terms that are helpful to annotate as *positive* tweets. We distinguish between infection verbs, detection verbs, medical terms, and additional terms with the help of examples given in Table 3.1.

3.1.2 Negative Disease-Reporting Micro-Message

Any tweet which confirms that there is no case or which contains text that is unrelated to a case is labeled *negative*. A negative tweet is any tweet similar to the following examples:

1. A tweet is a question, e.g., *What is this Bieber-fever Thing?*
2. It contains a condition, e.g., *If i have the flu again i will kill someone.*
3. It offers advices like *#Kids health: you should prevent your child from getting #dengue fever.*
4. It negates an infection, e.g., *I do not have measles.*

Word Category	Example
Infection Verbs	affect, infect, got, come down, suspect, down with, have, has
Detection Verbs	find, confirm, detect, discover
Medical Terms	death, fatality, case, hospital, patient, victim, clinic, pain, ill, sick, ache, doctor, outbreak, hurt, inflammation, negative test
Additional Terms	start up, hurt, running, stricken, squeaks, soreness, think

Table 3.1: Examples of useful words for labeling a tweet as positive

5. It contains a disease definition, e.g., *Dengue fever, also known as breakbone fever, is an infectious tropical disease caused by the dengue virus*; statistics, e.g., *Cervical cancer killed 900 a year in PNG* , describes past outbreaks; or jokes about diseases or outbreaks, e.g., *I love your links, Your links infected me with the fever.*

6. It is outside the disease-outbreak domain.

In addition, there are some verbs and terms that are helpful to annotate tweets as *negative*, e.g., education verbs, examination verbs, other verbs, and medical terms. For detailed examples, see Table 3.2.

These rules are applied by human annotators to annotate a dataset for the experiments. Furthermore, an inter-annotator study is performed with results reported in Section 3.3.1.

3.1.3 Annotation Tool

The annotation tool for manually annotating Twitter messages has been implemented by Java Servlet with JavaServer Pages (JSP), as shown in Figure 3.1. The

Word Category	Example
Education verbs and nouns	lecture, seminar, report, teach
Examination terms	examine, check, screen, diagnose, prognose
Medical Terms	drug, medicine, research, tablets, study, positive test, treatment, vaccine, immunization
Additional Verbs	cure, survived, get well, feel better

Table 3.2: Examples of useful words for labeling a tweet as negative

tool allows a user to login, to provide annotations for Twitter messages, to search for messages matching a query, and to look at annotated postings. It removes retweets automatically.

More specifically, the annotation user interface (see Fig. 3.1) provides the below mentioned facilities.

1. To search for up-to-date Twitter messages by using one or multiple keywords in the language of preference. This enables any user to label tweets that he/she gets directly from Twitter.

2. To read and annotate files that have been collected through the Spinn3r API.

3. To go through the annotated twitter messages.

4. To get and annotate tweets from specific users such as official sources (WHO, ProMed-mail, CDC, and UNICEF).

As well as the annotation page (see Fig. 3.2) provides the following facilities:

1. To save the annotated tweets into separate files.

Figure 3.1: Annotation tool GUI with search functionality and facilities like file and language selection.

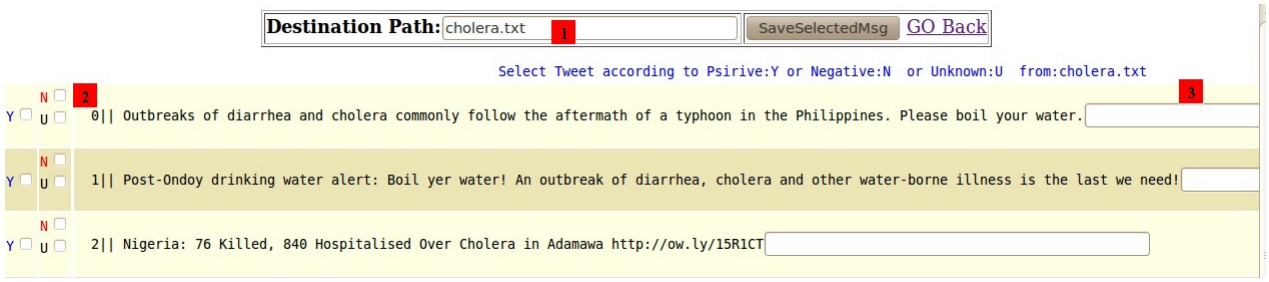

Figure 3.2: Annotation view of the Annotation tool.

2. To mark tweet into positive(Y) or negative (N) according to annotation guidelines, as well as the annotator is allowed to mark a tweet as unknown (U) when he/she cannot make a decision.

3. To make comments on the tweet to explain his/her decision. In this way, the annotator gets the opportunity to describe when he/she was unclear in deciding.

3.2 Data Collection and Processing

This section describes the data collection and preprocessing steps performed before running the classification experiments.

In this study, Twitter is used as a data source for identifying data relevant to monitoring medical conditions. For collecting Twitter data, Spinn3r, which is a web service for indexing the blogsphere is used. Spinn3r provides access to every posting on blogs, twitter, etc. that are being posted in real time. The data collection process comprises three steps:

1. **Data collection through the Spinn3r API.** The Spinn3r API[1] offers possibilities to collect content from weblogs, microblogs, social media, forums, and mainstream news in real time. Within 24 hours, the Spinn3r API returns roughly between 30000 and 35000 files, from which each file contains between 700 and 1000 objects. In total, the storage size required for incoming data within 24 h is between 90 and 100 Gb.

2. **Indexing.** Because of this immense volume of daily incoming data, Apache Lucene is used to index the files collected in step 1 to allow efficient retrieval and reduce the required storage size.

3. **Extracting Twitter data**. In the last step, data of a specific source (e.g., microblogs, weblog data, social media data, and forum data), in a particular language, or matching a specific keyword is extracted from indexes into text files where one text file contains a set of postings. Furthermore, data from specific users can be collected. This was applied to get Twitter messages in English from official sources, more specifically from the World Health Organization, ProMED-mail, and Center of Disease Control.

In addition to the Spinn3r API, Twitter Search API is also used to get tweets for one or more keywords. Subsequently, given the presence of misspellings and slang on Twitter, there are many undesired words and special characters that might have a negative effect on a classifier. Therefore, regular expressions are used within the tokenization process to clean the tweets. In particular, Twitter user names, URL links (e.g., http://bit.ly/dAuNZh), emoticons, numeric values, and words starting with a digit are removed from tweets. In addition, some special characters are were removed from every tweet, such as brackets.

3.3 Experiments

In this section, the results of the inter-annotator study are presented first. The experimental set up is then described to evaluate the classifier. Furthermore, the results of the classification experiments are presented.

[1] http://spinn3r.com/

User	Positive	Negative	Unknown
User 1	46	149	5
User 2	37	161	2
User 3	50	141	9

Table 3.3: Number of positive, negative, and unknown labeled tweets out of 200

3.3.1 Annotation Study

This section describes the annotation study and the observations that made during data labeling. The objective of this experiment was to ensure that the use of the annotation guidelines support the annotation strategy of a larger training set. We assessed whether annotators could label different datasets (if they agree to a large extent) or whether they need to label the same dataset, because of high disagreement. Three subjects were asked to label 200 tweets using the annotation tool. These tweets were randomly collected from the dataset described in Section 3.2. From this dataset, tweets have been selected such that it matches at least one of the keywords listed in Table 3.4, (e.g., smallpox, dengue fever, and flu).

The annotators were asked to label the tweets by considering the annotation guidelines described before. All annotators are computer scientists, who are working in the area of medical informatics. The results are shown in Table 3.3.

The numbers show that the annotations differ to a certain extent. The Kappa metric[1] is used to measure the agreement between the three annotators, and the Kappa value obtained was 0.793333. Based on the interpreted guidelines reported in [64] we concluded, for the final annotation, that the agreement was sufficient to ask the subjects to label different data.

The disagreement mainly occurred for tweets that made the confused annotator to choose a category class because of missing criteria in the annotation guidelines. For example, for the following tweets the annotators disagreed:

- *"Brent tested positive for influenza a...Alexis tested negative but with her symptoms and the fact that we all have it".* should be labeled positive because the case is confirmed.

- *"mom says she hopes I get malaria. Obvz she loves me lots".* should be

[1] http://justus.randolph.name/kappa

labeled negative because there is no confirmation.

- *"U want me to get malaria"* should be labeled negative because there is no case.
- *"My daughter is laid up . I wish she isn't influenza".* should be labeled negative because it is not a confirmed case.

The final annotated dataset comprises 5880 tweets:- 2130 positive and 3750 negative tweets.

3.3.2 Methods

- **Naive Bayes**

  The naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. The classifier has been successfully used in the text classification system [39]. Let $C=(c_1,...,c_m)$ be m document classes. Given a new unlabeled document D and its corresponding word-list $W=(w_1,..w_d)$. The naive Bayes approach assigns D to a class c^*_{NB} as follows:

$$C^*_{NB} = argmax_{c_j \varepsilon C} P(c_j) \prod_{i=1}^{d} P(w_i|c_j) \tag{3.1}$$

  where $P(c_j)$ is the a priori probability of class c_j, and $P(w_i|c_j)$ is the conditional probability of the word w_i given class c_j. The underlying assumption of the naive Bayes approach is that for a given class c_j, the probabilities of words occurring in a document are independent of each other.

- **Support Vector Machines**

  Support Vector Machines (SVMs) are a technique used in supervised machine learning and are typically used for classification problems (text categorization, handwritten character recognition, and image classification, etc.). Given a set of categories, which contain an arbitrary number of items, SVMs

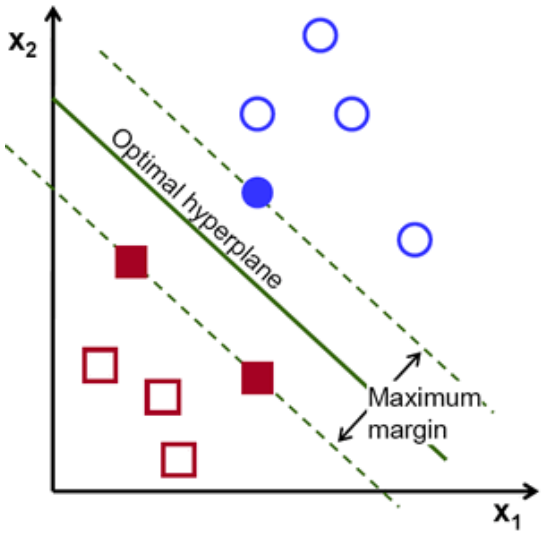

Figure 3.3: Separating hyperplane in SVM.

predict which category a new item belongs to. The theoretical background of SVMs is explained detailed in [10].

As illustrated in Figure 3.3: Given 2 categories (red items and blue items in the figure). Then, the operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. This distance receives the important name of margin within SVM's theory. Therefore, the optimal separating the hyperplane maximizes the margin of the training data.

- **Feature Selection Methods**

  The feature selection algorithm is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Moreover, the algorithm makes training and applying a classifier more efficient by decreasing the size of the effective terms and increases classification accuracy by eliminating noise features [24]. Feature selection combines two parts:

  1. a search method to find a set of features that is a good predictor of what class a sample belongs to?

2. an evaluation method, after having a good set of features from search methods, that tests how well those features used to predict the class?

- **K-fold Cross-Validation**

  Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: the first is used to learn or train a model, and the second is used to validate the model [21]. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold cross-validation or involve repeated rounds of k-fold cross-validation. In more details, in k-fold cross-validation, the data is first partitioned into k equally sized segments (folds). Subsequently, k iterations of training and validation are performed such that within each iteration different fold of the data is held-out for validation, whereas the remaining k-1 folds are used for learning.

- **Part-Of-speech Tagger**

  Part-of-speech tagging is a process wherein tokens are sequentially labeled with syntactic labels. In fact, it is a program that reads text in some language and assigns parts of speech, such as noun, verb, adjective, and adverb, to each word. These detailed feature sets produced from the POS tagger are considered to be the most relevant feature subset to enhance tasks of text classification. An example of POS tagger is the Stanford Log-linear Part-Of-Speech Tagger[1] that was developed by Stanford University and used for most of the studies.

- **Performance Measures**

  The performance measures are computed as follows [68]:

  - $Recall = \frac{TP}{(TP+FN)}$
  - $Precision = \frac{TP}{(TP+FP)}$

[1] http://nlp.stanford.edu/software/tagger.shtml

- $Accuracy = \frac{TP+TN}{(TN+TP+FP+FN)}$
- $F - Measure = \frac{(2*Precision*Recall)}{(Precision+Recall)}$

where TP=true positive, TN=true negative, FP=false positive, and FN=false negative.

- True positive tweets are those positive tweets that are correctly classified as positive by the classifier.
- False positive tweets are those tweets that are labeled positive but incorrectly assigned by the classifier.
- True negative tweets are those negative tweets that are correctly classified as negative by the classifier.
- False negative tweets are those tweets that are labeled negative but incorrectly assigned by the classifier.

3.3.3 Experimental Goals and Setting

The purpose of the evaluation is to prove the following hypothesis: A feature set considering the peculiarities of tweets is better suited for classification purposes than a simple bag of words approach. To analyze the quality of the classifier, the dataset is divided into 10 folds; each fold has 375 negative tweets and 213 positive tweets. Different feature sets are used in this experiment, and they are as follows:

1. Term frequency of all terms (referred to as baseline).
2. Term frequency of cleaned term set (referred to as baseline with cleansing).
3. Baseline with a cleansing feature set reduced by information gain.
4. All named entities, pronouns, verbs, adverbs, and nouns.
5. Medical conditions and treatments, pronouns, verbs, and adverbs.
6. A 1000 terms with highest frequency, medical conditions and treatments, nouns, verbs.

A baseline result is produced by exploiting term frequencies of all words as features (bag of words).

WEKA toolkit has been used to select attributes automatically. In more detail, for feature selection a ranking is used as a search method and information gain as the evaluation method. The top 1000 features are selected for the feature set 3. To assess and compare the classification quality for the various feature combinations, accuracy, precision, recall, and F-measure are determined.

To detect tweets relevant to disease or outbreak monitoring, two machine learning algorithms are exploited to build a classifier, i.e., Naive Bayes and SVMs.

Both algorithms are analyzed for their suitability on the data collection. First of all, the feature sets are determined. After that, a 10-fold cross-validation is performed to assess the accuracy of the classifier on the dataset. The experiments are implemented under a Linux operating system using WEKA[1] software for both Naive Bayes and SVMs.

3.3.4 Results

3.3.4.1 Guideline Exploitation and Annotated Data

Using the annotation guidelines reported in Section 3.1, the collected tweets have been manually labeled *positive* or *negative*.

Table 3.4 summarizes the keywords or phrases used for collecting tweets. The focus of this work is to classify tweets in English. The number of collected tweets per term differs, because not all topics have the same popularity or relevance every time. The decision is made for relevant keywords by their frequency in a random sample: from 2000 tweets, tweets with medical conditions were extracted. The most frequent ones were then used as keywords for collecting the dataset. The table also shows the number of tweets per keyword or phrase. There are tweets that contain more than one search word, (e.g., *"I'm so SICK! Infected with..... some flu:("*) which contains words "Infected", "with", and "SICK". This labeled data will provide the training material for the classifier. In order to include data that is totally unrelated to health and medicine, 1144 tweets have be included into the dataset that fall into another domain (e.g., into sports and technologies).

[1] http://www.cs.waikato.ac.nz/ml/weka/

The complete data collection was comprised of 5880 tweets, in which 2130 tweets were labeled *positive* and 3750 tweets belonged to the class *negative*.

3.3.4.2 Medical Case-Driven Feature Analysis

For the classification experiments, several feature sets exploited, that are words of specific word classes and named entities of specific categories using OpenCalais[1], an open-source named entities extractor. One shortcoming of OpenCalais is that it does not perform well with short texts. Often, no results are produced for such texts. According to a study from Maricel [40], the text should be more than 100 words and less than two pages to get optimal result from OpenCalais. Because tweets are very short texts that do not fulfill this requirement, OpenCalais software is applied to sets of 20 tweets to be processed at one time. Then, this operation is repeated on the whole dataset. For annotating the words with their part-of-speech the Stanford part-of-speech tagger is used. Table 3.5 shows the frequencies of parts-of-speeches, word classes, and all types of named entities for the annotated dataset. It can be seen that in negative labeled tweets all parts-of-speech occur more frequently than in positive-labeled tweets.

The same holds true for named entities. A reason for this is that lots of negatively labeled tweets originate from news agencies or are announcements. Thus, they contain complete sentences. Conversely, most of the positively labeled tweets are posted by normal people, and the tweets are characterized by short phrases and enumerations of nouns. It is known to us that existing tools such as OpenCalais and the Stanford Tagger have been developed to process complete documents. Their performance on short messages such as tweets has not been assessed so far. It remains a future issue to evaluate their accuracy on tweets and microblogs. All nouns and verbs that have been explained in the annotation process for labeling tweets as positive or negative are used as features for the classifier.

[1] http://www.opencalais.com

3.3.4.3 Micro-Message Classification

Table 3.6 summarizes the results of the experiments. Without any preprocessing or cleaning of the tweets, an accuracy of 74% (Naive Bayes) or 81% (SVMs) could be achieved. The cleaning process described before led to a significant increase in the accuracy of both the classifiers. In more detail, accuracy values of 83% (Naive Bayes) and 87% (SVMs) were achieved for a cleaned feature set. When applying information gain to reduce the features further to 1000, the accuracy values improved slightly. The highest impact of such features selection is a significant increase of the recall for the SVM classifier.

For all other feature sets, a significant improvement can be recognized on the accuracy when comparing to the baseline. However, compared to the results with a cleaned feature set, the improvement of other feature combinations is insignificant. It is concluded that use of named entities and restriction of word classes for building feature sets for classification purposes does not help in improving the classifier. The reasons are discussed in the following section.

3.4 Discussion

In this chapter, annotation guidelines are developed, a tool to realize the annotation process and learned from an annotation process on how to well classify Twitter messages that are relevant to monitoring medical conditions. Several points that were discussed in this chapter are now highlighted.

Annotation Tool: There is an increasing need for collecting data from Twitter, especially of disease-reporting tweets. Important factors to consider for such data collection are the time and reliability of tweets. Therefore, the annotation tool that has been developed was fast, very easy to handle, and helpful for manual annotation with respect to any specific topic.

Annotation Guidelines: Well established guidelines have been adopted for defining clinical cases to build a dataset for a disease-reporting classifier and to learn about the characteristics of the dataset. An empirical experiment implemented on 200 tweets that were annotated by three subjects showed a good agreement according to Kappa performance. The problems reported by the an-

notators in the empirical experiment helped improve annotation guidelines.

Disease-Reporting Classifier: Building a machine learning classifier is the major target of this chapter, which distinguishes between relevant and irrelevant diseases postings on Twitter. In this chapter, the classifier has been developed to resolve the problem definition which is to filter disease-reporting posts. Two classification algorithms have been exploited and tested on several feature sets. In the experiment, the performance of the classifier achieved up to 89% of accuracy for support vector machines.

Indeed, the classification was a hard classification problem in comparison with Web Search Domain Disambiguation. As have been seen in the results of the experiment in Table 3.6, the feature set did not perform well because most of the tweets from the same domain and different classes share a lot of common medical terms, which implies that the data are much less discriminative. Three reasons for wrongly classified messages could be stated as follows:

1. Shared terms
2. Short messages
3. Slang words and lingos

- **Shared terms**

  As aforementioned in the annotation guidelines, words that are used to write about symptoms and diseases were identified. Nevertheless, some of them are also used in negative tweets, which make it difficult to consider these terms characteristically for positive tweets. For example, the verbs *have* and *got* are two popular verbs related to positive tweets like *"I got ...", "I have ..."*. The frequency of those two verbs in positive tweets is 886 (have) and 559 (got). In negative tweets the frequency is 757 (have) and 227 (got). For the verb *have*, the frequency is similar; thus, the observation that these two words are not representative features with respect to different domains is confirmed. Moreover, there is not much lexical difference between the tweet *"I have measles"* that is annotated as a *positive* tweet and the tweet *"I don't have measles"* that is annotated as a *negative* tweet. Therefore, it is difficult for machine learning algorithms to discard this confusion.

- **Short messages**

  Tweets by nature are very short messages. Therefore, the number of terms per tweet is very limited, and it is sometimes insufficient for the classifier to decide if the tweet is *positive* or *negative*. It still needs to be tested whether other classifiers are better suited. Support Vector Machines (SVMs) algorithm was chosen as other works and show that this algorithm can deal with few irrelevant features and sparse document [31].

- **Slang words, lingos, and writing errors**

  Language in tweets is very noisy and comprises grammatical errors. Words are written incorrectly and slang words are used, e.g., *ah, ohh, omg, aghh.* This makes the preprocessing by named entity recognition, part of speech tagging, and classification difficult, thus leading to errors. Lingos remained unconsidered in that work. Their influence on the classification accuracy still needs to be analyzed.

3.5 Conclusions

The study in this chapter focused on identifying disease-reporting Twitter messages. An annotation tool has been developed to collect and annotate Twitter messages. Furthermore, criteria to label tweets as positive or negative have been documented. Several feature sets have been tested with two classification algorithms to automatically filter tweets. Although tweets are very short and noisy, support vector machines performed with an accuracy of up to 89%. The main outcome of this chapter is a dataset of annotated twitter data as a "gold standard" benchmark. There are still open research questions that need to be addressed, and the dataset can ensure comparability in experiments. Therefore, in the next chapter, The preprocessing of classification task was improved to manage the challenges and make the classification model work in a streaming scenario in real-time with the Twitter stream.

Keyword or phrase	**Positive labeled**	**Negative labeled**	**Total**
Malaria	157	291	448
Fever	64	89	153
Dengue	67	125	192
Yellow Fever	54	79	133
Influenza	60	86	146
Measles	107	200	307
Poisoning	130	93	223
Cholera	63	112	175
Typhoid	82	95	177
Hepatitis	86	152	238
Smallpox	14	90	104
Headache	181	57	238
Tuberculosis	33	32	65
Polio	28	57	85
Tetanus	19	63	82
Otitis	17	12	29
Ebola	37	64	101
Rash	38	19	57
Gout	33	34	67
Tonsillitis	92	22	114
Allergies	18	18	36
Arthritis	44	53	97
Plague	61	67	128
Diabetes	50	98	148
HIV/AIDS	17	86	103
Cancer	42	64	106
Syphilis	20	19	39
infected+with +sick	19	42	61
case+Outbreak	78	10	88
Pneumonia	82	32	114
Appendicitis	52	34	86
Asthma	55	31	86
Kidney Failure	49	21	70
Tumors	26	10	36
Anemia	6	6	12
Norovirus	27	37	64
Diarrhea	49	105	154
infect+fever+ sick+virus	73	101	174
Non-medical	-	1144	1144
Sum	2130	3750	5880

Table 3.4: Annotated data showing the type of medical annotation and the corresponding number of positive and negative labeled tweets.

Type or Word class	**Frequency in positive documents**	**Frequency in negative documents**	**Total**
Verbs	6509	10397	16906
Adjectives	2480	4477	6957
Pronouns	1843	3051	4894
Adverbs	1910	2828	4738
Medical condition and treatment	829	1142	1971
Products	19	41	60
Location	66	166	232
Person	40	90	130
Technology	8	32	40
Organization	20	82	102
Industry Terms	22	213	235

Table 3.5: Frequency of named entity types and word classes in positive and negative labeled texts and POS tags.

	Naive Bayes				**Support Vector Machines**			
Features	**Acc**	**Rec**	**Pre**	**F-M**	**Acc**	**Rec**	**Pre**	**F-M**
Baseline	74.88	62.68	75.77	68.61	81.40	73.52	76.01	74.74
Baseline with cleansing	83.20	75.09	80.23	73.73	87.47	82.56	82.91	82.73
InfoGain with 1000 Features	83.45	75.07	80	77.46	89.18	88.79	80.28	84.32
all NEs + Pronouns +verbs+ adverbs + N	81.89	74.00	77.09	75.51	86.12	83.21	77.29	78.82
Medical condition and treatment + Pronouns + verbs + adverbs	81.8	74.01	76.85	75.4	85.30	81.84	76.38	79.02
1000 Term MC+MT+N+V	81.33	72.49	78.08	75.18	86.12	82.49	78.31	80.35

Table 3.6: Performance results of the classification experiments (Accuracy (Acc), Recall (Re), precision (pre), F-Measure (F-M), Medical Condition (MC), Medical Treatments (MT), Nouns (N), and Verbs (V) listed in the annotation guidelines).

Chapter 4

Streaming Scenario

In the previous chapter, a significant type of classification task was introduced. Specifically one that is crucial for medical intelligence gathering. A more general approach that discovers many different diseases, outbreaks, and symptoms has been presented. In addition, it was analyzed whether domain or task specific trends are better suited for this classification task. Importantly, in Section 2.2.2, the types and characteristics of public health data on Twitter such as relevant disease-, vaccination-, or Spam-related messages were presented. Furthermore, the Section 3.4, presented most of the challenges that had an effect on the classification process of disease reporting in Twitter messages. In fact, automatically identifying Twitter content associated with disease-reporting messages is a challenging problem because of the heterogeneous and noisy nature of the data.

Therefore, in this chapter, the design, implementation, and evaluation a real-time system is developed for collecting and filtering disease-related postings. The system tracks peoples' status updates in real-time and works with an input Twitter stream. The metric is using the classifier that was built in Chapter 3 to filter postings into disease-related or disease-nonrelated. It has the ability to distinguish between real disease-, shot or vaccination-, and Bieber-fever-related postings.

In summary, the contributions in this chapter are as follows:

- Describing the components of the real-time filtering system that works over-time with a Twitter stream(Section 4.1).

- Describing the scenarios that illustrate all functions of the filtering system that will apply on Twitter messages(tweets) from the time of posting by the user to the time that the post is considered as a real disease-related message (Section 4.2).
- Implementing the real-time system by using different data mining techniques, and a multi-threading technique are applied to make the system's components work concurrently in real-time (Section 4.3).
- Evaluating the run-time performance of the filtering system with respect to latency and throughput (Section 4.4).

The bulk of this chapter appeared in [54] and [55].

4.1 The Real-time Collecting and Filtering System

This section describes all components of the real-time system that are working concurrently with incoming postings from Twitter. Figure 4.1 depicts that the system comprises nine components. The lines between components indicate the communications between the components with some relative functions labeled on those lines. The functions are used to run specific tasks during communication.

4.1.1 Crawler

Web crawling is a process used by search engines to collect pages from the World Wide Web. Topical crawlers are also used to collect pages that are relevant to a particular topic(s) [42]. Twitter is used as a source to crawl the public health data. However, Twitter Streaming API[1] allows high-throughput near-real-time access to various subsets of public and protected Twitter data. The crawler component in the system uses Twitter API to crawl public status updates of Twitter users that mentions any medical term in its content. Some of these medical terms are listed in the Table 4.1, which contains disease and symptom names, and the

[1]https://dev.twitter.com/

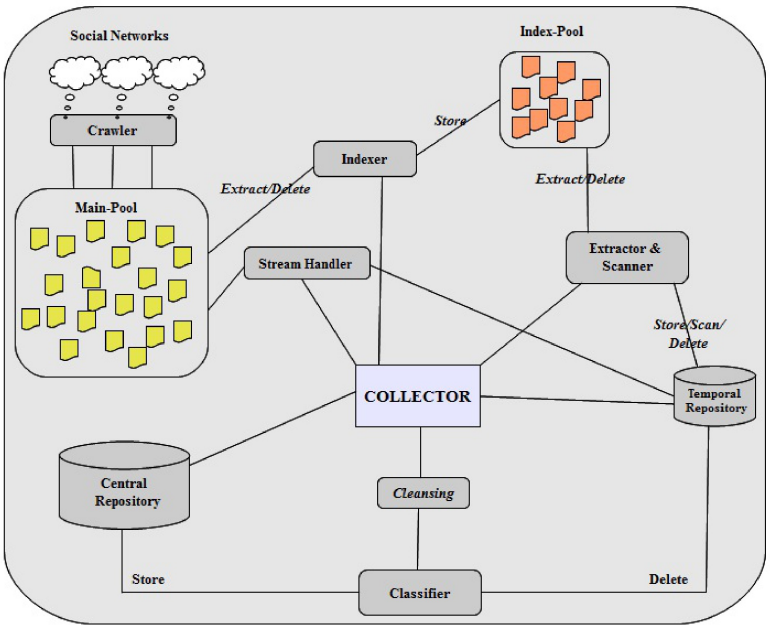

Figure 4.1: The real-time collecting and filtering system.

outbreak-related terms. In part of the second s, the component gets N tweets that differ from one second to another in amount. The real-time property of Twitter Streaming API are exploited to build the system with the same stream and filtering disease-related postings as quickly as possible. Therefore, this component is used to collect data immediately and save them in JSON format into a pool known as Main-Pool.

4.1.2 Indexer

Because of the immense volume of incoming data from the crawler component, the indexer uses Apache Lucene[1] to index JSON files data to allow efficient, fast, and accurate information retrieval. This component works concurrently with the crawler component, and checks whenever the crawler gets new data. Subsequently, the component starts indexing those data to a small size and makes them searchable and analyzable. The output is stored in a pool known as Index-Pool. In addition, the indexer is used to structure the meta-data of each post.

[1]http://lucene.apache.org/

Thus, post text and its meta-data are structured in the following manner:

- POST_TITLE: The content of the post.
- TIME_PUBLISHED: Timestamps of the post.
- AUTHOR : The account name of the author.
- A_LINK : Hyperlink of the author.
- T_LINK : Hyperlink for title of the post, if found.
- POST_LANG: The language code of the post.
- USER_LANG: The language code of the user.
- USER_LOCATION: The location of the user.
- GEOLOCATION: The geolocation of the post, if found.
- PUBLISHER_TYPE: Type of publisher that is MICROBLOGS.
- RT: Stores the value 1 if the post is a retweet, otherwise stores the value 0.

USER_LOCATION field represents the location of the user, which is in free text form as Twitter users can assign valid or invalid location information in their profiles. Conversely, geolocation contains the longitude and latitude of tweets, when the user uses any mobility devices such as iPhone. The field RT in meta-data represents whether the post is a retweet or not. In fact, retweet posts that could be used for some studies in the future are indexed. Indexer component extracts each JSON file from the Main-Pool, indexes it into the Index-Pool, and after the indexing operation is performed, the file is removed to free up space for storage of new incoming data.

4.1.3 Extractor and Scanner

Crawler crawls and the indexer indexes postings (microblogs) into the Main-Pool and Index-Pool, respectively. The extractor component is responsible for extracting postings with their meta-data from the Index-Pool by using medical

malaria	ebola	deaths	virus
fever	rash	ill	hurt
dengue	gout	pneumonia	stricken
influenza	tonsillitis	appendicitis	epidemic
measles	allergies	asthma	soreness
poisoning	arthritis	kidney failure	cholera
plague	tumor	illness	typhoid
diabetes	anemia	pain	hepatitis
HIV/Aids	diarrhea	fatality	headache
cancer	infection	patient	tuberculosis
syphilis	sick	ache	polio
flu	cough	inflammation	tetanus
shivering	chills	malaise	nausea
check	norovirus	otitis	squeaks
h1n1	doctor	outbreak	stricken
bacteria	death	strep	dizziness
disease	viral	infectious	pandemic
smallpox	outbreaks	hospital	vomiting

Table 4.1: Some of the medical patterns detected by the real-time architecture.

term-related queries, and stores results in the temporal repository. The medical terms are the same terms that are used by the crawler component and are listed in Table 4.1. These terms are used by an extractor as a query string to retrieve the data from the index files (Index-Pool). Indeed, the extractor component is actually a search engine to retrieve each tweet that contains at least one of the medical terms. Furthermore, the extractor works as scanner to scan each tweet whether it contains one or more undesired words. Its functions are summarized as follows:

- Reducing the massive data that enters into the classifier component. This operation will mitigate the heavy load on the classification model by ignoring tweets that are posted from official sources such as WHO [45], and ProMED-mail,[1]. Moreover, the extractor removes all postings that are retweets.

- Removing all postings that are not understandable (spam) to the classifi-

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

cation model. Therefore, many new postings are sent at every second that are unknown to the classification model. They contain a lot of slang words, chatter, and non-understandable postings that make false decision for the model regardless whether they contain medical patterns. Therefore, the extractor removes all tweets (spam) that contain medical conditions that do not related to the interest.

Some examples of those tweets are described as follows:

- tweets related to Bieber-fever, which are postings about the pop star Justin Bieber(e.g., *"I think I'm getting a case of Bieber-fever"*).
- tweets related to vaccine or shot (e.g., *"Getting my last shot of hepatitis B vaccine today"*).
- tweets related to Typhoid Mary who was the first person in the United States identified as an asymptomatic carrier of the pathogen associated with typhoid fever (e.g., *"This Virus-Copter Is a Digital Typhoid Mary http://t.co/UEt2AF9C"*).
- tweets related to fever or Beadles Measles (e.g., *"I HAVE BEADLES MEASLES lol"*).
- tweets related to love (e.g., *"I'm infected. I have a virus. My virus is serious and contagious. I am in love"*).

The extractor works synchronically with the indexer and classifier components. Furthermore, each post(tweet) that has been extracted and scanned will be stored in the temporal repository until the classification process initiated. Subsequently, the post should be removed from the Index-Pool. This operation avoids the duplication if the post contains more than one medical term, as well as free up space for storage of new incoming data.

4.1.4 Collector

The collector is the entry point in the system and is responsible for communication between all the other components except the crawler, which works independently.

Term(s)	Positive	Negative
Cough	42	12
Sick	54	-
Hurt	36	13
Vomiting	25	7
Dizziness	27	13
Outbreaks	66	-
Sum	250	45

Table 4.2: Some new data that was added as training data.

4.1.5 Temporary Repository

This repository is used to store all postings returned by the extractor component, regardless whether they are disease-related or unrelated. It stores the data that is checked by the classifier component. After the classification process is done, all data in the repository will be automatically deleted.

4.1.6 Cleansing Process

Because of misspellings and slang on Twitter, there are a lot of undesired words and special characters that affected the classification process. Therefore, regular expression with tokenization process are used in order to process the tweet text, break it into words, and follow the below process:

- Remove twitter user names in the tweet. The user name appears after the symbol @ for example "@xyz".
- Remove all URL links (e.g., http://bit.ly/dAuNZh).
- Remove all emoticons and special symbols or characters.
- Remove all numeric values and any word that starts with digits.
- Remove most of the slang words and words that have continually repeated characters (e.g., aaaaaach, huuuungry, and cooooooool).

4.1.7 Classifier

The classifier component represents a machine learning classification model that is used to filter posts into disease-related or unrelated. In Chapter 3, a classifier was built that distinguishes between postings that are relevant to diseases and those that are irrelevant. The dataset is collected after adhering to the guidelines adopted for defining clinical cases to build "gold standard" data for a disease-reporting classifier and for learning about the characteristics of the dataset. Evaluation of disease-reporting classification was presented as well. The experiment was implemented on a dataset of 5880 tweets. The classification model achieved good accuracy values of up to 89% on support vector machines(SVMs). In addition to the previous dataset, a new data is added and is listed in Table 4.2 that contains tweets about symptoms and outbreaks. The new dataset contains medical terms that were not considered before especially the popular medical terms for symptoms (e.g., Dizziness, cough, and ache), and this dataset is annotated according to the annotation guidelines that are presented in Chapter 3. Then, this new dataset is added into the previous dataset and the final dataset contains 2380 disease-related tweets and 3795 unrelated tweets.

The classifier component is working as a gate keeper to filter all disease-reporting postings. As shown in Figure 4.1, after the post is cleaned by the cleansing component, the classifier checks if the post is disease-related or unrelated. Furthermore, the metric of the filtering system is represented by the performance of the classifier that will be presented later in Section 4.4.1.

4.1.8 Central Repository

The central repository is used to store all disease-related postings that are relevant to diseases, symptoms, or outbreaks. In more detail, this database is used to store the pure diseases-related data that represents the output of the filtering system.

4.1.9 Stream Handler

Because of the huge amount of data coming from the crawler component, the stream handler component is used to control the scalability of the system. It is

responsible to determine two thresholds for the amount of data in the Main-Pool and temporal repository, respectively. If the amount of data is more than the threshold, the stream handler changes the Main-Pool to another new pool, and informs another machine to execute the same system on the new data in that new pool.

4.2 Filtering System Scenario

In the previous section, all the components of the real-time filtering system have been described. In this section, more details about the execution scenario of the system will be given, and describe how the system works. All components in the system will be described in one scenario that is shown in Figure 4.2. The figure shows all the processes required for the life cycle of posting, from the time it is published on Twitter until it is stored in the central repository component as a relevant disease posting.

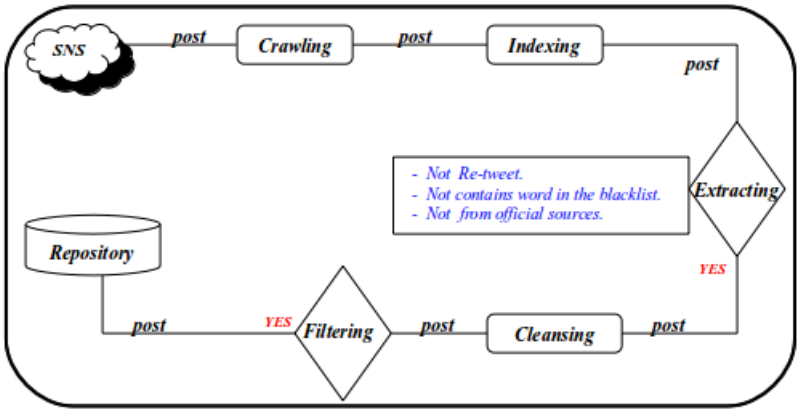

Figure 4.2: Filtering System.

The following steps describe the scenario:

1. From the left side of Figure 4.2, the crawler receives the Twitter post directly in a fraction of a second by using Twitter API.

2. The indexer takes the crawled post and indexes it with its meta information that is stored in a structured manner in a small space.

3. Each indexed posting is extracted via the extractor component by using medical term-related query to check if the post is not a retweet post, not from official source, and does not contain at least one or more words in the black list, that contains undesired words (e.g., Bieber-fever, shot, vaccine, Beadles Measles, and love).

4. If the post has not been removed in step 3, it enters into the cleansing process to remove slang words and special symbols from its content.

5. After cleansing, the post enters into the filtering process that uses the state-of-the-art classifier model to check if the post is disease-related or unrelated. Then, it stores only the relevant post in the central repository storage.

All above steps from one to five reflect all processes executed by the system that are crawling, indexing, extracting, cleansing, and classifying, respectively. Furthermore, in all decision stages that are presented in this scenario (e.g., extracting and filtering), if the decision result is false, then the post is automatically removed.

4.3 Implementation

The real-time collecting and filtering system has been implemented in Java under Ubuntu Linux 11.04 operating system. Each component in the system was implemented as an independent Java package and tested each of them independently. Subsequently, all of them were combined in one system else the crawler that is working independent overtime and returns data in JSON format each one contains 1000 postings with their meta-data. Apache Lucene has been used, which is a high-performance and full-featured text search engine library to parse and index data of JSON files into structured data as indexed files. Apache Lucene is also used by the extractor component as a search engine to extract or remove postings from indexed files. Furthermore, the SVMs' classifier model has been trained and evaluated by using 10-fold cross-validation via the Weka toolkit and then the Weka code was merged into system's code to make the classifier work overtime.

Moreover, the multi threading technique has been used to make the components of the system work simultaneously as a non stop process. Whenever the collector component starts, four threads will be started concurrently, i.e., indexing, extracting, streaming handler and classification, with each one is working overtime. Whenever each thread gets new data, it starts performing its job. MySQL is also used to create two databases, the first to store temporary data until filtering process starts and the second to store disease-related data, which represents the output of the system. Besides, Java vectors have been used to transport and store data during processing (e.g., load medical-related terms, list of official sources, or black list patterns from text files). Briefly described, all components are working together overtime whenever each gets new data.

4.4 Evaluation

In this section, the accuracy of the classifier and the run-time performance of the system will be evaluated. Indeed, to evaluate the system, it is mandatory to evaluate the performance attribute of its components to ensure if they fulfill the goals. Therefore, the classifier that is the stone corner of the system is evaluated. The results show that the system can identify medical tweets with up to 88% accuracy, and each tweet takes on an average 12.46 milliseconds(ms) to process from start to finish.

4.4.1 Classifier Performance

Chapter 3 presented the building of a machine learning classifier that distinguishes relevant from irrelevant disease postings on Twitter. The classifier has been developed to meet the goals of this work, and two classification algorithms have been trained by using a dataset that comprises 5880 tweets evaluated by 10-fold cross-validation. The performance of the classifier achieved up to 89% of accuracy for SVMs. Moreover, a new dataset is added that comprises many types of symptoms and infectious disease outbreaks (Table 4.2). Thus, the final dataset comprises 6175 tweets, 2380 disease-related, and 3795 disease-unrelated tweets. On the same methodologies that are used in Section 3.3.3 (Chapter 3),

the performance of the new classifier has been evaluated. In this evaluation, 10-fold cross-validation method was used with SVMs algorithm to assess the accuracy of the classification model on the whole dataset. After cleaning the dataset by the cleansing process, the performance that was achieved showed an accuracy of 88.261% with respect to the "bag of word" features. In addition to classification performance, the classifier is evaluated by testing it on the crucial dataset annotated manually (Chapter 6).

4.4.2 Run Time Performance

In addition to the performance of the classifier component in the real-time filtering system, the overall performance to execute is an important metric. In this subsection, the latency and throughput are measured. It was found that each tweet takes on average of 12.46 ms for processing and a throughput of six million disease-reporting tweets per day with respect to the characteristics of the machine that is used to run the system. The performance experiment has been run on Intel Xeon 2.40 GHZ (8 processors) with 2G of memory and running a 32-bit version of Ubuntu Linux 11.04.

4.4.2.1 Latency

The time delay in the system has been measured from when the system receives a tweet from the crawler, until the real-time system decides whether that tweet is disease-related or unrelated. Table 4.3 shows the processing time for a single tweet. In each operation, the indexer needs 0.487 ms to process a tweet from being taken from the Main-Pool until storing it in the Index-Pool. To extract a single tweet from the Index-Pool until saving it in the temporal repository, the extractor needs 11 ms. The classifier needs 0.975 ms to get a tweet from the temporal repository, crossing cleansing process until saving it in central repository. All the components were ran at the same time, which means that all the components are sharing the same memory and CPU except the crawler, which works independently. Therefore, the performance time of each component is different from time to time but the total time stills the same and the classification time is constant. Furthermore, the classification time includes the time that is needed

Component	Run Time(millisecond)
Indexer	0.487
Extractor	11
Classifier	0.975
Total	12.462

Table 4.3: The processing time for a single tweet.

for the cleansing process, as well as the classification process was speeded up by sending N postings each time into the classification model as 3000 postings need 2925 ms. It was found that each tweet takes on average 12.46 ms to process from start to finish. Surely, it was believed that this time will be reduced by using high-speed parallel servers.

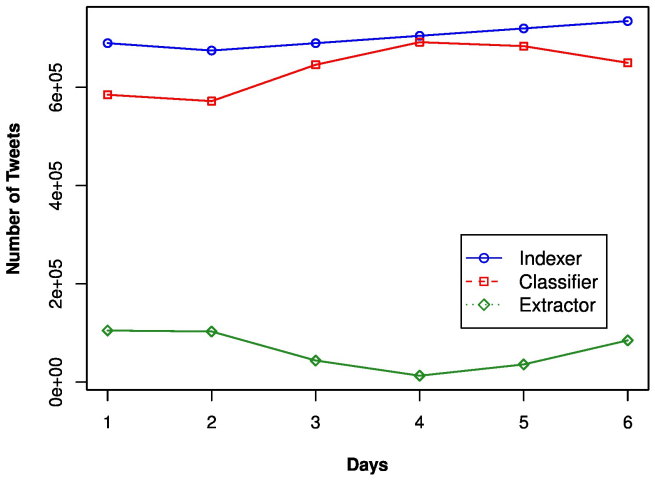

Figure 4.3: Number of tweets per day.

4.4.2.2 Throughput

The aim here is to satisfy the expected throughput of large social data. The throughput of the system has been measured with respect to the characteristics of the system that were mentioned previously. In fact, the scalability is enough to satisfy the throughout expected of large Twitter data as the system can process

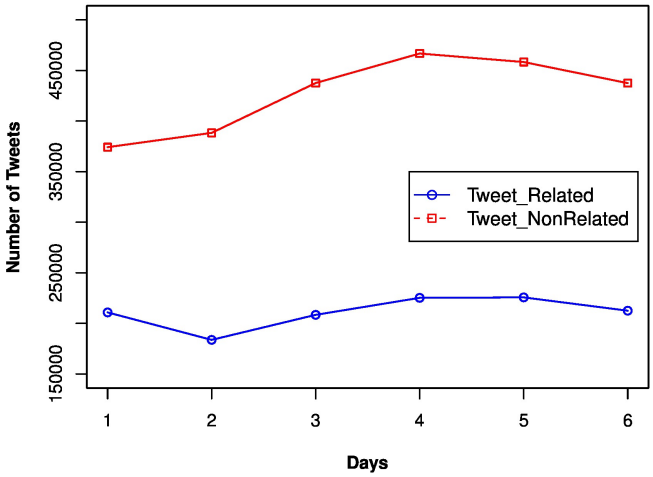

Figure 4.4: Number of disease related and unrelated tweets per day.

millions of tweets per day. Conversely, the stream handler component is responsible for determining the threshold of the data that can be processed. There is no exact number of tweets returned by the crawler each second as it depends on the tweets that contain medical conditions posted by the users. The tweets returned by the crawler are processed directly in batches to speed up the processing. The system was run overtime for six days and measured the number of tweets during each process of the system. Figure 4.3 shows the number of tweets for each day. The blue line shows the number of tweets processed by the indexer, which are the same tweets returned by the crawler. The red line is the number of tweets that enter the classification process, and the green line represents tweets that are removed by the extractor. In addition, Figure 4.4 shows the number of relevant or nonrelevant tweets per day produced from the classification process, the red and blue lines represent nonrelated and related tweets, respectively. Indeed, the total time for processing a single tweet is 12.47 ms; therefore, the system can process roughly more than six million tweets each day.

4.5 Conclusions

In this chapter, a real-time system has been developed for the collection, identification, and filtration of medical cases of Twitter messages. Symptoms and disease outbreaks can be detected as well. The system focused on detection of diseases on Twitter instead of traditional surveys in public health, which can be costly and time consuming. However, the system tracks the status updates of the population in real-time, and has ability to remove spams (e.g., shot or vaccination-related tweets, and Bieber-fever-related tweets). Furthermore, the system uses the state-of-the-art text classification to filter pure disease-related tweets. All components of the system and their tasks have been described, and the scenario of the system was then explored that reflects all operations that were performed on each tweet. In particular, the system has been implemented, and the valuable performance of the system was showed an it is represented by the performance of the classification model that achieved up to 88% of accuracy for SVMs. In addition, the scalability of the system system has been seen when the number of incoming tweets is large. Finally, the overall performance of execution has been explored. It can achieve a throughput of more than 6 million tweets per day with performance time of 12.462 ms for every single tweet.

Chapter 5

Entity Extraction and Event Recognition

Short messages posted on social web sites such as Twitter can typically reflect events as they happen. Therefore, the content of social web sites is particularly useful for timely, disease, or outbreak event identification, which is the problem that will be addressed in this chapter. Collecting and filtering methods are important to detect and track public health data such as disease outbreak or infection in social web. In previous chapters (Chapter 3 and Chapter 4), significant techniques were presented for tracking status updates of people on Twitter and identified medical cases of microblogs in real-time. The collected data should be understandable and valuable. Information Extraction (IE) is a task of natural language processing (NLP) that extracts useful structured information such as entities, relationships between entities, and attributes describing entities from unstructured text [50].

Named entity recognition (NER) is a very important subtask of IE that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, dates and times, and quantities [48]. However, NER has been extensively studied on formal text, such as the news. There has been much research in NER on news articles whereas on social web documents it is still in the preliminary stage. In fact, the task of NER and NLP in general become much harder, when the text is not written in standard

English (e.g., social web text). Thus, it is necessary to have a new system of NER that takes into account the nonstandard language used in the informal sources (Twitter data).

The task of NER with respect to social networks such as Twitter and Facebook is a new area of study for researchers. Currently, NER tools that are available on the web fail when applied on tweets [67] [3]. As shown by by the disease-reporting classification experiment in Section 3.3.4.2. The use of current NER techniques (such as OpenCalais and Part-of-Speech Tagger) do not improve the performance of the classification model. The reason being the differences between Twitter text and standard English text (e.g., news articles) because social posts are of limited length, contain incorrect grammar, and contain slangs.

In particular, disease-related posts may contain popular medical terms instead of correct medical terms. For instance, some users may write the term *stomach flu* instead of *gastroenteritis*. All of these factors make NER difficult to extract entities. In this chapter, an specific entity extraction process will be performed on disease-reporting postings that are relevant to disease surveillance. A named entity tool is built to understand the tweets meaning by analyzing their types and extract medical entities to construct events. All data used in this chapter are positive disease-related postings, which might be personal or general posts.

In summary, the contributions of this chapter are as follows:

- Distinguish between different types of disease-related postings (medical microblogs), and extract all medical entities from each post such as virus, bacteria, normal disease, or symptoms. Furthermore, other terms (e.g., infection, deaths, and fatality) that indicate an occurring outbreak or infection are extracted as well.

- Tracking the source of each post, where this post came from. This could be include the country, state, city, neighborhood, street, or GPS data.

- Creating disease events by capturing disease names, timestamps, location, and number of cases.

5.1 Motivation

5.1.1 Syntactic Processing (Parsing)

Syntactic analysis or parsing is the task of analyzing sentence structure and the dependencies between its parts. For instance, the grouping of words together as phrases, and if those words are the subject or object of a verb. A natural language parser is a program that performs the task of parsing. Usually, the input to a parser is morphologically disambiguated word-tag pairs (e.g., the input has been tagged). The output from the parser is a parse tree that reflexes a structural description of the sentences and tags denoting various dependencies.

5.1.1.1 Simple Rule Language (SRL)

SRL[1] is a parser developed by Google for extracting entities from the plain text based on regular expression that is manually defined for interest events. It is used in the BioCaster project[2] to identify disease outbreaks from news. Furthermore, it has ability to make domain experts developing their own text mining systems. The SRL editor provides many features to create trend named entity application. Before creating SRL rules the user should use a set of word lists that helps to organize specific categories of words comprising two types of rules, i.e., entity rules and template rules both of that are main extraction processes in SRL.

- Entity Rules

  Entity rules are used for detecting entities in the plain text by matching each token in the text with predefined word lists. The syntax of SRL named entity rule is given below:

  $$entityType(entityVal,var) \quad body$$

  The *entityVal* is the name of the entity that returned and bounded to the variable *var* when the body is a match. The body is several functions that are word-base regular expression matching, literals matching, prede-

[1] http://code.google.com/p/srl-editor/
[2] http://born.nii.ac.jp/

fined list of matching words, or word list that do not match. Examples of entity rules are as follows:

- *name(medviruses,V) list(@viruses)*
- *name(medbacteria,B) list(@bacteria)*
- *name(meddiseases,D) list(@diseases)*
- *name(medsymptoms,S) list(@symptoms)*

The rules are applied to the plain text to identify viruses, bacteria, normal diseases, and symptoms, respectively. In more detail, *@viruses* in the first rule are included in word list containing medical virus names, and the rule shows that the variable *V* could be *medviruses* if an entity in the text matches the term in the predefined list *@viruses*.

- Template Rules

  The named entity extraction in this thesis depends on template rules, which are applied after the entity rules are set to return a list of facts about the interest from the text. The template rule can combine more than one entity rule.

  head ":-" body

  The above syntax represents the general syntax of the SRL template rule. The *head* expression is represented as *id(var)*, which is a sequence of alphanumeric that should be output and the *var* variable should appear again in the body of the rule. A real example of template rule is given below:

  - *personalpost(X); pronoun(Y) :-name(pronouns,Y) "infected" "with" name(medviruses,X)*

  which indicates that the rule matches any entity returned by the entity rule *name(pron,Y)*, followed by *"infected with"* and then match any entity from the entity rule *name(medviruses,X)*.

5.1.2 Event Extraction

Text mining refers to the tasks that include text categorization, text clustering, entity extraction, sentiment analysis, and entity relation modeling (e.g., learning relations between named entities). By means of text mining, often using natural language processing (NLP) techniques, information is extracted from texts of various sources, such as news articles, and is represented and stored in a structured manner (e.g., databases). A particular type of information that can be extracted from text by means of text mining is an event, which ideally identifies who did what, when, and where. Automatically extracting events is a higher-level information extraction (IE) task, and relies on identifying named entities and relations holding among them.

Research in the field of event extraction has been an active area for the past ten years. Moreover, studies focused on the processing of human language in text form a variety of sources, such as news document, broadcast conversation, and weblogs, as part of a National Institute of Standards and Technology (NIST) initiative for automatic content extraction (ACE)[1]. The ACE event extraction task explicitly define a set of event types (e.g., conflict) and subtypes (e.g., attack) to be extracted from various text sources (e.g., newswire, blogs, and conversation transcripts), using a set of predefined templates that include event attributes (e.g., attacker and target).

Event extraction has been extensively applied within the medical domain [8], where event parsers are utilized for extracting medical or biological events, such as molecular events from corpora (e.g., clinical text), but did not consider the medical document of the social web (e.g., Twitter messages). Many event extraction systems have been reported. For example, systems capable of extracting disease outbreaks [49] and the BioNLP'09 shared task [30] focuses on the extraction of nested-bimolecular events.

Moreover, Lancet[2] is a supervised machine-learning system that automatically extracts medication events, comprising medication names and information pertaining to their prescribed use (dosage, mode, frequency, duration, and reason) from lists or narrative text in medical discharge summaries. Design Lancet in-

[1] http://www.itl.nist.gov/iad/mig/tests/ace/
[2] http://code.google.com/p/lancet/

corporates three supervised machine-learning models, i.e., a conditional random fields model for tagging individual medication names and associated fields, an AdaBoost model with decision stump algorithm for determining which medication names and fields belong to a single medication event, and a support vector machines disambiguation model for identifying the context style (narrative or list).

5.1.3 Mining Tweets in Literature

Several studies have demonstrated that information shared on Twitter has some essential value, for example facilitating predictions of box office success [6]. Recent work has leveraged the collective behavior of Twitter users to gain insight into a number of diverse phenomena. Analysis of tweet content has shown that some correlation exists between the global mood of users and important worldwide events [26], including stock market fluctuations [27]. Moreover, similar techniques have been applied to infer the relationships between media events, such as presidential debates, and affective responses among social media users [15]. Sakaki has successfully approximated the epicenter of earthquakes in Japan by treating Twitter users as a geographically-distributed sensor network [52].

5.1.4 Traditional NER Tools on Tweets

In this part, the performance of traditional NER tasks on tweets is described in brief. The current part-of-speech taggers trained on non-tweets perform poorly on tweets [67]. According to the experimental study of Ritter [3] that demonstrated that existing tools for POS tagging, chunking, and NER perform quite poorly when applied to tweets since the state-of-the-art Stanford POS tagger obtained an accuracy of 80% on tweets that is a huge drop from the 97% when applied on news, as well as the Stanford NER drops from 90.8% F1 to 45.88% when applied to a corpus of tweets [67]. There are several reasons for this drop in performance, due to unreliable capitalization; common nouns are often misclassified as proper nouns, and vice versa. Furthermore, interjections and verbs are frequently misclassified as nouns. In addition to differences in vocabulary, the grammar of tweets is quite different from edited news text. For instance, tweets

often start with a verb (where the subject 'I' is implied), as in the tweet *"having headache"*.

Examples of traditional NER tools that are available on the web and used by many studies are as follows: Stanford Named Entity Recognizer (NER)[1] and OpenCalais[2]. The Stanford NER labels sequences of words in a text that are the names of things, such as person, places, and company names, but it does not label the medical condition inside text. Conversely, OpenCalais is an API web service used to tag places, people, location, medical conditions, facts, and events in the content of the text. It can process up to 50,000 documents per day (blog posts, news stories, web pages) free of charge. In reality, the goal is to extract medical entity (such as name of virus, bacteria, normal disease, or symptom) from content of the tweet, which is a disease-reporting tweet. An example for the real tweet *"Both my parents have cancer :("* that is a short text but representing the real case of the cancer disease. However, this tweet has been tested by OpenCalais API, but there is no any result. Moreover, in the classification experiments (Section 3.3) the classifier has been examined on features extracted by OpenCalais and Stanford POS tagger, but the performance was not improved. Nevertheless, the important challenges were presented in Section 3.5.

The limitation of this section is that the tweet is very short in length, often containing grammatical errors and simple medical term(e.g., stomach hurt, stomach pain, stomach ache, head hurt, feet hurt, and tummy hurt) that are difficult to be tagged by standard NER techniques. As has been seen that the traditional NER techniques failed to extract the medical entity from the Twitter post. Therefore, to address those challenges, significant practical techniques are required to be able to extract the medical entity in each single disease-related post (tweet). The next sections describe the scenario of entity extraction and extarct event of a social posts.

[1]http://nlp.stanford.edu/software/CRF-NER.shtml
[2]http://www.opencalais.com/

5.2 Event Extraction on Tweets

An event is a unique thing that happens at some point in time. A disease infection outbreak at a particular place is an example of an event. Event extraction is a task that involves identifying instances of specific types of events and their associated entities.

This section describes the extraction of disease-related events from Twitter posts. However, a disease spread event is modeling as given below:

E=(medical entity, location, time period , posts)

where medical entity is the disease name, such as normal disease, virus, bacteria, or symptom. Location refers to the place where the disease occurs. Time period refers to time window, and posts are set of postings within the time period with restricted disease and location.

In fact, one cannot say from a single post that an event has occurred. In order to make the event detection reliable, each single general or personal post is defined as a micro-event as shown in the equation 5.1.

$$microE \stackrel{def}{=} (medicalEntity, location, time, type) \tag{5.1}$$

The difference between these events is that event E contains a collection of micro-events that occur in a time range with a specific disease and location as shown in the equation 5.2. Conversely, micro-event is only a single post that occurs at a particular time.

$$E \stackrel{def}{=} \sum_{i=1}^{n} microE_i \tag{5.2}$$

To extract an event on tweets, NER uses extract features of that event. Therefore, three questions are important in the entity extraction process: what, where, and when a disease incidence occurred. These questions will be answered for each post by three processes that are described later. The NER technique is useful to estimate the number of cases in each geographic location and decide if there is

an unusual event.

5.2.1 Medical-Entity Extraction on Tweets

5.2.1.1 Disease-Reporting Postings

From the experience in annotation, collecting, and filtering tweets in real-time, Twitter with public health information is considered as the following:

first a source of real-time information that is posted directly from the population (e.g., *"I'm very ill at the moment... I have Measles"*).

Secondly a distributor that means that Twitter users share or link near real-time information from another media (e.g., *"H1N1 outbreak kills 2 in Mexico: Official bit.ly/wWr3V3"*).

From these two dimensions, the distinguish between different types of the disease-related postings is taken into account.

- **General Epidemic Posts**

  General posts are posts that mention disease outbreaks somewhere on the earth. The general epidemic means where Twitter users do not post epidemic information about themselves, families, colleagues, friends, or someone around, but they post and link information that announces the occurrence of an epidemic as shown in the following examples:-

  1. *Measles outbreak on Merseyside: more than 200 cases http://t.co/3dnhcg0f,*
  2. *Four confirmed cases of dengue in Brazil.,*
  3. *18 suspected swine flu cases reported at Carnegie Mellon ...: Eighteen suspected cases of H1N1 influenza have pr..,*
  4. *23 horses on Maine farm die in rare botulism outbreak http://t.co/Pct8peRh,* and
  5. *2 Robertson County toddlers infected with E. coli outbreak http://t.co/kFfPGgmw*

As aforementioned, Twitter users are distributing web content to Twitter. Most of the examples(1-5) are from another social media sources such as news sites and some of them contain hyperlinks that indicate to the source of the information in other media on the web as in examples (1),(4), and (5). The important objective to distinguish these types of postings is the location entity that appears in the content of the post as in examples (1), (2), (3), and (5).

- **Personal Epidemic Posts**

  Personal epidemic posts express a personal epidemics for Twitter users since the users write health information about themselves, family, colleagues, friends, neighbors, or someone around.

  1. *I am a Streeper and I am infected with Strep Throat. i think i have a severe case of insomnia,*
  2. *I got breast cancer ugh man..,*
  3. *my sister has the norovirus I am overtired and rundown its only a matter of time fml.,* and
  4. *your kids have the plague too fool.*

  All of the examples listed above are express the health situation of Twitter users, and the syntax in example (1) is very close to the syntax of a general epidemic posting except that English pronouns have been used to distinguish these kinds of postings. Furthermore, in each personal post, the location of the Twitter user is considered as the location entity instead of extracting the location from content of the text.

- **Negation Epidemic Posts**

  Negation posts include all postings that mention that there is no outbreak in a specific location, or negate that the user does not has any disease or infection. Examples of negation epidemic posts are as follows:

  1. *there is no more virus infections in Haiti*,
  2. *I don't get a flu.*

These types of postings are distinguished because there are some negation posts that are difficult to filter out during collecting and filtering processes. Surely, there is not much lexical difference between the tweet *"I have measles"* that is a disease-related tweet and the tweet *"I don't have measles"* that is not related to disease monitoring. Therefore, it is difficult for machine learning algorithms to discard this confusion. Furthermore, adding additional process during collecting and filtering processes to check negation posts is time consuming especially in real-time.

The summary is that the separation between personal and general disease-related postings depends on the nature of the tweet content. Two reasons are making us do this separation:

first to determine the location entity of the epidemic that is extracting from text content or from location field of Twitter user for general and personal posts, respectively.

Secondly number of disease cases in personal postings are known in contrast to general epidemic postings where the cases are almost unknown.

5.2.2 Location-Entity Extraction on Tweets

The location entity is a very important metric to make disease-related post valuable. As it has been explained before, the location estimation for public health data in social networks depends on the type of post, that could be personal or general. For personal posts, the status updates of people are tracked and answer the question, where the post came from?. Conversely, estimation of location from the content of the post is used for general posts. Therefore, two techniques are required to estimate location of the Twitter user and location entity in the text.

5.2.2.1 Location of the User

The location field of user on Twitter is a free-text form. Most Twitter users write valid location information in their profiles, whereas some of them write invalid

location information or nothing. According to Hecht study, [7] 66% of Twitter users enter their correct locations, 16% enter invalid locations, and 18% do not enter anything. The valid location information could be street, neighborhood, famous place, city, state, or country. Most Twitter users specify their location at the city scale, and they may enter the city and the country name. Conversely, invalid location information is such as *Heaven*, *Hell*, and *behind you*. 3-G technique has enabled mobility devices such as the iPhone to make use of mobile devices to access and post social content. A Twitter user has the option to make his location available to the public through mobile devices. Therefore, GPS information (e.g., longitude and latitude) of the post will be available as meta-data. Judging from this brief description about user location on Twitter, a preprocessing operation is required in preparing the input via cleansing special symbols and slang. Then, the geocoding process is used to get information about textual location or GPS data.

Geocoding is a process to get geographic coordinates (latitude and longitude) from geographic data. Current geocoding techniques such as google[1], Unlock[2], and Yahoo! BOSS PlaceFinder[3] are available on the web, and they take textual location such as a street address as input and transform it into coordinates. In addition to the coordinates, geocoding returns more information about input places including city, state, country name, and the Where-on-Earth (WOEID) which is the identifier to determine each place in the world without repetition, for example *New York City* and *NYC* have only one WOEID. Conversely, reverse geocoding technique is a service that allows converting coordinates into place names. Actually, this process is used for tweets that have GPS data.

5.2.2.2 Location in the Tweet Text

This subsection describes the extraction of location entity from the content of the tweet. As has been described before, the current NER techniques that are available on the web do not perform well on tweets. Therefore, this work uses

[1]https://developers.google.com/maps/documentation/geocoding/
[2]http://unlock.edina.ac.uk/places/api/
[3]http://developer.yahoo.com/boss/geo/docs/free_YQL.html#table_pf

the geoparsing service, which is pure NER to extract only the entity of location from free text.

Geoparsing is a service that identifies, disambiguates, and extracts places from structured and unstructured text such as web pages, blogs, status updates, and other data sources. This service is used to extract the location from tweet text. Geoparsing also gets all geographic information for the extracted location. Examples of this service are Unlock text[1] and Yahoo! BOSS PlaceSpotter[2] that are available on the web.

5.2.3 The Scenario of NER

Entity and event extraction need a framework to analyze the semantic of a social post and extract medical and location entities from disease-related postings. However, Figure 5.1 shows the scenario of the framework of NER. Pre-process operation takes place before the processes have started and are represented by the cleansing process that is applied on each post to remove undesired tokens (e.g., numbers and emoticons). The NER scenario is represented by three processes that are performed to analyze the semantic of positive disease-related postings.

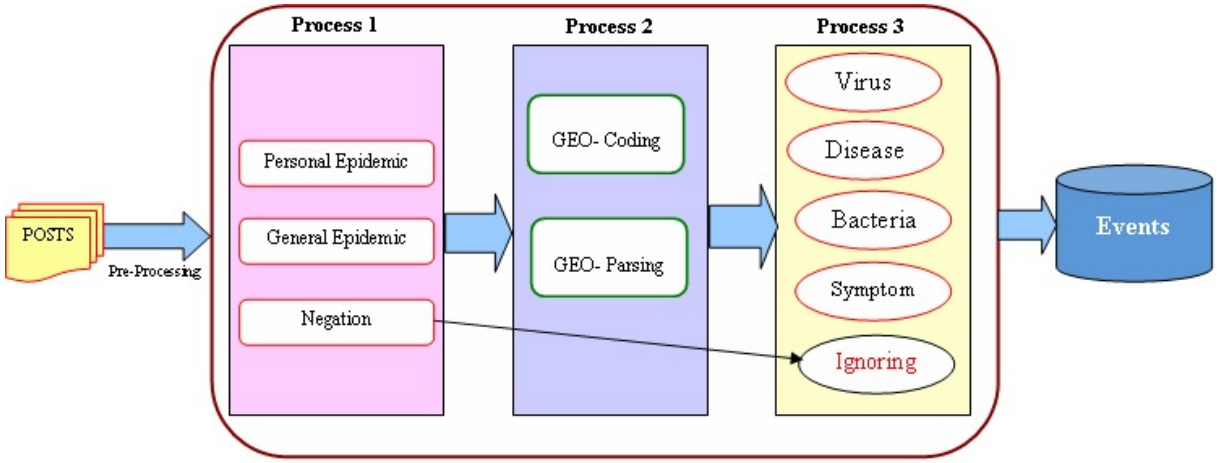

Figure 5.1: Name Entity Recognition Framework for Disease-related Postings.

[1] http://unlock.edina.ac.uk/texts/introduction
[2] http://developer.yahoo.com/boss/geo/docs/free_YQL.html#table_pm

First analyzing a piece of post text to determine the syntactic structure that could be general epidemic, personal epidemic, or negation post. The negation post is directly removed.

Secondly after determining the type of the post, the second process is responsible for getting location entity depending on the type of the post that is produced in the first process. To do so, geoparsing and geocoding are required, but only one is executed for each post. For general epidemic post, geoparsing is used to extract location entity from the text and to get related geographic information. Conversely, geocoding is used for personal posts to get geographic information for the content of location field in the user profile on Twitter.

Thirdly this process is responsible for extracting medical entity from the content of the post into correct classes such as disease, virus, bacteria, or symptom category. The symptoms will be extracted if the post does not contain any term from the other categories.

Finally, the outputs of the processes are the post's type, medical entity, timestamp, and geographic information. All these information are stored as an event in a database.

5.3 Experiments

This section presents first the data collection that is used for the experiment, describes the evaluation metrics, and then describes the methods. Furthermore, the results of the experiments are presented.

Three different sets of experiments are performed:

- Using the machine learning classifier of support vector machines (SVMs), and feature extractors such as unigram, bigram, and unigram and bigram to recognize types of disease-reporting postings.
- As well as using parsing rules (SRL rules) to recognize types of disease-reporting postings.

Category	Tweets
General Epidemic	35
Personal Epidemic	15
Negation	15
Bacteria	16
Disease	40
Virus	39
Symptom	40
Sum	**200**

Table 5.1: Experimental Annotated Dataset

- Extracting medical and location entities by using SRL rules, and geocoding and geoparsing techniques, respectively.

5.3.1 Data Selection

185 positive-disease-related tweets are randomly selected from the corpus (Section 3.2) by using the annotation tool that is described in Section 3.1.3. All of those tweets were distributed over many different types of diseases and outbreaks. In addition, 20 negation tweets are collected by using Topsy,[1] and randomly selected 15 tweets from them. The total number of collected data were 200 tweets that are shown in Table 5.1 as the personal tweets have very close syntax to the general epidemic tweets, but the tweets related to bacteria, disease, virus, and symptom are personal tweets. Conversely, the negation tweets contain personal and general postings. The criteria that have been used for labeling is as follows: the tweets are annotated into the three different types according to the description in the Section 5.2.1.1.

Moreover, each medical term inside the whole dataset (200 tweets) is annotated into related category (e.g., virus, bacteria, normal disease, symptom, and other), the medical terms that are known as "other" is used for post that does not contain a specific disease name (e.g., infection and death). This dataset does indeed seem small but it covers different syntactic structures of the tweets.

[1] http://topsy.com/tweets

5.3.2 Evaluation Metrics

this sub section briefly describes the scenario of evaluation. Three significant measurements that are used to evaluate the effectiveness of a NER program are given below.

1. **Precision(P)** is the proportion of the number of named entities that were identified correctly by the program to the total number of entities identified (the total number of the correctly identified and incorrectly identified named entities).
$$P = \frac{Ncorrect}{Nresponse} \tag{5.3}$$

2. **Recall(R)** is the proportion of named entities that were identified correctly to the total number of named entities that were assigned manually(Nkey).
$$R = \frac{Ncorrect}{Nkey} \tag{5.4}$$

3. **F-Measure(F)** is a measurement involving both P and R as combined in the following equation.
$$F = \frac{2RP}{R+P} \tag{5.5}$$

5.3.3 Methods

5.3.3.1 Support Vector Machines with N-gram

SVMs algorithm and the feature extractors(e.g., unigram, bigram, and unigram and bigram) are used to distinguish between different types of disease-related postings.

5.3.3.2 Parsing Rules

SRL rules are built to detect the correct medical terms in the dataset, but before that word lists are created as categories that contain medical terms(e.g., virus, normal disease, bacterium, symptom and others). Wikipedia is used to distinguish between all categories, and consider the health behavior of the user (e.g.,

feel sick and *have pain*) as symptoms. Then, SRL entity and template rules are constructed to distinguish between seven different categories.

1. **Personal Epidemic Rules** are constructed to distinguish the personal epidemic tweets that have high similarity in their syntax structure with the general epidemic tweets. An example of these rules is given below:

   - *persepidemic(X); pronoun(Y) :- name(pron,Y) "have" "infected" "with" words(,3) name(medcond,X)*

   since the parts *name(pron,Y)* and *name(medcond,X)* are SRL entity rules that check pronouns and medical condition, respectively, and the medical condition could be any term in the word lists. Furthermore, *"have"* and *"infected"* are literals between the pronoun and the medical term in the text as well as the part *words(,3)* that represents any text not more than three words. In fact, the pronouns are represented the important metric to distinguish the personal posts.

2. **General Epidemic Rules** are built to distinguish the general epidemic tweets from other tweets. Examples of these rules are as follows:

   - *epidemics(X) :- words(,4) "infected" "with" name(medcond,X)*
   - *epidemic(X) :- "cases" "of" name(medcond,X)*

   These rules do not contains any entity rules about pronouns. Considerably good rules have been constructed for general outbreak by checking all the syntax of general epidemic tweets in the corpus and also used Google news and some disease outbreak reports to know the syntax of outbreak titles.

3. **Negation Rules** are used to determine all negation tweets. Examples of the rules are given below:

   - *negation(X) :- "don't" "have" name(medcond,X)*
   - *negation(X) :- "no" name(medcond,X) "infection"*

   these types of rules are constructed for both personal and general epidemic negation tweets.

4. **Virus Rules** extract name of viruses from a tweet.
   - *virus(V) :- name(medviruses,V)*
5. **Disease Rules** are used to extract disease names. An example of this rule is given below:
   - *disease(X) :- name(meddiseases,X).*
6. **Bacterium Rules** extract all bacterium from the tweet's text.
   - *bacteria(B) :- name(medbacterum,B).*
7. **Symptom Rules** are used to get names of symptoms in the post and are applied when there is no specific medical term in that post.
   - *symptom(S) :- name(medsymptoms,S)*

Rules (1), (2), and (3) are used to determine the type of the tweet. Conversely, (4),(5), (6), and (7) are used to extract medical entities from the content of the tweet. Also the rules can handle the tweets that contain multi medical terms that belong to different categories as the SRL rules extract all cases in the text. Some rules are also constructed to detect the entity in the tweet that does not contain specific disease name (e.g., *"His wife is died. too sad :("*). Moreover, all negation posts do not considered as a cases but automatically removed.

5.3.3.3 Identifying Location

The extraction of location entity from Twitter data depends on the type of postings. The content of the tweet is used to extract location entity for the general epidemic posting. Conversely, the location field of the Twitter user has been used as a location entity for personal posting. The methodologies to extract geographic location information are described as follows: all country names are stored with their geographic information such as latitude and longitude in a relational database. During the processes of parsing rules, Yahoo! PlaceSpotter[1] is

[1]http://developer.yahoo.com/boss/geo/docs/free_YQL.html#table_pm

used as a geoparser and Yahoo! Placefinder[1] as a geocoder to extract geographic location information for general and personal postings, respectively. Then, the geographic location information is stored for the related tweet as a city-level, which is the most entered location by Twitter users according to Hecht study [7]. If the location information that returned is only a country name, then data as a country-level is stored. Some personal tweets have geographic information (GPS) as meta-data, which are provided by the Twitter API. Therefore, before starting the geocoding process it was checked if there is geographic information associated and ignoring information in location field of Twitter user. Furthermore, the geoparsing process returned location entity in the tweet text, after that geocoding for those entities is used to get geographic information, such as latitude, longitude, and country code.

5.3.4 Experimental Results

This subsection begins by examining the performance of the SVMs technique, and then shows the performance of the parsing rules (SRL rules) on the collected dataset to determine the types of posts. The performance of extracting of medical entity is described as well.

5.3.4.1 Using SVMs Algorithm with N-gram Features

Test runs used 5-fold cross validation with SVMs algorithm on unigram, bigram, and unigram and bigram features to classify tweets types. Table 5.2 shows the accuracy, precision, recall, and F-Measure rates.

Feature	Accuracy	Precision	Recall	F-Measure
Unigram	92.31%	92.3%	92.3%	92.1%
Bigram	75.38%	83.1%	75.4%	73.1%
Unigram and Bigram	92.31%	93.3%	92.3%	92.1%

Table 5.2: Evaluation rates for experiment dataset using SVMs.

- **Unigram** The unigram feature extractor is the simplest way to retrieve features from a tweet. The algorithm performs with an accuracy of 92.31%.

[1] http://developer.yahoo.com/boss/geo/docs/free_YQL.html#table_pf

- **Bigram** The bigram feature extractor is useful to distinguish negation phrases in the tweet such as *"don't have"* or *"not infected"*. In the experiment, the precision for negation tweets was 100%, but the bigram does not improve the accuracy for all dataset. However, using bigrams as features is not useful because the shared terms between tweets.

- **Unigram and Bigram** Both unigrams and bigrams are used as features. As compared with unigram features, only the precision is improved from 92.3% to 93.3%.

5.3.4.2 Using SRL rules

164 SRL rules were applied, manually developed to separate tweet types and extract medical entities from the dataset of the experiment. In fact, the rules extract all the strings that match patterns in the word lists, previously defined. The performance is broken down into two divisions:

- *First*, the performance of the system when using SRL rules to distinguish between tweet types. Table 5.3 shows the precision, recall, and f-Measure rates for personal, general-epidemic, and negation tweets. The denominators in recall and precision denote the number of correct posts detected by SRL rules and the number of terms manually annotated in the dataset. The recall performance for detecting personal epidemic tweets was 80%, which is a small rate in the results since two tweets were detected incorrectly because of person entity. One of those tweets is *"Justin is infected with HIV!"*. Epidemic tweet that related to a person was annotated as a personal and the person entity in this study was not extracted. Therefore, SRL considers these types of tweets as general epidemic. Moreover, in the NER framework, if the SRL rules did not determine the type of the tweet, it was automatically considered as a personal tweet.

- *Secondly*, the performance for extracting medical entity from the dataset excluded the negation tweets. Table 5.4 shows the recall and F-Measure rates, and the numerators in the parenthesis are the term numbers that extracted correctly. According to the nature of SRL rules with medical

Category	Precision	Recall	F-Measure
Personal Epidemic	80%(12/15)	80%(12/15)	80%
General Epidemic	100%(30/30)	85.71%(30/35)	92.31%
Negation	86.67%(13/15)	86.67%(13/15)	86.67%
Average	**91.67%(55/60)**	**84.61%(55/65)**	**86.33%**

Table 5.3: Evaluation rates for experiment dataset using SRL.

Entity	Recall	F-Measure
Bacteria	89.29(25/28)	94.34%
Disease	95.45%(63/66)	97.67%
Virus	98.99%(98/99)	99.49%
Symptom	88.24%(45/51)	93.75%
Average	**95.45%(231/242)**	**96.31%**

Table 5.4: Evaluation rates for Entity Categories.

name categories, there is no incorrectly identified named entity. Therefore, the precision rate was 100% for all categories. The average recall rate on the other hand was 95.45%.

5.3.5 Summarization

It was found that the performance of SRL rules do not like SVM performance but using SRL rules is more practical, fast, and do not need a training; So they can be applied to tweets where training data is not available as in corpus-based approaches, where if there is a disease name re-designated into another name (e.g., Swine flu re-designated into H1N1) the retraining is needed. In addition, SRL rules are used to extract medical terms, recognize slang terms (e.g., shot, vaccine, drug, jab, vacc, doctor, doc, and dr), and mitigate sparseness problem (e.g., *"I got a fluuuuu"*). Besides, SRL is easily working with many different languages, and it can be updated when needed.

5.4 Discussion

This chapter presented the semantic meaning of disease-related postings by defining a specific event with named disease, location, and time. Several points that

have been discussed in this chapter are highlighted.

- **Posting Type** during collection and filtering of disease-related postings, it was noted that there are many different types of posts that should be distinguished to track the location entity and determine the event type. In this chapter, two different methods were used: machine learning and named entity parser to distinguish between three different types of postings, i.e., general, personal, and negation postings. It was found that parser rules are more practical to deal with tweets because of slang and sparsity of data.

- **Entity and Event Extraction** SRL parsing rules are constructed to extract the medical entity from each post such as virus, bacteria, normal disease, and symptom. In the experiment, good results are obtained for extracting entity. Conversely, location entity extraction was depending on the type of post. Geocoding techniques have been used on the user's location of personal post to get related geoinformation. Geoparsing has been used to extract location entity from a text for general epidemic posts. Furthermore, the named entities are transformed to construct the events.

  Besides, in the following sections, potential determination of popular medical entity and event summarization will be discussed.

5.4.1 Is it Influenza or a Stomach Flu

Flu-related tweets were collected for four months from December, 2011 to March, 2012. After that, the filtering techniques in Chapter 4 were applied to get pure positive flu-related tweets which are tweets of positive cases of flu, influenza, or H1N1 virus. It was found that 23% of the tweets are positive flu-related tweets. Only tweets that were posted from USA were extracted, and these tweets are accounting to 17% of all related tweets. Then, NER techniques that presented in this chapter have been applied to detect personal tweets and detect entity names. The result was that 99% of the tweets were personal. Some Twitter users write about influenza as a flu (e.g., *"i got the flu :(been sneezing all day stuffy and eyes running again"*) that is a viral infection. Conversely, stomach flu is written by some users, which is not the true medical term but a popular medical term for

gastroenteritis. An example of a stomach flu tweet is *"My baby sick with stomach flu :("*. The NER technique has the ability to distinguish between flu (influenza) and stomach flu, as well the ability to extract different types of human pain (e.g., knee pain, teeth pain, and tummy ache). In addition, it was found that 13% of the positive flu tweets that were posted from USA are related to stomach flu.

5.4.2 Event Summarization

In practice, the NER framework was utilized to detect disease-related events. Features such as medical term, location, and timestamps are extracted and stored as a micro-event with the type of post that might be personal or general. This operation is very helpful in getting a number of symptom or disease cases in specific location, and depends only on personal postings that are posted by Twitter users. These cases can be effectively used as an input to the spatio-temporal model for prediction of the disease spread. Therefore, disease statistics in social networks are supposed to be cheaper and faster than traditional statistics, but still the challenge that whether each patient posts about his/her health or not. The alerting event on the other hand, depends on the micro-event numbers of personal or general postings. In fact, the alerting event is determined by the threshold according to the location. For example, the number of HIV cases that are needed to alert in Germany should be different from those in Nigeria.

Moreover, the collecting and filtering system (Chapter 4) has been used for collecting disease-related tweets of April 2012. 114 Gigabytes of data that contains 46,360,000 tweets were collected (Figure 2.8); 20.21% of those tweets are automatically filtered as positive disease-related tweets, including tweets mentioning that the user is sick or has pain(e.g., *"I feel sick to my stomach"* , and *"i just got a sharp pain in my ear, it really hurts"*).

The NER framework has been executed on 700,000 positive disease-related tweets and the statistic results are shown in Table 5.5. More than 80% of the tweets were observed to be about symptoms and 58.36% of them mention that the users are only sick since twitter users write about their symptoms more than any other diseases. According to this data, it was found that 65.05% of the tweets have a valid geographic information. The results are extremely close to Hecht study [7]

Category	Total	Personal	General	with Loc	without Loc
Virus	2.78%	95.51%	4.49%	67.83%	32.17%
Bacteria	0.79%	98.77%	1.22%	64.73%	35.27%
Normal Disease	7.50%	98.54%	1.46%	67.89%	32.11%
Symptom	80.24%	100%	-	64.34%	35.65%
Others	7.63%	100%	-	68.65%	31.35%
Average	**98.95%**	**99.75%**	**0.25%**	**65.05%**	**34.95%**

Table 5.5: Disease statistics for 700,000 tweets posted in April 2012.

that showed that 34% of Twitter users did not provide real location information. From the results in Table 5.5, More notes about the nature of tweets' syntax are gained and they were very useful to add or modify the SRL rules for improving the framework.

The negation tweets and the tweets containing noise were 1.05% of all data and have been automatically removed. Tweets that did not contain any disease names accounted up to 7.63%, but they were about terms such as deaths, fatalities, and unknown infection. In a frequency analysis of the trend cases, The flu cases were observed to represent the top cases in a virus category with 56.23% of all virus cases. Plague and strep throat were the top cases in the bacteria category with 35.81% and 19.28%, respectively. Cancer and diabetes were the top cases for the normal disease category with 39.87% and 10.57%, respectively. Conversely, the symptom category has the largest number of cases as compared with other groups; 26.47% of symptom-related tweets mention that the users have headache that represented the second top cases after the cases that mention that the user is only sick. Furthermore, it was noted that the three top countries that contain the top cases in all categories are USA, UK, and Canada respectively.

5.5 Conclusions

This chapter focused on NER on disease-related postings. The chapter presented the challenges when using the traditional NER tools for tweets and described the approaches for extracting entity and event. The postings have been classified into three categories personal, general, and negation. The medical entity, such as virus, bacteria, normal disease, or symptom, from each post has been then

extracted. The results were good enough for the NER framework to work in practice. Conversely, location entity extraction was dependent on the type of the post.

Chapter 6

Performance Evaluation of Whole System

The whole system that is described in this dissertation has been developed for tracking diseases by monitoring Twitter data. The first and the second stages in it are microblog annotation and classification, the classifier is working in real-time with an input Twitter stream, and the third stage is named entity and event recognition. Therefore, the evaluation of the whole system is represented by the evaluation of accuracy of the micro-message classifier and its run-time performance, as well as the performance of the entity identification system. Furthermore, in order to provide proof of the system, the system's output are compared to national health statistics. The results include quantitative correlations with government data, as well as qualitative evaluations of the system's output.

In summary, the contributions of this chapter are as follows:

- Evaluating the classifier model that distinguishes between positive disease-related positing and those that nonrelated by using a dataset annotated manually (Section 6.1).
- Comparing results of the system to official government data (Section 6.3).
- Exploring whether the system can predict values of the disease rate before reported by official sources (Section 6.4). Also the alerting methods are described in Section 6.5.

Some parts of this chapter appeared in [53], [54], and [55].

6.1 Classifier Performance

A machine learning classifier has been built to distinguish between postings that are relevant to diseases and those that are irrelevant (Chapter 3). Well established guidelines have been adopted for defining clinical cases, as well as an empirical experiment implemented on 200 tweets that were annotated by three subjects who showed a good agreement according to Kappa performance. However, Figure[1] 6.1 shows how the quality and correctness of the classification model is determined by testing the creating model on testing data via confusion matrix.

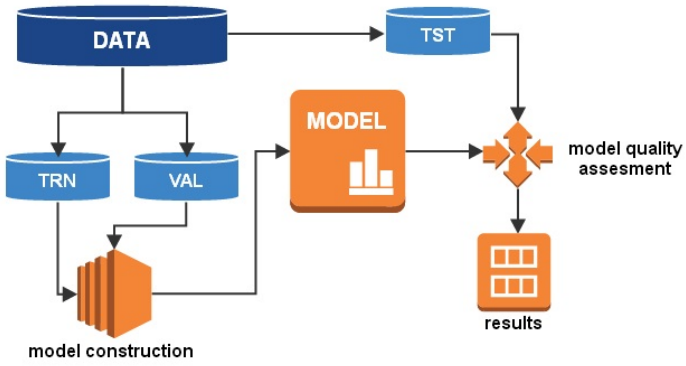

Figure 6.1: Model Quality Assessment

The classifier is evaluated in two different ways:

- **Using 10-fold Cross-Validation**

  The annotated dataset collected by the annotators comprises 5880 tweets, 2130 positive ones and 3750 negative ones. Then, this dataset was used as training data to build the classifier by using two different machine learning algorithms: Naive Bayes and SVMs. The accuracy evaluation of the classifier is conducted on that data by using 10-fold cross-validation and the classifier achieved up to 89% of accuracy.

[1] http://www.datasciencecentral.com/

Moreover, in Section 4.1.7 a new dataset was annotated in order to make the training dataset cover large number of medical conditions, which contain tweets covering symptoms and infectious disease outbreaks (see Table 5.1). This new dataset is added into the corpus. The final dataset contains 2380 disease-related tweets and 3795 nonrelated tweets, and was used to train the final classifier that achieved up to 88% of accuracy by using SVMs as the training model with 10-fold cross-validation.

- **Using a Dataset Annotated Manually**

  In addition to cross-validation, 200 tweets are collected by the annotation tool, that was developed in Section 3.1.3. The dataset has been annotated according to the annotation guidelines described in Section 3.1. All of those tweets in the dataset belong to the same domain of public health with respect to various names of symptoms and diseases. The result of manual annotation was as follows:

  - 111 tweets manually annotated as disease-related (positive), and
  - 89 tweets annotated as nondisease-related.

  Both tweets, i.e., related and nonrelated contain ambiguous tweets that may lead to false decision through the classification model. In fact, vague tweets are collected to see the performance achieved by the classification model. The following examples are some real tweets, which are collected to test the performance of the classifier that may get confused while making a decision.

  1. *Fever :(,*
  2. *headache ! Need aspirin,*
  3. *I hate having asthma,*
  4. *i have NEVER suffered from influenza or any viral infections, fungal infections etc*
  5. *Oh yeah , my brother didn't go to work today cause he has diarrhea and vomiting too !!*

The implied meaning of the tweets 1 and 2 is that the user has fever in (1) and headache in (2). Thus, they are manually annotated as positive but the classification model assigned them as negative because the model understands only the term frequency. The tweet 3 is very ambiguous; therefore, it is manually assigned as disease-related only for testing. The classification model assigned it as disease-related as well. The tweet 4 is a negation tweet that is manually annotated as negative. At times, the model has confusion for taking a decision on a negation sentence. For example, the tweet (4) has the terms *have, suffered, influenza, viral, and fungal* that refer to positive and the only one term that refers to negation is *NEVER*. Therefore, this tweet is incorrectly assigned by the model as disease-related. Furthermore, the tweet (5) is correctly assigned as positive disease-related. A confusion matrix is used to measure the performance of the classifier model. The result were as follows:

- For the positive tweets, 99 tweets were assigned correctly by the model and 12 tweets were assigned incorrectly.
- Conversely, for negative tweets, 72 tweets were assigned correctly by the model and 17 tweets were assigned incorrectly.

Therefore, the accuracy for classification model was 85.5% for the crucial dataset that contains ambiguous tweets. Finally, this dataset is added into the whole training set, and the tweets containing one word or negation have been excluded.

The limitation here is that this classification is a hard classification task in comparison with the web search domain disambiguation. The reason is that, all these tweets came from the same domain (Health domain), are short in length, and share a lot of common medical terms regardless whether relevant or nonrelevant into disease surveillance. This means that evaluation data are much less discriminative. Indeed, there is not much lexical difference between the tweet *"I have the plague"* that is annotated as a related tweet, and the tweet *"I do not have the plague"* that is annotated as a nonrelated tweet. It is difficult for machine learning algorithms to dis-

card this confusion. Regardless of the confusion tweets for the model, good results have been obtained for the performance of the classification model. Regardless of that challenge, the entity recognition technique is used to remove negative tweets.

6.2 Run-time Performance

The run-time performance was represented by the performance of the real-time filtering system, which is supposed to work with Twitter stream overtime. However, the latency and throughput were two important metrics to evaluate the run-time performance (more details in Section 4.4.2). According to the characteristics of the machine that the system was implemented and evaluated on, it was found that each tweet takes on average 12.46 ms to process from start to finish. In addition, the system can process more than 6 million tweets per day.

6.3 Comparison to Gold Standard Data

The goal of this dissertation is the using of sophisticated techniques to track human diseases via the social web. A real-time short-text mining system has been developed to detect and track diseases using Twitter messages. The system used the state-of-the-art text classification to filter postings into disease related or nonrelated, the real-time tracking of peoples' status updates, as well as an evaluated entity extraction to make them understandable and valuable. In order to provide proof of the system, the system's output are compared to national health statistics. Many different disease cases were detected by the system but the rate of Influenza cases that were reported by Twitter were chosen to compare with ILI(Influenza-like illness) rate that reported from CDC(Centers for Disease Control and Prevention).

6.3.1 Gold Standard and Twitter Data

The CDC (CDC[1]) publishes national and regional ILI rates based on weekly reports from all states in USA. The rate value reported by the CDC is the percentage of positive cases for all weekly tested cases. In more details, the CDC produces FluView[2] that is a weekly influenza surveillance report because the CDC collects information from eight different data sources. The U.S. Influenza surveillance system is a collaborative effort between CDC and its many partners in USA, local and territorial health departments, public health and clinical laboratories, vital statistics offices, health-care providers, clinics, and emergency departments.

In this dissertation, the ILI rate for the USA is considered as the gold standard data. The system is run for 17 weeks on Twitter data that was collected between December, 2011 and March, 2012 (Figure 6.2). All tweets in the dataset mentioned about flu, influenza, and H1N1, and this dataset is named with flu-related data. The filtering system (Chapter 4) has been applied to automatically remove all retweets, the tweets related to flu vaccine and flu shot, tweets that share hyperlinks, and tweets posted more than once from the same user in the same week. In addition, the filtering system filters tweets into positive or negative tweets.

Then, by applying the NER system (Chapter 5), the system extracted all type of flu cases (such as personal and general) from positive tweets that were posted only from USA because flu outbreaks in the United States usually occur from December to April. The percentage of personal positive cases of flu is used for each week (by using the total number of tweets in that week) as a Twitter data for the comparison with the ILI rate of CDC (gold standard).

6.3.2 Evaluation measures

Pearson product-moment correlation coefficient[3] is used to measure the correlation between the number of flu cases produced by the system that is described in this dissertation and the CDC data on ILI rates (gold standard). The Pearson

[1]http://www.cdc.gov
[2]http://www.cdc.gov/flu/weekly/
[3]http://en.wikipedia.org/wiki/Pearson_productmoment_correlation_coefficient

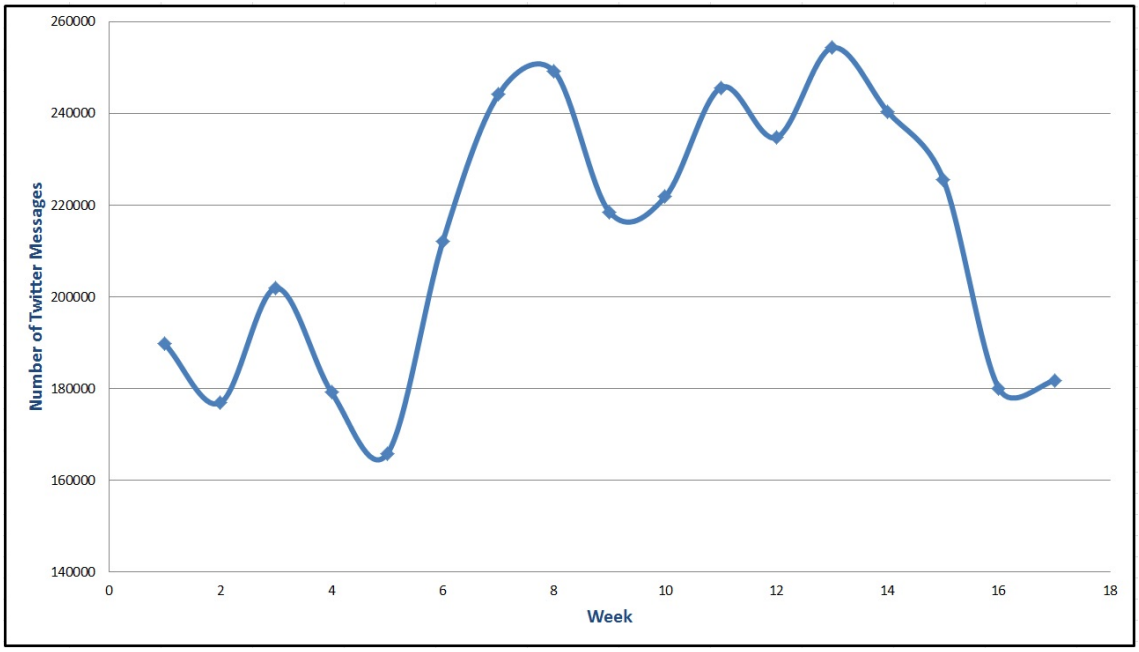

Figure 6.2: Number of Twitter messages per week for 17 weeks (week 1 starts on December 5th, 2011, week 17 ends on March 31, 2012)

correlation coefficient between two variables X and Y is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2(Y_i - \bar{Y})^2} \tag{6.1}$$

In the experiments in this work, X represented the ILI rate from CDC data (the percentage of positive cases for all tested cases each week), and Y represented the percentage of personal positive tweets of flu each week (by using the total number of tweets in that week) that were filtered by the system. Because CDC data are reported weekly (starting on Sunday and ending on Saturday), in the same order of the CDC, Twitter messages are arranged. Moreover, X_i and Y_i were the number for the ILI rate and the number of filtered Twitter messages for the week number i (i = 1, ...,17), respectively. The $\bar{X}$ and $\bar{Y}$ represent the mean for each variable. The value of r represents the linear relationship between the variables X and Y, and it ranges from -1 to +1. It is +1 or -1 in the case when two variables are perfectly correlated or anti-correlated, respectively. If it is 0, there is no linear correlation between the variables and the high correlation

is from 0.5 to 1.0 or -0.5 to -1.0.

6.3.3 Results

The aim here is to measure the correlation between real-personal-positive cases of flu, influenza, and H1N1 that produced from the system, and gold standard data. A free statistics and forecasting software[1] has been applied in this experiment. The result of a Pearson correlation coefficient was 82%, which was statistically significant ($P-value < 0.05$). All statistics results are shown in the table 6.1, as well as the correlation was between flu cases reported by CDC and real personal cases filtered by the text-mining system.

Pearson Product Moment Correlation		
Statistic	**CDC - X**	**Twitter - Y**
Mean	11.7529411764706	10.4739242355882
Biased Variance	69.6060207612457	2.07332368363962
Biased Standard Deviation	8.34302227980039	1.43990405362289
Correlation(r)	0.817228852319202	
p-value	6.19760259863433e-05	
Degrees of Freedom	15	
Number of Observations	17 Weeks	

Table 6.1: Statistical Results of the Correlation

Comparing the strong correlation result with related work: Culotta [12] and Lampos [37] used respectively simple and complex methodologies of regression to fit data from the government and focused only on flu tweets. The key idea in these two studies was to choose keywords to filter and aggregate influenza-related messages. Culotta's approach filtered tweets using four keywords ("flu", "headache", "sore throat", and "cough"). Thus, the syndrome cases have been considered. Moreover, Lampos's approach selected 73 keywords to filter Twitter messages and compared it with national health statistics in the UK. Both Culotta and Lampos used keyword-based methods in their correlation. However, many ambiguous cases have been not considered. Conversely, several ambiguous cases have been filtered out via the novel methods in this work. For example, tweets

[1] http://www.wessa.net

about flu vaccination, shared information (URL or Retweets), negation (*e.g, I don't have the flu, I just slept wrong xD*), stomach flu, or tweets about flu related to opinions or wondering (*e.g, What if I get swine flu?*). In fact, social web data is vast, noisy, distributed, and unstructured, which poses novel challenges for data mining. Therefore, novel filtering methods have been developed for disease-related messages to get quality data. The correlation in this work only focused on self-reporting tweets (personal messages) that are real cases of influenza.

Furthermore, the system in this work is a text mining system that uses the state-of-the-art text classification (Support Vector Machines). This is in order to filter not only a single ailment, but also all disease-related tweets, using entity and event extraction system to discover many different pure cases such as normal diseases, viruses, bacteria and symptoms, as well as geographical locations with their latitudes and longitudes. Also the dataset that was used for correlation was very small compared with other studies, but it represents real personal flu cases. In fact, it is difficult to get the full Twitter data since the amount of data that returned by Twitter streaming API for specific keywords is limited.

One of Twitter's limitations is that the amount of data returned by Twitter streaming API for specific keywords (e.g., flu) is less than 1% of the total fire-hose volume (all tweets happening on Twitter at the moment of streaming). It is very expensive to access to all the volume of fire-hose but Twitter returns rate limited messages by telling how many tweets are missed.

6.4 Forecasting of Disease Rates

Another validation is that, the system can be evaluated in a forecast setting, by testing how well the text-based system can predict values of the disease rates. In this section, the system was tested to predict values of ILI rates before they are reported by the CDC. In fact, ILI values that were reported by the CDC are only available for one to two weeks after a patient is diagnosed, because the CDC collects information from many different official sources. For example, data for patients seen in week 9 in 2013 (February 24 - March 2) was published on March 17. Section 6.3 showed the strong correlation between ILI values reported from the CDC and the Twitter messages. Furthermore, the system (Chapter 4) was

designed to work in real-time with the Twitter stream and each tweet needs 12.46 ms to process from start to finish. In addition, the framework for disease entity extraction (Chapter 5) can be integrated to work in real-time with the filtering system. Therefore, Twitter can be used to estimate ILI rates in the real-time before they are reported from official sources more than a week.

To summarize, Twitter can be used to find early cases of an outbreak before reporting by traditional systems. It was also found that the automatic prediction with machines has a much lower cost [17], as well as they can be useful for public health events. It can process greater amounts of data and provide responses quickly when unusual events occur such that the volume of Twitter messages posted for the Influenza in the time period exceeds some expected level of activity. Most importantly, it was proved that the social web can be used as an important resource in surveillance systems for tracking infectious disease outbreaks and symptoms, calculating the disease rate (e.g., ILI rate), showing that public health information can be extracted from Twitter, and demonstrating that because of real-time data filtering there is an opportunity for low cost time-sensitive sources to be exploited to supplement existing traditional surveillance systems.

6.5 Alerting

The previous section focused on forecasting disease rates in real-time before it is reported from official sources via monitoring disease cases of Twitter messages. But with increasing medical cases, the automatic detection of the unusual(unexpected number of cases for a given place and time) is required. However, aberration detection algorithms (alerting methods) [28] are designed to look for alerts when the volume of disease cases is increasing. The detection algorithms use the selected statistical method(s) on all cases and send notification (e.g., emails) to whomsoever it may concern if any alerts are raised. For example, the Early Aberration Reporting System (EARS)[1] methods developed by the CDC are used by many local health departments across the United States and have been widely used by public health officials for traditional disease surveillance. All

[1] http://www.bt.cdc.gov/surveillance/ears/

methods used a history window of length seven days. As described in Fricker et al. [20], the current EARS' detection methods are known as "C1," "C2," and "C3." The C is likely an abbreviation for the cumulative sum (CUSUM) methodology from that the EARS documentation claims these methods were derived. The comparison between the three methods represented by the sensitivity that is the probability of detection an outbreak signal since the C1 method has the lowest sensitivity and hence is known as C1-MILD, the C2 method (C2-MEDIUM) is more likely to flag high consecutive values, and the C3 method(C3-ULTRA) is considered to have the highest sensitivity of the three methods. Details of EARS (C1, C2, and C3) algorithms used in this section are provided.

- The C1 method calculates a standardized cases daily count for day t using the sample average and sample standard deviation estimated from the previous seven days of daily counts,

$$C_1(t) = \frac{Y(t) - \bar{Y}_1(t)}{S_1(t)} \tag{6.2}$$

  Where: $Y(t)$ is the observed cases count for day t, $\bar{Y}_1(t)$ is the sample mean based on the previous seven days of data, $\bar{Y}_1(t) = \frac{1}{7}\sum_{j=t-1}^{t-7} Y(j)$, and $S_1(t)$ is the sample standard deviation based on the previous seven days of data, $S_1(t) = \frac{1}{6}\sum_{j=t-1}^{t-7}[Y(j) - \bar{Y}_1(j)]^2$.

- The C2 method is very similar to C1 but calculated using a seven day sliding window baseline with a two day guard on the target day t being assessed. Specifically,

$$C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{S_3(t)} \tag{6.3}$$

  Where, $Y(t)$ is the observed cases count for period t, $\bar{Y}_3(t)$) is the moving sample mean, $\bar{Y}_3(t) = \frac{1}{7}\sum_{j=t-3}^{t-9} Y(j)$, and $S_3(t)$ is the moving sample standard deviation, $S_3(t) = \frac{1}{6}\sum_{j=t-3}^{t-9}[Y(j) - \bar{Y}_3(j)]^2$.

- The C3 method combines current and historical data from day t and the previous two days, and it calculates the statistic at time t in a similar

manner to C2 as follows:

$$C_3(t) = \sum_{j=t}^{t-2} max[0, C_2(j) - 1] \tag{6.4}$$

The C1, C2, and C3 methods signal an alarm at time t when their statistics exceed a fixed threshold h, which occurs when $C_1(t) > 3$, $C_2(t) > 3$, or $C_3(t) > 2$, respectively as well as the threshold may change according to the location.

In order to detect unexpected rises in the volume of Twitter messages for each disease, the three methods have been implemented on positive cases of influenza that were detected by the text mining system when applied on dataset of 121 days that presented in the Figure 6.2.

Figure 6.3 shows all alerting flags in two different weeks. C3 alerted on December (29 and 30, 2011), and all three methods alerted on January 19, 2012, when the high volume of cases took a place. Because of wide variability in regional level data, the EARS methods are currently used by states, counties, and cities[1], and focus on applying alerting methods on city-level, produces a good detection of sensitivity[2]. But the alerting methods were applied on all Influenza cases of Twitter messages posted from all cities in USA with default thresholds. Therefore, applying the alerting methods on country-level could produce the false alarm, as well as the epidemic threshold is determined according to each disease and the population in a specific location[3]. For example, for the areas where the disease epidemic is high (e.g., HIV/AIDS in Africa), the threshold should be different to areas where the epidemic is more low (e.g., Europe).

In fact, it was difficult to find published studies that have systematically compared detection methods using real syndromic surveillance data. This lack of comparisons on actual syndromic surveillance data makes it difficult to select aberration detection methods objectively. Multiple algorithms have seldom been compared on the same data [43], which is problematic because algorithms that work well for one data source may not do as well for another. In CDC, the data

[1]http://www.bt.cdc.gov/surveillance/ears/

[2]http://www.cdc.gov/mmwR/preview/mmwrhtml/su5301a16.htm

[3]http://www.who.int/infectious-disease-news/IDdocs/whocds200527/whocds200527chapters/4_Outbreak_control.pdf

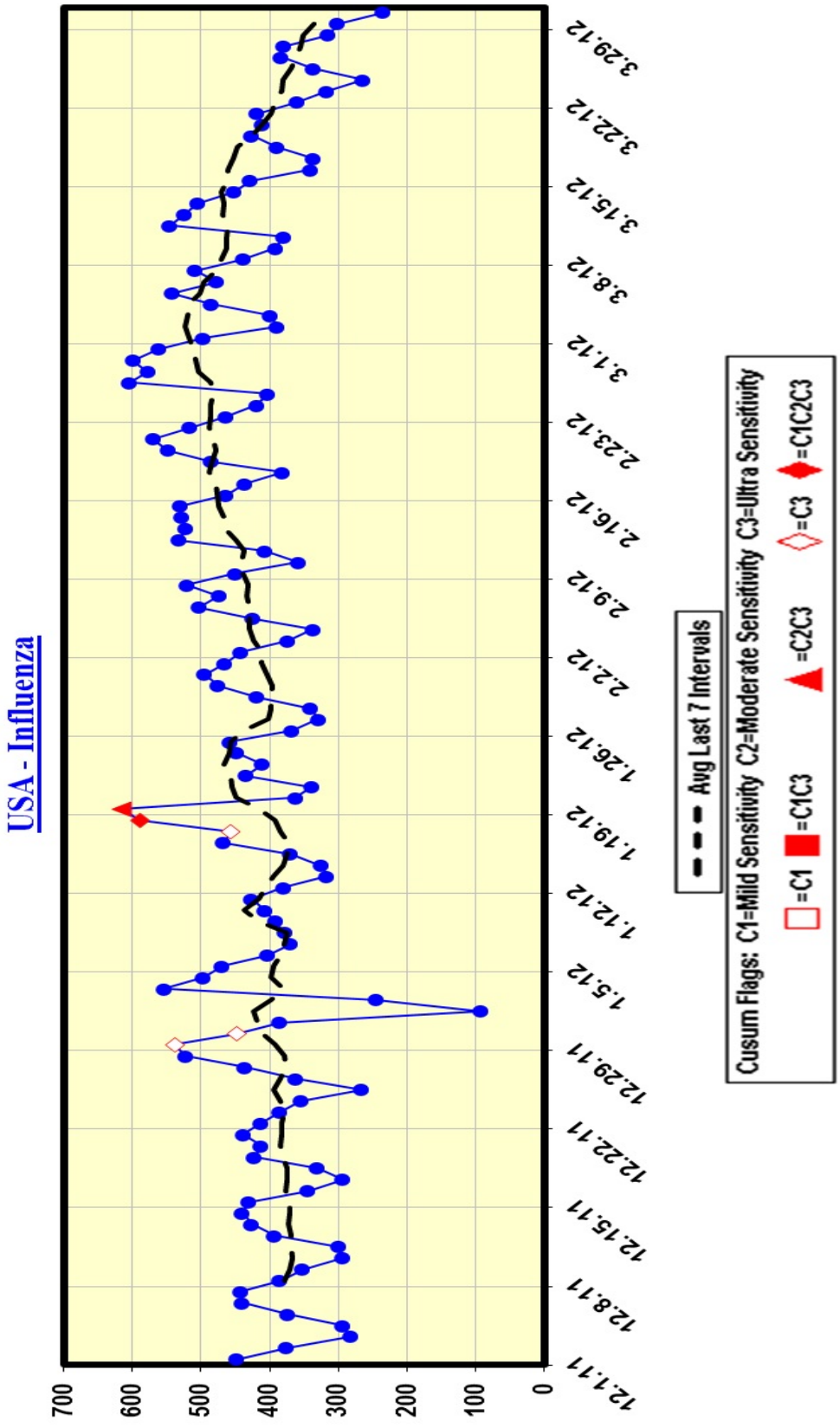

Figure 6.3: Daily Flags from Alerting Methods.

collected from public health providers in all states in USA. According to weekly laboratory-confirmed cases of influenza, ILI activity is measured for all states via percentage of patient visits to healthcare providers for ILI reported each week that is weighted on the basis of state population[1].

6.6 Conclusions

The previous chapters presented a text mining system that tracks peoples' status updates in real-time while the post classifier works with an input Twitter stream overtime. In addition, novel approaches for the extraction of disease-related events by determining the type of the post and extracting medical entity, timestamp and location have been presented. In order to provide a proof of the system, this chapter focused on evaluating of the whole system by evaluating the post classifier that distinguishes between post-related or not related to diseases, and showed the performance of the classification step with crucial dataset that contains ambiguous postings. Then, an evaluation was conducted for the real-time filtering system while the classifier works with the Twitter stream overtime. The latency and throughput have been presented as important metrics for the evaluation. Furthermore, the important evaluation of the system has been represented by the comparison between output of the system with national health statistics. Specifically, the results showed strong correlation between information on diseases that the system produced from Twitter data, and ILI values that were reported by the CDC.

[1] http://www.cdc.gov/flu/weekly/overview.htm

Chapter 7

Identification of Public Health-Related Topics

The previous chapters presented and evaluated techniques to track and identify diseases and outbreaks via Twitter. This chapter addresses the problem of how to effectively identify and browse health-related topics within Twitter data. The topic of smoking is chosen as an example for this study. The tobacco epidemic is one of the biggest public health threats the world has ever faced. It kills nearly six million people a year[1]. However, giving smokers some health guidelines about the risks of smoking is very important in stopping or reducing smoking. Health-related behaviors such as smoking are a trending topic of discussion for social network users. Studying the behavior of users towards smoking helps health organizations to understand the behavior of smokers and provide smokers with advices to avoid bad health habits. Smoking behavior relates to people using nicotine, cigarettes, marijuana, shisha, or any smoking product. Measuring health behaviors, such as smoking in populations overtime and space, is valuable to identify unusual behavior and target areas for providing health guidelines; however the traditional smoking control measures are generally cost-effective because information is collected from several sources such as articles and reports from WHO (World Health Organization), ministries of health, national statistical offices, and smoking control organizations. The social networks appear to play important

[1] http://www.who.int/mediacentre/factsheets/fs339/en/index.html

roles to measure the sentiment analysis on smoking among their users. Detection and tracking of smoking behavior on social networks can help to prevent or reduce the habit of smoking among people. Furthermore, studying the smoking behavior on social networks helps the governments combat not only the spread of tobacco use but also weed or marijuana. This chapter studies how Twitter messages(tweets) can be used to analyze opinions of the population toward smoking, discovers smoking-related themes on Twitter, as well as uses Twitter as a tool to identify smokers in different locations. The approach is to use n-gram models to extract features from the annotated dataset and use state-of-the-art techniques to build a classifier after adapting well annotated guidelines that distinguish the opinions of Twitter users by analyzing their smoke-related messages. A smoke-related message refers to a tweet that mentions any smoke-related keyword in its text (e.g., smoke, cigarette, nicotine, tobacco, and marijuana).

In summary, the contributions of this chapter are as follows:

- Describing the nature of smoke-related data on Twitter (Section 7.1).
- Using state-of-the-art methods that use Twitter data to identify relevant postings (Section 7.2).
- Validating the effectiveness of the methods on the different features set using annotating dataset (Section 7.3).
- Studying the behavior of smokers on Twitter and discovering the type of smoking over spaces(Section 7.4 and Section 7.5).

The methods and results in this chapter appeared in [56].

7.1 Tweets on Smoking

This study goes beyond the monitoring of diseases, and detect health-related behavior as Twitter provides a wealth of information about a user's behavior and interests (e.g., smoking behavior). This information can be used to raise awareness among the population, and the information offers an opportunity to survey the smoking prevalence in real-time. Some examples of tweets are as follows:

Category	Tweets
Positive	250
Negative	200
Neutral	50
Total	**500**

Table 7.1: Experimental Annotated Dataset

- *Hurry up! I want to smoke my cigar!* ,
- *I smoke so much weed* ,
- *going to smoke shisha :D* ,
- *Smoking is so disgusting, hate it so much.*

These examples are positive smoke-related messages that are worth considering for health behavior interests and a valuable source of peoples' opinions as in tweet (1), (2), and (3). The last tweet is a totally nonpositive smoke-related message.

7.2 Methods

7.2.1 Data Collection

Twitter API were used to collect tweets about smoking by crawling tweets containing smoke-related keywords (e.g., smoke, cigarette, nicotine, tobacco, shisha, hashish, weed, hookah, and marijuana) that are likely to be of significance and used by the crawler component (Section 4.1.1) to retrieve publicly available posts that have at least one of those keywords mentioned. The experiment dataset was selected from a full collection of one week. Each tweet is timestamped and geolocated using the author's self-declared home location. Actually, the internal annotation tool (Section 3.1.3) was used to annotate the experiment dataset by using the annotation guidelines that are described in Section 7.2.2. The final annotated dataset comprised 500 tweets: 250 positive, 200 negative, and 50 neutral, as shown in Table 7.1.

This work is only interested in smoke-related tweets that are positive. However, in the experiment, the neutral tweets are considered as negative. Besides, a

comprehensive dataset that matched a set of perspecified keywords are collected. The dataset was collected for over two months from July 1, 2012 to August 27, 2012, and the result was 20,239,490 smoke-related tweets that could be positive or negative, all the data is shown in Figure 7.1.

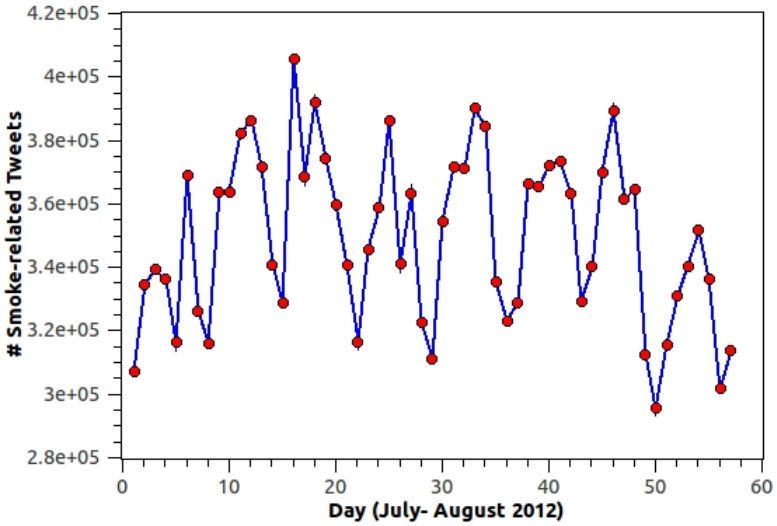

Figure 7.1: Smoke-term-Related Tweet Volume

7.2.2 Smoke-related Tweet Annotation

According to the nature of the smoke-related tweets that are collected, the tweets are annotated into positive, negative, or neutral.

7.2.2.1 Positive Smoke-related Tweet

Any tweet that confirms that the user is a smoker, likes or loves smoking, going to smoke anywhere, and mentions that the smoking is good. (e.g., *"Come on, lets smoke some marijuana in my room and relax"*, *"Going to smoke a cigarette or a pack"*, and *"smoking feels so good"*).

7.2.2.2 Negative Smoke-related Tweet

Any tweet that confirms that the user is not a smoker, hates or does not like smoking (e.g., *"I don't smoke"*, *"I hate when people smoke in public places"*, *"don't even wanna smoke"*). In addition, anti-smoking tweets that stating that smoking is dangerous, bad, or undesirable (e.g., *"Smoking is so pointless"*, *"Ain't smoking no more"*, and *"Cigarette smoking is dangerous to your health"*).

7.2.2.3 Neutral Smoke-related Tweet

Any tweet in which no clear opinion can be observed, like questions, conditions, or the tweet is not about smoking. (e.g., *"If you think a smoke alarm has stopped going off, don't hold it up to your ear"*).

7.2.3 Classifier Model

The approach in this Chapter uses machine learning methods for sentiment analysis on smoke-related tweets. Support vector machines (SVMs) classifier is a state-of-the-art supervised kernel method for machine learning. It was successfully adapted to text classification. Conversely, maximum entropy models (MaxEnt) [34] are feature-based models and have an alternative technique that is effective in natural language processing applications.

7.3 Evaluation and Results

Because of misspellings and slang in tweets, the cleansing task(Section 4.1.6) is used to remove undesired tokens. Then, using 5-fold cross-validation, Support vector machines (SVMs) and maximum entropy models were tested on term-based without cleansing (baseline), unigram, bigram, and both unigram and bigram as features. Both algorithms were performed using the machine learning toolkit Weka. The results are shown in Table 7.2. All unigram, bigram, and both features are extracted after cleansing the dataset from undesired words.

Without any preprocessing or cleansing of tweets (baseline), an accuracy of 85.6% (SVMs) and 84.8% (MaxEnt) could be achieved. Unigram features did not

		SVMs				MaxEnt			
Features Set	**#Features**	**Acc**	**Rec**	**Pre**	**F-M**	**Acc**	**Rec**	**Pre**	**F-M**
Baseline	1508	85.6	85.6	86.3	85.5	84.8	84.8	85.2	84.8
Unigrams	1131	84.8	84.8	85.5	84.7	86.8	86.8	87.2	86.8
Bigrams	2980	79.6	76.9	82	79.2	82	82	83.3	81.8
Unigrams and Bigrams	4117	85.6	85.6	86.6	85.5	86	86	86.7	85.9

Table 7.2: Evaluation Results (Accuracy (Acc), Recall (Re), precision (pre), and F-Measure (F-M))

improve the accuracy with respect to support vector machines classifier; however, they improved the maximum entropy classifier. Bigram features are useful to capture phrases such as "don't smoke" and "no smoking"; however, bigram, and both unigram and bigram did not improve the performance compared with baseline and unigram features because of sparseness problems (e.g., *"I donttttttt liiiikee smoking weeeed"*) and shared terms among tweets. Finally, it was found that the use of baseline features with SVMs is a bit better than unigram because, in baseline, the emoticons are considered as features. However, the emoticons are considered as noisy labels because they are not perfect to determine opinions in the tweet text (e.g., *"I got headache :)*").

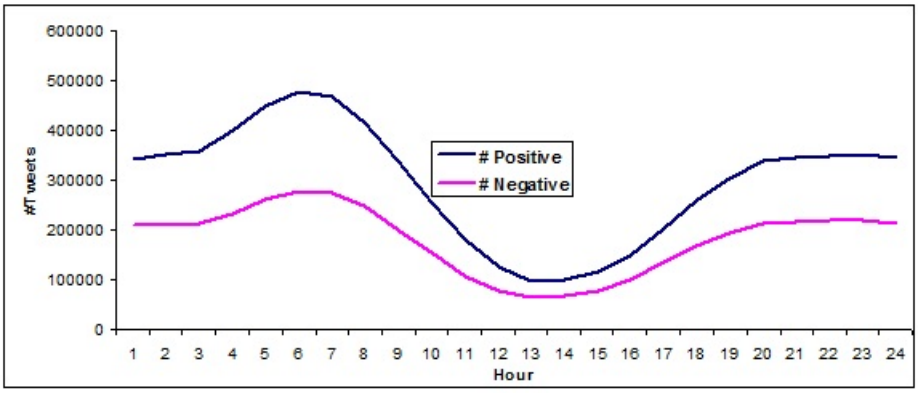

Figure 7.2: Diurnal smoke-related Messages

7.4 Sentiment Analysis on Smoking

The dataset of 20,239,490 tweets that were observed between July and August, 2012 have been analyzed. 7,183,211 tweets that were retweets have been removed. Retweeting is the process of repeating information previously tweeted by another user. Also 1,408,331 tweets that contained URL's have been removed. Subsequently, unigram is used as an extractor to retrieve features from 11,489,102 tweets. Then, by applying the MaxEnt classifier, 62% tweets were classified as positive and 38% as negative. The results were divided into 24 bins, one for each hour of the day. Figure 7.2 presents the daily behavior of Twitter users over a 24 h period for both positive and negative tweets. It was observed that the positive and negative tweets seem to be the same in activity movement during daily hours; however, the number of users that like smoking is more. An assumption is that the people talking about their personal feelings toward smoking are smokers. This means that the users who posted negative tweets could be smokers, but they hate smoking. Specifically, the activity of smokers decreases a bit from midnight onwards but increases a lot from the early morning. After 7 am posting activity gradually decreases until 3 pm. The lower peak appears between 12 pm and 2 pm. Then, the activity quickly increases and becomes stable from early night until midnight. These results reveal several implications about how people express their behavior toward smoking in context: people generally express their behavior towards smoking early in the morning, which indicates that people are generally more likely to smoke before work. In addition, they also report much in the night.

7.5 Discussion

This chapter proposed an approach to identify opinions on smoking from Twitter messages. It introduced two machine learning classifiers (SVMs and MaxEnt) that applied several types of features, and these classifiers were used to discover the sentiment on many different themes of smoking (e.g., cigarette, weed, and shisha). In the experiment, the performance of the classifier achieved up to 87% accuracy. It was found that using bigram, and both unigram and bigram did not

drastically improve the performance because of short messages, shared terms, and sparseness. Regardless of the challenges, the performance result was good enough to track the behavior of social web users toward smoking and discover smoke-related themes (Section 7.5.1). Moreover, Twitter could be used as a better tool to identify smokers by location(Section 7.5.2).

7.5.1 Tracking Public Interest with Twitter Data

This subsection analyzes several common topics related to smoking from the entire dataset. All results are reported as a percentage from all observed tweets, excluding retweets and tweets containing URL's. Figure 7.3 shows all daily positive smoke-related tweets (orange line) as a percentage of all observed tweets. By analysis of the frequency of smoke-related terms (e.g., smoke, cigar, weed, and marijuana), It was found that 88% of tweets contain the term smoke. The positive tweets that mention smoke-related terms were analyzed as a percentage of observed daily tweets that mention those terms.

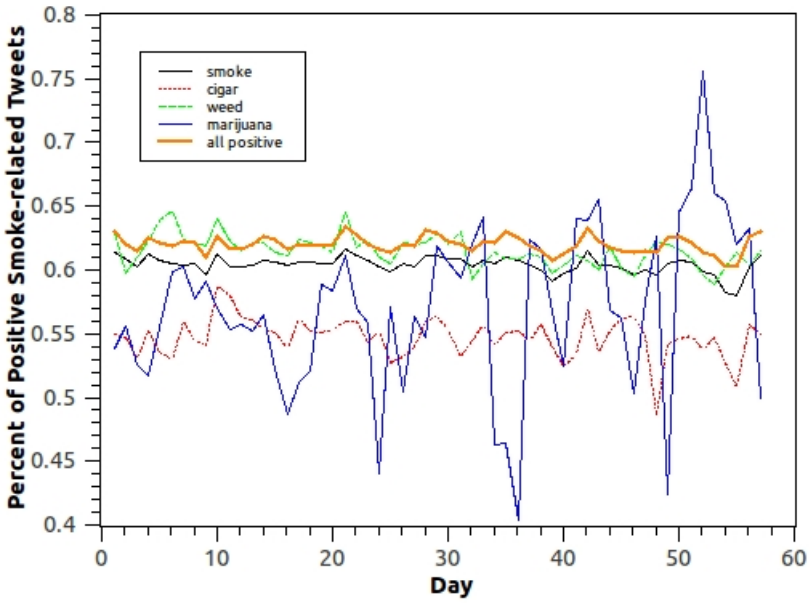

Figure 7.3: Smoke-term-Related Tweet Volume

From the analysis perspective, Twitter messages provide good information

for daily smoking activities, and they can be used to observe the daily behavior of marijuana users since the use of marijuana is illegal in most countries. Figure 7.3 also shows the percentage of positive smoke-related tweets that contain marijuana-related tweets (blue line) and weed-related tweets (green line). Realistically, most Twitter users use the slang word "weed" instead of the term "marijuana"; however, their related tweets are showed independently. The spikes that are related to an increase of total weed and marijuana-related messages can be very useful to governments or organizations that are fighting the spread of illegal smoking; moreover, it will help them provide health guidelines to the population so as to stop or reduce smoking.

7.5.2 Tracking smoke-related Tweets in spaces

For public health organizations, it is important to discover locations of the large percentage of users who report a behavior such as smoking. The location of smoke-related tweets can be tracked and distribute the location of smokers across geographic locations. Twitter allows users to specify their geographic location as meta-data. The location data is manually entered by the user or updated by a GPS enabled device. Comprehensive dataset is used to study the distribution of smokers across cities by matching the meta-data of location with city names and extract tweets from all positive smoke-related tweets that have been filtered in Section 7.5.1. Most Twitter users posted about their smoking behavior repeatedly. Thus, to study the prevalence of smoking using Twitter, only unique messages should be calculated.

This work thoroughly considered the number of users instead of the number of messages that they reported. Therefore, in all positive tweets, it was found that 54% users posted more than one message and unique users represent 46% of all positive datasets. Twelve percent of unique users mentioned that they were smoking weed and marijuana, 2% smoking cigarettes, and 72% of users mentioned that they just smoke (e.g., *"I just wanna smoke all day"*). A user that posted a positive tweet is considered as a smoker. Moreover, the number of smokers for some cities where English is the official language is calculated. Figure 7.4 shows the number of unique smokers for 11 cities from a dataset of two months (July and

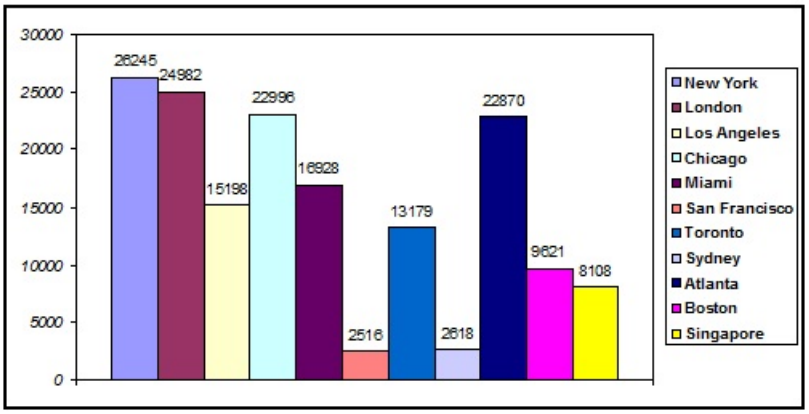

Figure 7.4: Number of Unique Users (smokers) for July and August 2012.

August 2012) that is not enough to compare with official statistics; however, the results demonstrate that Twitter can be used to track smokers across locations.

7.6 Conclusion

This chapter examined people's behavior toward public health-related topics on a popular social medium. Twitter was used as an example of a high-traffic social network for measuring the behavior of the population towards smoking over-time and space. Several feature sets have been tested with the machine learning model to automatically detect relevant smoking tweets. The methods used in this chapter can be used as a possible tool for researchers of public health and practitioners to better understand public health hazards, such as smoking, through large datasets of Twitter messages.

Chapter 8

Related work

This chapter reviews the literature relevant to this dissertation. Section 8.1 outlines prior research on public health identification in textual news documents. Section 8.2 describes works on single disease detection and analysis in social web. This was addressed in Chapter 3 and Chapter 4. Section 8.3 discusses related efforts on named entity recognition of tweets that were addressed in Chapter 5. Section 8.4 provides an overview of related work that estimates location entity on Twitter. Finally, Section 8.5 discusses research related to health topics on social web (such as smoking), which is related to the task in Chapter 7.

8.1 Disease Identification in Textual News

There has been growing interest in monitoring disease outbreaks using the Internet. Previous approaches have applied data mining techniques to analyze news articles. Grishman et al. [49] described a system known as Proteus-BIO, which is capable of searching documents about infectious disease outbreaks on the web. The system gathers web pages, extracts information about outbreaks, and presents the extracted information in a tabular form with links back to the documents. Mawudeku et al. [41] explored the Global Public Health Intelligence Network (GPHIN) that is a secure Internet-based early warning system. The system gathers preliminary reports of public health significance on a near real-time, 24 h a day, 7 days a week basis. This unique multilingual system gathers and

provides relevant unverified information on disease outbreaks and other public health events by monitoring global media sources in six languages, namely: Arabic, Chinese, English, French, Russian, and Spanish. This automated process, which includes filtering for relevancy and categorizing of information, is complemented by human analysis. News articles in English are posted in the system and translated into the other five languages. News articles in any of the five non-English languages are posted in the system and translated into English.

Brownstein et al. [33] presented the HealthMap system which is a freely accessible, automated real-time system that monitors, organizes, integrates, filters, visualizes, and disseminates online information about emerging diseases. The goal of HealthMap is to deliver real-time intelligence on a broad range of emerging infectious diseases for a diverse audience, from public health officials to international travelers. Collier et al. [9] developed an ontology-enabled text mining system known as BioCaster. The system was designed to detect and track the distribution of infectious disease outbreaks from Internet news.

Moreover, MedISys gathers data from global media sources such as news wires and web sites to identify information about disease outbreaks [38]. PULS(Pattern-based Understanding and Learning System) is a system integrated with MedISys, and extracts event data from the English MedISys articles and produces searchable outbreak data in table format.

8.2 Disease Identification on the Social Web

While disease detection and tracking in textual news documents have been studied in depth, the identification of disease on the social web is still in its infancy. Several related works explored the disease and outbreak detection on the social web.

Ginsberg et al. [22] developed the widely used system Google Flu Trends, which uses query logs from users in the Google search engine. It reported a high Pearson correlation coefficient of 97% with the CDC-observed ILI percentages. Google Flu Trends tested each query among over 50 million candidates and finally established the top 45 query terms. However, the drawback is that although the information could be integrated into Epidemic Intelligence (EI) systems since it

is stored in commercial search query logs and tends not to be freely available. Additionally, Google Flu does not show which search terms have been identified. Quincey et al. [14] have demonstrated the potential of Twitter in outbreak detection by collecting and characterizing over 135,000 messages pertaining to the H1N1 virus over a one week period, though no attempt is made to estimate influenza rates. Both Lampos [37] and Culotta [12] used linear regression to detect keywords that correlate with influenza rates, then combine these keywords to estimate national influenza rates. Chew et al. [13] collected tweets containing the keywords "H1N1", "swine flu", and "swineflu" from May 1 to December 31, 2009. They categorized the tweets into many categories (resource, personal, jokes, spam, etc) depending on the content of the tweet. They used two concepts to assign each tweet into a related category. These were the manual and automated concepts. In the manual concept, they assigned the tweets of 9 days into the related category. Conversely, they used SQL queries for the automated concept, assigning the same tweets of 9 days. SQL was used in order to search for keywords and phrases that match the content categories. The results of both the concepts were then correlated. The study demonstrated the potential of using social media to conduct "infodemiology" studies for public health.

Gomide et al. [29] analyzed how the Dengue outbreak reflected on Twitter. They proposed a methodology based on four dimensions: the amount of tweets, location of tweets, publishing time of tweets, and content of tweets that is the overall population perception/sentiment about dengue epidemic. Signorini et al. [5] tracked the rapidly-evolving public sentiment with respect to H1N1 or swine flu. They were interested in monitoring influenza-related traffic within the United States, and they excluded non-English tweets and tweets published outside the U.S. They used two datasets. The first contained 951,697 tweets, and the second contained 4,199,166 tweets. All of them contain "swine", "flu", "influenza", or "h1n1". They used the search-term concept to distinguish between tweets (e.g., H1N1-related tweets and drug-related tweets). Indeed, they studied the possibility of using the Twitter stream to make real-time estimation of weekly ILI values that were produced by the CDC.

ARAMAKI et al. [18] collected 300 million tweets, starting from November, 2008 to June, 2010 via Twitter API. Then, they extracted 0.4 million influenza-

related tweets using the look-up of influenza. Five thousand tweets were annotated manually and used as training data, while the remaining data was used for testing. The SVM was used to filter out the negative influenza tweets. Their experiment with test data has been applied to detect influenza epidemics with correlation performance 0.89% compared with the Google Flu Trend. However, they did not compare their results with national official statistics.

Paul et al. [47] collected 2 billion tweets from May, 2009 to October, 2010; they filtered those tweets by 20,000 key phrases related to illness, diseases, symptoms, and treatment.

Most related works that have been explored in this section focused on a single disease, i.e., the "flu" or "dengue fever". However, This dissertation went beyond one single disease via detection, tracking, and analysis many different types of diseases and outbreaks, such as normal diseases, viruses, bacteria, and symptoms, as well as human behavior (e.g., sick, tired, pain, and ache), are detected as well. Moreover, the text mining system collects, identifies, filters, and analyzes the real medical cases of disease in real-time, and the comparison with the national health statistics is different from other works of related studies because the pure personal medical cases have been used to predict the national health statistics. Previous work that also went beyond a single disease is the work of Paul et al.

However, the current study differed from the study by Paul et al, in many aspects. First, the studies differ in the way disease-relevant tweets are segregated from irrelevant tweets. The current study applied annotation guidelines to distinguish disease-relevant tweets from nonrelevant tweets. Enough training datasets that covered different types of diseases were also collected. Conversely, Paul et al. labeled 5,128 tweets as related or unrelated to health. The related tweets were messages that indicated that the user was sick with an illness or the message was about health. Unrelated tweets included ambiguous text, or messages that were in a language other than English. It also included tweets irrelevant to a person's health(e.g., news). In addition, Paul et al. applied SVMs with 10-fold validation to obtain a classifier with 90.4% precision. The recall performance was also 32.0% that was poor compared with the recall performance reported in this study, as shown in Section 3.3.4.3. Moreover, Paul et al. applied their classifier to the 11.7 million messages and produced a corpus of 1.63 million health related

tweets. Finally, a probabilistic model known as the Ailment Topic Aspect Model (ATAM) was used to separate illness text from other text.

8.3 Named Entity Recognition on Twitter

A few studies have focused on named entity recognition on Twitter data. This section summarizes the related work of named entity recognition on tweets. Doerhmann [16] used a normalizer as a preprocessor to a named entity recognition (NER) tool to improve the productivity of the NER, his normalizer used to translate the text into standard English. Then, he used the voting CRF classifier system that receives results of many different classifiers on the same tweet text and chooses the best result. Finin et al. [60] showed how to efficiently annotate large volumes of data for information extraction tasks. They used the Amazon Mechanical Turk for annotating named entities on Twitter and used CRF model to evaluate the effectiveness of human labeling. Ritter et al. [3] addressed the challenge of the poor performance of current tools of Part-of-Speech tagging and named entity recognition when they are applied to Twitter data. They annotated 2400 tweets manually with 10 types of categories that are popular on Twitter and on that data they used a supervised approach based on LabeledLDA for evaluation.

In fact, the current studies of named entity recognition of tweets are still developing. The named entity recognition in this thesis focused only on disease entity recognition and the entity of location. However, there is no effective, practical, and available tool for recognizing biomedical terms in tweets at the moment because all related works that are presented in this subsection focus on general named entity, such as person, organization, location, and product. Thus, A practical NER system was used for disease entity recognition in Chapter 5. A text parser was used to extract the medical entity from each tweet into an appropriate predefined category.

8.4 Location Entity Recognition on Twitter

Extraction of the location entity is a major challenge because Twitter allows the users to specify their geographic location as meta-data, and the users may enter incorrect or unreal locations. However, a few studies focused on estimate named entity type of location from Twitter.

Hecht et al. [7] collected 62 million tweets using Twitter streaming API, and identified only English tweets which were 51% of the whole dataset. Consequently, they found that only 0.77% of 62 million tweets contained location information. They used a sample classification model to predict the location of users from tweet content. Java et al. [2] used Yahoo! Geocoding API to get latitude and longitude coordinates, and they showed the results of the geographical distribution of Twitter users in each continent. Li et al. [66] used a ranked approach to predict POI (Place Of Interest) tags of tweets using language modeling method and time dimension. Cheng et al. [69] proposed a probabilistic framework to predict user location based on the content of the user's tweets. Their methods use local word identification to predict locations at city level. Chandra et al. [59] estimated a user's geographic location information using public words inside the user's tweeted text. In Chapter 5, Two techniques (geocoding and geoparsing) were used to estimate the location of the Twitter user and location entity inside the text, respectively.

8.5 Behavior Analysis on Twitter

There are previous works that studied sentiment analysis. Pang and Lee [46] used different techniques and approaches to classify movie reviews (not from Twitter). Go [4] used also different machine learning algorithms to classify the general sentiment of Twitter messages. Twittratr[1] is using a simpler keyword-based approach to classify the sentiments in the tweets. Prier [36] studied the possibility of identifying health-related topics such as tobacco using the Latent Dirichlet Allocation (LDA) model. The work in chapter 7 focused only on the sentiment analysis of smoking-related tweets, and identified the smokers by examining their messages:

[1] http://apiwiki.twitter.com/Twitter-APIDocumentation

positive or negative, as well as discovering many different themes of smoking (e.g., marijuana or weed, nicotine, and cigarettes).

Chapter 9

Conclusions and Future Work

The birth and development of the social web has fundamentally changed how individuals interact in our life. Social web sites such as Twitter allow individuals to instantly communicate with large networks of friends, acquaintances, families, and colleagues. Twitter can be a source and distributor of public health information such as diseases and outbreaks, as well as health-related topics. Because of the huge and vast amount of data, publicly available, and real-time nature, Twitter messages are valuable information for monitoring disease outbreaks in real-time.

This dissertation presented tracking, identification, and analysis of diseases in real-time via Twitter. Different types of diseases have been identified such as names of normal diseases, viruses, bacteria, symptoms, and outbreaks, as well as peoples' behaviors towards health-related topics (e.g., smoking) have been identified as well. Furthermore, a proof of using the social web for disease monitoring has been provided by comparing the system's output with the national health statistics. This dissertation specifically outlined the characteristics and types of public health information that are available on Twitter. The analysis provided a good knowledge that exploited to develop suitable techniques for identifying the diseases on Twitter (Chapter 2). For the disease-reporting messages classification scenario, where diseases are classified in a supervised manner, well established guidelines have been adopted for defining clinical cases to build a dataset for a disease-reporting classifier and to learn about the characteristics of the dataset. Besides, two classification algorithms have been exploited and tested on several

feature sets (Chapter 3). To track and detect the diseases in real-time, a real-time filtering system was developed that uses data mining techniques to crawl, index, extract, cleansing, and classify postings in real-time overtime (Chapter 4). Then, in the Named Entity Recognition on disease-related postings scenario, a framework was developed to determine the meaning of the Twitter message, as well as extract entities to create the events (Chapter 5). All techniques were integrated into one system, the whole system has been evaluated, and its output was used to predict the disease level before they were reported by official sources (Chapter 6). Finally, approaches for detection of public health-related topics on Twitter were explored. Specifically, smoking behaviors among the population by using Twitter messages were identified (Chapter 7)

To summarize, this dissertation presents a variety of significant techniques for tracking, detecting, and analyzing Twitter content that represents a valuable source of health information. Specifically, this dissertation provided important insights regarding the types of diseases that exist on Twitter and the characteristics of their associated topics (e.g., smoking). Then, developed methods for annotating and classifying different types of public health information namely, symptoms, normal diseases, bacteria', viruses, and outbreaks. As the number of Twitter messages is large, containing spam (e.g., retweets, Bieber-fever-related messages), and noisy. However, many techniques were designed to meet these challenges, and identify pure cases of disease in real-time overtime. Overall, this dissertation provides a system for tracking and identifying diseases on Twitter, and offers contributions for improving the utility of Twitter messages through disease identification, and analysis. Promising directions for future work can be built on these contributions to provide a powerful web interface to show the public health activities of Twitter, as well as shows the statistics of diseases, and public health events around the world.

Bibliography

[1] CDC 2003. Syndrome definitions for diseases associated with critical bioterrorism-associated agents dated. October 23, 2003. Accessed at www.bt.cdc.gov/surveillance/syndromedef/ on November 21, 2003.

[2] Tim Finin Belle Tseng. Akshay Java, Xiaodan Son. Why we twitter: Understanding microblogging usage and communities. Joint 9th WEBKDD and 1st SNA-KDD Workshop 07 ACM, 2007.

[3] Mausam Alan Ritter, Sam Clark and Oren Etzioni. Named entity recognition in tweets: An experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK.

[4] Richa Bhayani Alec Go and Lei Huang. Twitter sentiment classification using distant supervision. Stanford University, 2009.

[5] Philip M. Polgreen Alessio Signorini, Alberto Maria Segre. 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*, 2011.

[6] S. Asur and B. A. Huberman. Predicting the future with social media. Technical Report arXiv:1003.5699, CoRR, 2010.

[7] Bongwon Suh Ed H. Chi. Brent Hecht, Lichan Hong. Tweets from justin bieber's heart: The dynamics of the "location" field in user profiles. Northwestern University, Palo Alto Research Center, 2011.

[8] Verspoor K. Johnson H.L. Roeder C. Ogren P.V. Baumgart-ner-Jr. W.A. White E. Tipney H. Hunter L.: Cohen, K.B. High-precision biological event extraction with a concept recognizer. Workshop on BioNLP: Shared Task collocated with the NAACL-HLT 2009 Meeting. Association for Computational Linguistics, 2009.

[9] Matsuda Goodwin R Conway M.-Tateno Y Ngo Q. Dien D Kawtrakul A. Takeuchi K. Shigematsu M. Collier N. Doan S., Kawazoe A. and Taniguchi K. Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics, 24(24):2940-2941, Oxford University, 2008.

[10] N. Cristianini and J. Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, 2000.

[11] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *In Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0217-3. doi: http://doi.acm.org/10.1145/1964858.1964874. URL http://doi.acm.org/10.1145/1964858.1964874.

[12] Aron Culotta. Detecting influenza outbreaks by analyzing twitter messages. 2010. URL http://arxiv.org/pdf/1007.4748.pdf.

[13] Gunther Eysenbach Cynthia Chew. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 2010.

[14] E. de Quincey and P. Kostkova. Early warning and outbreak detection using social networking websites: The potential of twitter. In *Electronic Healthcare.*, pages 21–24. Springer, Berlin Heidelberg, 2010.

[15] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *In Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9. doi: http://doi.acm.org/10.1145/1753326.1753504. URL http://doi.acm.org/10.1145/1753326.1753504.

[16] Cassaundra Doerhmann. Named entity extraction from the colloquial setting of twitter. 2011.

[17] D. Apostolou E. Bothos and G. Mentzas. Using social media to predict future events with agent-based markets. IEEE Intelligent Systems, vol. 25, no. 6, pp. 50-58, Nov, 2010.

[18] Mizuki MORITA. Eiji ARAMAKI, Sachiko MASKAWA. Twitter catches the flu: Detecting influenza epidemics using twitter. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK, 2011.

[19] Centers for Disease Control and Prevention(CDC). http://www.cdc.gov/.

[20] Jr. Hegler B. L. Fricker, R. D. and D. Dunfee. Comparing biosurveillance detection methods: Ears' versus a cusum-based methodology,. Statistics in Medicine, 27, 3407–3429, 2008.

[21] Seymour Geisser. Predictive inference. New York, NY: Chapman and Hall. ISBN 0-412-03471-9, 1993.

[22] Patel RS Brammer L Smolinski MS. Ginsberg J, Mohebbi MH. Detecting influenza epidemics using search engine query data. 2009.

[23] US Government. Homeland security presidential directive 21: Public health and medical preparedness. Accessed on-line at www.fas.org/irp/offdocs/nspd/hspd-21.htm, 2007.

[24] Isabelle Guyon. An introduction to variable and feature selection. Journal of Machine Learning Research 1157-1182, USA, 2003.

[25] Hosung Park Haewoon Kwak, Changhyun Lee and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th International World Wide Web Conference (WWW '10), 2010.

[26] H. Mao J. Bollen and A. Pepe. Determining the public mood state by analysis of microblogging posts. In Proc. Of the Alife XII Conf. MIT Press, 2010.

[27] H. Mao J. Bollen and X.-J. Zeng. Twitter mood predicts the stock market. Technical Report arXiv:1010.3003, CoRR, 2010.

[28] Painter I Duchin J: Jackson ML, Baer A. A simulation study comparing aberration detection algorithms for syndromic surveillance. Medical Informatics and Decision Making , 7(6), BMC, DOI: 10.1186/1472-6947-7-6, 2007.

[29] Wagner Meira Virgilio Almeida Fabricio Benevenuto Fernanda Ferraz Janaina Gomide, Adriano Veloso and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. WebSci, 2011.

[30] Sampo Pyysalo Yoshinobu Kano Jin-Dong Kim, Tomoko Ohta and Jun'ichi Tsujii. Overview of the bionlp'09 shared task on event extraction. In Proceedings of the NAACL-HLT 2009 Workshop on Natural Language Processing in Biomedicine (BioNLP'09), 2009.

[31] T. Joachims. Learning to classify text using support vector machines: Methods, theory and algorithms. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[32] F. Jordans. Who working on formulas to model swine flu spread. 2009.

[33] B.Y. Reis J.S. Brownstein, C.C. Freifeld and K.D. Mandl. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. PLoS Medicine , 5:1019–1024, 2008.

[34] J. Lafferty K. Nigam and A. Mccallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999.

[35] Coulombier D. Kaiser, R. Different approaches to gathering epidemic intelligence in europe. EuroSurveill, 2006.

[36] Matthew S.Smith. Kyle W.Prier. Identifying health-related topics on twitter,an exploration of tobacco-related tweets as a test topic. Springer-Verlag Berlin Heidelberg, 2011.

[37] V. Lampos and Cristianini. Tracking the flu pandemic by monitoring the social web. Information Processing (CIP 2010)., 2010.

[38] Steinberger R. Weber T.P. Yangarber R. van der Goot E. Al Khudhairy D.H. Stilianakis N. Linge, J.P. Internet surveillance systems for early alerting of health threats. EuroSurveill, 2009.

[39] C. D. Manning and H. Schutze. Foundations of statistical natural language processing. Foundations of MIT Press, 1999.

[40] Căciulă Maricel. Study about opencalais api practical usage in linked data context.

[41] Mawudeku and M. Blench. Global public health intelligence network (gphin). In 7th Conference of the Association for Machine Translation in the Americas, 2006.

[42] F. ARACHNID: Menczer. Adaptive retrieval agents choosing heuristic neighborhoods for information discovery. Proceedings of the 14th International Conference (ICML97). Morgan Kaufmann, 1997.

[43] Ian Painter Michael L Jackson, Atar Baer and Jeff Duchin. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Medical Informatics and Decision Making, USA, 2007.

[44] Global Health Monito. http://born.nii.ac.jp/.

[45] World Health Orginazation(WHO). http://www.who.int/csr/alertresponse/epidemicintelligence/en/index.html.

[46] Lillian Lee Pang and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[47] Michael J. Paul and Mark Dredze. A model for mining public health topics from twitter. Technical Report. Johns Hopkins University,, 2011.

[48] Beth Sundhei. Ralph Grishman. Message understanding conference. A Brief History In Proceedings of the 16th International Conference on Computational Linguistics,, 1996.

[49] Silja Huttunen Ralph Grishman and Roman Yangarber. Real-time event extraction for infectious disease outbreaks. In Proceedings of the second international conference on Human Language Technology Research, pages 366–369, San Francisco, CA, USA, 2002.

[50] E. Riloff. Information extraction as a stepping stone toward story understanding. Understanding Language Understanding: Computational Models of Reading (MIT Press, Cambridge, MA, 1999.

[51] O'Connor Rolka, H. Real-time public health biosurveillance: Systems and policy considerations. in: Infectious disease informatics and biosurveillance: Research, systems and case studies. Springer.

[52] Matsuo Y. Sakaki T, Okazaki M. Earthquake shakes twitter users: Realtime event detection by social sensors. 2010.

[53] Mustafa Sofean and Matthew Smith. Medical case-driven classification of microblogs: Characteristics and annotation. In Proceedings of the 2012 Conference on International Health Informatics Symposium. ACM, USA, 2012.

[54] Mustafa Sofean and Matthew Smith. A real-time disease surveillance architecture using social networks. In Proceedings of the 24th European Medical Informatics Conference(MIE) Pisa, Italy, 2012.

[55] Mustafa Sofean and Matthew Smith. A real-time architecture for detection of diseases using social networks: Design, implementation and evaluation. In Proceedings of the ACM Conference on Hypertext and Hypermedia (HT2012) Milwaukee, USA, 2012.

[56] Mustafa Sofean and Matthew Smith. Sentiment analysis on smoking in social networks. In Proceedings of the 14th World Congress on Medical and Health Informatics(MedInfo) Copenhagen, Denmark, 2013.

[57] D.M. Sosin. Syndromic surveillance: The case for skillful investment view. Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science, 1, 247–253.

[58] Robert Steffen. Travel medicine prevention based on epidemiological data. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85 (2): 156–162, 1991.

[59] Latifur Khan Swarup Chandra and Fahad Bin Muhaya. Estimating twitter user location using social interactions - a content based approach twitter users. IEEE International Conference on Social Computing, 2011.

[60] Anand Karandikar Nicholas Keller Tim Finin, Will Murnane and Justin Martineau. Annotating named entities in twitter data with crowdsourcing.

[61] Ratard R. Straif-Bourgeois S. Sokol T. Averhoff F. Brady J. Staten D. Sullivan M. Brooks J.T. Rowe A.K. Johnson K. Vranken P. Toprani, A. and E. Sergienko. Surveillance in hurricane evacuation centers - louisiana. morbidity and mortality. Weekly Report, 55, 32–35.

[62] Google Flu Trends. http://www.google.org/flutrends/.

[63] Luis Gravano Vasileios Hatzivassiloglou and Ankineedu Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval (SIGIR), 2000.

[64] A. Viera and J. Garret. Understanding interobserver agreement: the kappa statistic. Family medicine, 2005.

[65] Jing He Yang Song Palakorn Achananuparp Ee-Peng Lim Wayne Xin Zhao, Jing Jiang and Xiaoming Li. Topical keyphrase extraction from twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11), 2011.

[66] Arjen P. de Vries Carsten Eickhoff Martha Larson Wen Li, Pavel Serdyukov. The where in the tweet. CIKM'11, Glasgow, Scotland, UK, 2011.

[67] Furu Wei Ming Zhou. Xiaohua Liu, Shaodian Zhang. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistic, Portland, Oregon.

[68] Y. Yang. An evaluation of statistical approaches to text categorization. Information Retrieval J, 1:69-90, 1999.

[69] Kyumin Lee. Zhiyuan Cheng, James Caverlee. You are where you tweet: A content-based approach to geo-locating twitter users. CIKM'10, Toronto, Ontario, Canada, 2010.