TEXT MINING STUDY OF MICROBLOG ACTIVITY ON HEATWAVES

by

FARANAK JALALZADEHFARD

(Under the Direction of Lakshmish Ramaswamy)

ABSTRACT

Evidence shows that heatwave events are increasing all around the world. They cause massive impact on public health, human constructions, and economy. As a small but progressive step, we performed data mining techniques on the famous microblog, Twitter, to find messages sent about this environmental phenomenon. We focused on heatwave-related tweets, and collected all containing the keyword #heatwave. The collecting process started from September 5th for nine weeks. We applied eight classification algorithms (Bayes Network, Naïve Bayes, Multinomial Naïve Bayes, Decision Tree, Random Forest, KNN, SVM, Maximum Entropy) to learn the patterns of related tweets and create proper classification models to classify new tweets. We achieved a high f-score (more than 90) in classifications. Our findings confirm that social media reflects the severe heatwave events.

INDEX WORDS: tweet, classification, heatwave, severe weather

TEXT MINING STUDY OF MICROBLOG ACTIVITY ON HEATWAVES

by

FARANAK JALALZADEHFARD

B.Eng., Azad University of Tehran, 2009

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2018

TEXT MINING STUDY OF MICROBLOG ACTIVITY ON HEATWAVES

by

FARANAK JALALZADEHFARD

Major Professor:	Lakshmish Ramaswamy
Committee:	Hamidreza Arabnia
	Khaled Rasheed

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

ACKNOWLEDGEMENTS

This thesis has been possible with the help and support of many individuals.

I would first like to thank my advisor Professor Lakshmish Ramaswamy who patiently taught me how to look deep and scientifically even into simple topics. He allowed this project to be my own work while he led me on the correct path whenever it was needed. I would also like to thank Professor Khaled Rasheed who was always willing to help me with his valuable guidance. Also, I am thankful to Professor Hamidreza Arabnia who has been always supportive, and has advised me personally and professionally during these years.

In addition, I would like to thank all faculty and staff members who provided me the opportunity to study at Computer Science department of University of Georgia.

I would like to express my gratitude to my parents and my brothers for their encouragement and believing in me. I would particularly like to thank my parents whose love and support are with me in whatever I pursue. This accomplishment would not have been possible without them.

Finally, I am grateful to all my friends who warmed my heart while I was far away from home.

Thank you,

Faranak Jalalzadehfard

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

1. INTRODUCTION

Wide spread of electronic devices, implementing sensors on almost any appliance, and collecting their data through the Internet promises a revolutionary future for IoT (Internet of Things). In many cases, especially when the human is a subject of interest, it makes sense to combine the data directly generated by people on social media with sensor data to acquire higher efficiency and accuracy in the analysis of complex issues. Indeed, citizen sensing or social sensors is an emerging research topic, which is often studied under more general category of crowdsourcing and can play an important role in environmental and public health surveillance and crisis/disaster informatics [1, 2, 3] .

Twitter, as a major social network, has been frequently used in different disaster detection and crisis management scenarios in the recent years. The real-time nature of Twitter makes it a very good data source of social sensors to quickly learn about the situation in the area of a crisis/disaster and affected people in order to improve crisis management accordingly; for example, Twitter played a key role in disaster management of Japan Earthquake in 2011 [4], the Deep Water Horizon disaster 2010 [5], explosions in Boston Marathon 2013 [6], and Typhoon in the Philippines 2013 [7].

In this work, we are interested in a particular kind of natural phenomena, namely heatwave. In general, heatwaves are short-term extreme temperatures that can cause health risks for the citizens. Although crisis related to heatwaves are not given as much attention as other natural disasters, they can be very dangerous for public health and cause vast mortality. For example, UK experienced a heatwave in July 2013 which lasted for three weeks. The

rate of death related to this heatwave was estimated 540-760 in England and 60-100 in Wales [8].

In comparison to the other natural disasters, heatwave has special characteristics that makes it different. First, there is no common global metric for detecting heatwave. For example, while 90 degrees Fahrenheit is an ordinary temperature for a summer day in Atlanta-USA, three consecutive days of this temperature in London-UK will be a level three warning [8]. Second, the impact of heatwave may not be the same on different people depending on their age, health condition, and their dwellings. The most vulnerable people in case of a heatwave are elders, kids, and chronically ill people. Moreover, it can be more dangerous for someone who is homebound in an un-air-conditioned apartment. These characteristics indicate that impact of heatwave on public health is more dependent on the feeling of people about the temperature rather than the physical temperature of the environment. While this property suggests that social sensors should be very helpful in detecting and analyzing heatwaves; surprisingly, there is very limited research work on this issue.

The focus of this work lies in improving the accuracy of near real-time disaster detection based on the data from Twitter. Early detection of heatwave like any other natural disaster can help in crisis management; e.g., to send early warnings, prepare the emergency facilities and supplies accordingly. Moreover, crowdsourcing can improve the credibility of alerts, and collect feedback about services for quick action.

The United States National Oceanic and Atmospheric Administration (NOAA) publishes the detailed information of unusual weather in a database named Storm Events. Looking to the available information during the time we collected tweets (5th September to 31st

October) shows16 rows that report heatwave events in three states of California, Missouri and Nevada. Table 1 displays the data extracted from the Storm Events dataset [9].

Table 1: Heatwave events in the US during our research from Storm Events dataset

	Type of Heat	Location	Beginning	Ending	Deaths (directly-related)	Deaths (indirectly-related)
1	Heat	LAS VEGAS VALLEY	9/8/2017 0:00	9/11/2017 23:59	1	2
2	Heat	ST. LOUIS	9/11/2017 15:30	9/11/2017 16:30	1	0
3	Excessive Heat	SAN DIEGO COUNTY COASTAL AREAS	10/23/2017 10:00	10/23/2017 17:00	0	0
4	Excessive Heat	SAN DIEGO COUNTY VALLEYS	10/23/2017 10:00	10/23/2017 17:00	0	0
5	Excessive Heat	SAN BERNARDINO AND RIVERSIDE COUNTY VALLEYS - THE INLAND EMPIRE	10/23/2017 10:00	10/23/2017 17:00	0	0
6	Excessive Heat	ORANGE COUNTY INLAND	10/23/2017 10:00	10/23/2017 17:00	0	0
7	Excessive Heat	ORANGE COUNTY COASTAL	10/23/2017 10:00	10/23/2017 17:00	0	0
8	Excessive Heat	SAN DIEGO COUNTY COASTAL AREAS	10/24/2017 10:00	10/24/2017 17:00	0	0
9	Excessive Heat	SAN DIEGO COUNTY VALLEYS	10/24/2017 10:00	10/24/2017 17:00	0	0
10	Excessive Heat	SAN BERNARDINO AND RIVERSIDE COUNTY VALLEYS - THE INLAND EMPIRE	10/24/2017 10:00	10/24/2017 17:00	0	0
11	Excessive Heat	ORANGE COUNTY INLAND	10/24/2017 10:00	10/24/2017 17:00	0	0
12	Excessive Heat	ORANGE COUNTY COASTAL	10/24/2017 10:00	10/24/2017 17:00	0	0
13	Excessive Heat	SAN DIEGO COUNTY COASTAL AREAS	10/25/2017 10:00	10/25/2017 17:00	0	0
14	Heat	ORANGE COUNTY INLAND	10/25/2017 10:00	10/25/2017 17:00	0	0
15	Heat	SAN DIEGO COUNTY VALLEYS	10/25/2017 10:00	10/25/2017 17:00	0	0
16	Excessive Heat	ORANGE COUNTY COASTAL	10/25/2017 10:00	10/25/2017 17:00	0	0

We collected tweets starting from 5th September for nine weeks using #heatwave keyword.

There are challenges in tweet classifications. First of all, heatwave is a polysemous word and have several meanings. Second, there is a limited number of character in each tweet. Furthermore, many acronyms are used in tweets which make the classification more complicated. Some tweets are sharing links, or photos without providing enough texts which leads to ambiguity in concept of a tweet.

We searched for the possible use of heatwave on Wikipedia. It is found out that in addition to severe hot weather, the term of heatwave has been also used in other contexts such as the name of a music band, a character in comic books, a song and album, movie and novel. We need to distinguish weather relevant tweets. Therefore, we labeled some tweets manually and used them to find suitable features for classifications. We evaluated the result of several classifiers and used the best ones to label the unseen tweets (Figure 1).

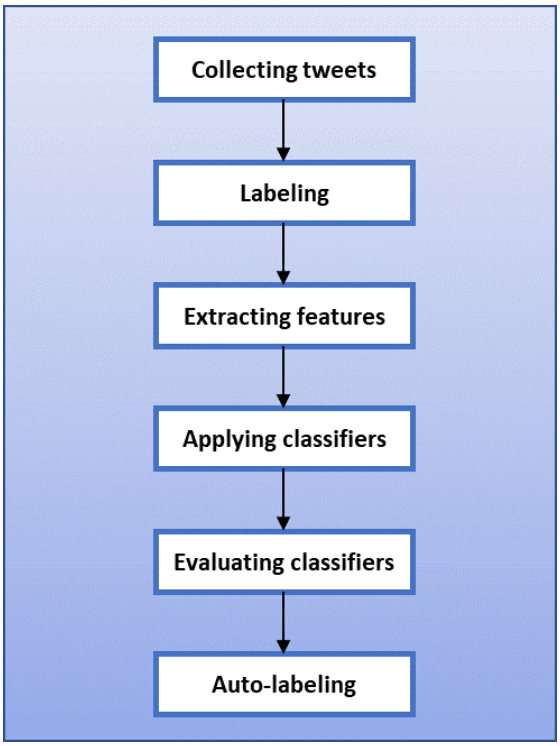

Figure 1: roadmap of using tweets

As the contribution of this project, we show that social media reflects the real word heatwaves. We designed a system with a high accuracy in f-measure (over 90%) to label tweets automatically. The result of our classifications provides a dataset of heatwave related tweets that can be used by other researchers.

The content of this thesis is structured as follows. A background about Twitter and classifiers is provided in chapter 2. In chapter 3, previous studies on social networks are discussed. In chapter 4, we explain our approach for collecting tweets and employed techniques for data preparation, analysis, and evaluation. Then, we discuss about the outcome of data analysis. Finally, we explain the conclusion in the 5th chapter.

CHAPTER 2

2. BACKGROUND

In this chapter, a brief review on our selected social media, Twitter, and heat wave is provided. It is followed by a brief definition of eight classifiers that we used in this project.

2.1.Twitter

Among all social networks, we decided to gather information from the popular social network, Twitter. In contrast to most social networks that tend to hide the information and keep the communication private by default, Twitter communication is public and searchable. This is one of the most significant characteristics that motivates researchers to choose Twitter over the other social network platforms for their studies.

Twitter is a microblog service that was founded in 2006 [10]. Microblog is a combination of blogging and short messaging systems. It is a platform that empowers users to send brief instant messages to many online users. Twitter has grown fast in recent years and been transformed to one of the most well-known online social networks.

Today, Twitter has over 328M monthly active users that includes 68M of US users [11]. Statistics on Pew Research Center shows that 24% of online adults use Twitter and a notable fraction of them (21%) are from US. The countries with the most number of Twitter accounts are respectively USA, Brazil, Japan and Mexico. Pew Research Center has a report containing the result of a demographic research on Twitter [12]. It includes the information about age ranges of adult Twitter users. As it is illustrated in Figure 2, 36

percent of Twitter users are between 18 and 29. Also, there are 23 percent between 30 to 49, 21 percent between 50 to 64, and 10 percent over 65 years old.

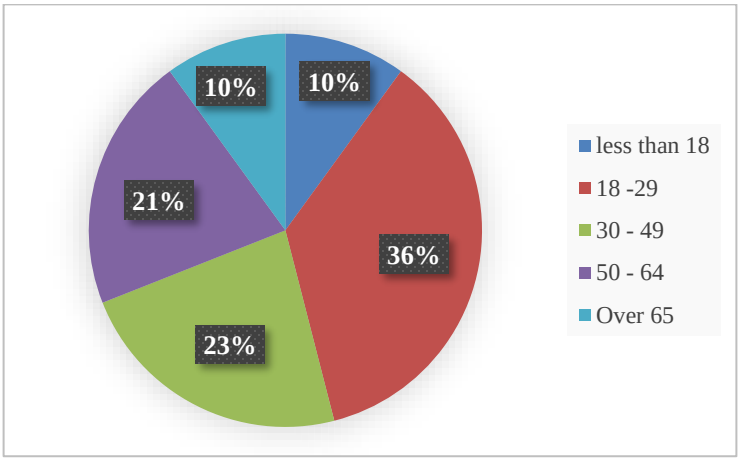

Figure 2: age distribution of Twitter audience from Pew Research Center

It also describes that most of Twitter users are among higher educated people (25% with some college degree and 29% over college degree). Another statistic, which is shown in Figure 3, claims that Twitter users from urban areas are 26%, from suburban 24%, and from rural areas 24%.

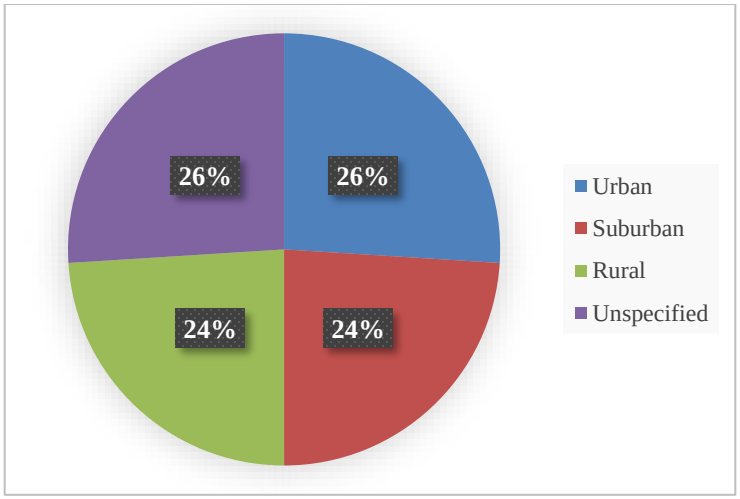

Figure 3: region distribution from Pew Research Center

Each posting message on Twitter is called a tweet. A user can send 140 characters in each tweet (Recently, it has been increased to 280 characters for some users). They can choose to post tweets in public or protected modes. Protected tweets are visible only for user's followers. A user follows other users to get updates about their tweets. Unlike other social platforms, the connection on Twitter is not reciprocity. It means a user can follow another one without being followed back. Twitter provides an easy tweet search by specifying the language, time, location and some other filters. There are three markup vocabularies in tweet language. A user can re-share any desired tweet to spread the word; this ishave called retweet (RT). A @ symbol followed by a username is used to address its user, and finally a # symbol at the beginning of a word creates a hashtag [13]. Twitter enables users to share their idea, run a poll, share photos or videos, share their locations and add emojis in composing each tweet (Figure 4).

Figure 4 : an example of a tweet

Every day 500M tweets are posted. Users take the benefit of using Twitter in various aspect of their lives. The result of an online survey shows that 86% of users "use Twitter for news" [14].

In this thesis, we would like to investigate tweets that contain #heatwave as their hashtag.

2.2. Heatwave

Extreme heat is a natural phenomenon that increases the rate of mortality, disease, violence, drought, wildfire (forest fire) and tornado. It decreases the production in livestock and agriculture. It causes huge consumption of electricity and water sources. It also leads to problem in transportation and constructions [15].

There is a high rate of mortality and damage cost of heatwave events (Table 2). Each year between 170 to 690 death of heat-exposure is reported only in the USA [16]. Some of the deadliest heat wave disasters around the world are listed here:

Table 2: the impact of heatwave on mortality

Year	Fatalities	Affected Area
1995	> 1,000	Midwest of US
1999	> 300	Midwest of US
2003	> 70,000	Europe
2010	> 20,000	Russia

Heatwave does not have a standard definition. It can be interpreted in different ways by considering the days in effect, temperature metrics of a region, threshold temperature, and the impact of the humidity [16]. For example, the exceeding of a temperature degree to 85F in 3 consecutive days can be interpreted severe weather forecast in the north of US but it is recognized as normal weather in southeast.

Researches show that the heat wave is getting more severe annually. Scientists extracted heatwave information of US cities in a period of 50 years and they concluded that heat wave is increasing significantly in frequency, duration, season and its intensity [16].

"Ready Campaigned" [17] has listed extreme heat as a topic that public needs to be educated about because of its impact on human health. "Ready", launched on 2003, is an official website of the department of Homeland Security to prepare and educate Americans for emergency disasters. It provides information about methods to receive alerts and responses properly. They also created a Twitter account (@Readygov) to inform Americans in term of disasters.

The US National Science Foundation (NSF) has also granted many projects to study solutions to decrease the impact of heat exposure, principally in urban areas. As an example, a group of researchers did a broad exploration on Brookline (a city in Massachusetts, USA). They examined many factors of the city and its residents. Through analyzing the historic data, they provided a risk map prediction of Brookline in the benefit of public health [18].

2.3. Classifiers

Following is a short description of the algorithms that we used both for training and testing our data (mostly collected from online material of NLP group of Stanford University [19]).

2.3.1. Bayes Network

Bayes network is a probabilistic classification based on Bayes Rule:

$$Bayes\ Rule: \quad P(x|y) = \frac{P(y|x)\ P(x)}{P(y)}$$

In text classification, we are dealing with a set of documents that each one is assigned to a class. Therefore, the probability of class c given document d is calculated as:

$$P(c|d) = \frac{P(d|c)\ P(c)}{P(d)}$$

$P(c)$ is the probability of class c, called as prior:

$$prior: \ P(c) = \frac{N_c}{N_d}$$

N_c is the frequency of c and N_d is the number of documents.

$P(d|c)$ is the probability of d given c, called as likelihood, and $P(d)$ is the probability of d in the training set.

For a given document d, Bayes Network maps d to class $\hat{c}$ if $\hat{c}$ has the highest probability of $P(c|d)$ among other classes:

$$\hat{c} = argmax_{c \in C}\ P(c|d) = \ argmax_{c \in C}\ \frac{p(d|c)P(c)}{P(d)}$$

P(d) has the same value for all classes, so we simplify formula by dropping P(d). We know that a document d is a set of features:

$$d = \{f_1, f_2, \dots, f_n\}$$

Thus, the formula to find the class of d using the Bayes Network is:

$$\hat{c} = argmax_{c \in C}\ P(f_1, f_2, \dots, f_n|c)P(c)$$

2.3.2. Naïve Bayes

This popular text mining classifier creates the training model fast and it is efficient when CPU and memory are limited. Naïve Bayes is a simple form of Bayes Network. It naively assumes $P(f_i|c)$ are independent probabilities.

$$P(f_1, f_2, \dots, f_n|c) = P(f_1|c).P(f_2|c).\dots.P(f_n|c)$$

Thus, Naïve Bayes finds the class of given d using this formula:

$$c_{NB} = argmax_{c\in C} P(c) \prod_{f\in F} P(f|c)$$

2.3.3. Multinomial Naïve Bayes

In Naïve Bayes classifier the distribution of features is unknown. Multinomial Naïve Bayes indicates that $P(f_i|c)$ is a multinomial distribution. Assume that each feature represents the existence of a word, then the likelihood is:

$$P(f_i|c) = P(w_i|\, c) = \frac{count(w_i, c)}{\sum_{w\in V} count(w, c)}$$

Think of a document which has a word that does not exist in one of the classes then the likelihood $P(w_i|c) = 0$ and using the Naïve Bayes classifier, the probability of class c given document d would be zero without considering the effect of other words in the document. This problem would be solved by using add-one smoothing (also called Laplace technique) that changes the formula to:

$$likelihood: \;\; P(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w\in V}(count(w, c) + 1)} = \frac{count(w_i, c) + 1}{(\sum_{w\in V} count(w, c)) + |V|}$$

2.3.4. C4

C4 is an implementation of decision tree ID3 that is commonly used for Natural Language Processing (NLP). The structure of a simple decision tree is a graph with a root node, branches, inner nodes and leaf nodes. The root node tests an attribute and based on the result, it goes to one of two (or more) branches. Next node is another control unit and a similar process would be repeated. A leaf node represents a class in the dataset. Decision tree algorithm can be implemented by nested if-else statements.

In training process, C4 selects those attributes for nodes that have the highest impact on splitting data into subgroup. And some less important attributes are ignored in the implementation of this tree (pruning). C4 handles missing values by assigning the most common value of the class.

2.3.5. Random Forest

Random Forest is an ensemble model of decision tree models. That is, a set of decision trees are trained independently, and a combination of all results would be used to determine the class. There are 2 randomness in training each tree: using Bootstrap aggregating technique, a randomly selected subset of training data is used; and for splitting each node of the tree, a random subset of predictors is considered.

2.3.6. KNN

KNN stands for K-Nearest Neighbors. It is a lazy classification (training happens each time at the time of prediction). K represents the number of neighbors to consider. To classify a

new input, KNN looks inside the training set and finds k closest instances to the new one; then, by considering the class of nearest ones, it selects the class of the new instance.

2.3.7. SVM

SVM or Support Vector Machine, is another discriminative classification that finds an optimal hyperplane to distinct instances of different classes. In SVM, some part of training set that are in decision boundary (called support vectors) are important and the remaining would be ignored. When classes are not separable linearly, a function is used to map data to N-Dimensional space. This function is called Kernel function.

2.3.8. Multinomial Logistic Regression

Multinomial Logistic Regression (also known as Maximum Entropy) is a discriminative classifier. It means, it learns which features of input should be used for distinguishing better among classes.

CHAPTER 3

3. RELATED WORK

The growth of using social networks, like Twitter, attracts many researchers to use these platforms as their data source in various disciplines. In this chapter we want to review several previous works in social networks related to disasters.

It is well known that people tend to share their feelings and thoughts when they are experiencing a tough situation. Free access to Twitter and its large number of audience, besides the fact that it provides large-scale data in real-time, makes it a powerful tool for sharing feelings and communicating news for individuals in crisis situations. Therefore, it can be a rich crowd source information for analyzing various aspects of a situation.

Nevertheless, analyzing text of tweets is more challenging in comparison to some other NLP projects. Tweet messages are short and posted often by people of diverse knowledge and cultural background. This leads to a large variety of slangs, abbreviations, and syntax style in each set of tweets of a language.

There are studies that leverage tweets as a surveillance service in public health [20] and for early detection of epidemic disease like Zika [21, 22] and influenzas [23, 24].

The growth of Zika virus feared many people in 2015. Therefore, a Twitter's live chat was held by The US Centers of Disease Control and Prevention (CDC) to answer questions among United States citizens. Glowacki et al [21] collected users' messages during CDC's live chat. They applied text mining techniques to assess the chat. They reported the most notable worries about Zika in messages.

Stefanidis et al [22] studied the global attention to Zika in 2015 on Twitter for 12 months. They demonstrated the propagation of Zika related tweets based on geographical locations, users' interactions with official announcement on Twitter and global involvement. They found out social media reflected the spread of Zika over the time in the real world. Their results showed how official announcement successfully attracted the public attention to this epidemic on social media.

Dredze et al [20] introduced a web service for public health officials to practically use the benefit of achieved knowledge of Twitter in public health, named HealthTweets.org. It randomly collected a sample of tweets and categorized them into health related and unrelated. A group of annotators was used to automatically find the trends in health-related tweets. Then data would be normalized based on the location of tweets. As a continuing study [23], they used the website to collect data of influenza in ten English speaking countries for three seasons of influenza to assess the findings of their system with official health announcement in these countries.

Grover et al [24] put the step even further and proposed an Epidemic Hint Algorithm for flu to determine the stage of epidemic using tweets. They used three Bag Of Words (BOW) and calculated the score hint for beginning, spread and decay stages. They claimed their method was more accurate than previous works.

In another study, Aramaki et al [25] collected flu-related tweets over a year. They used tweets of one month and trained a model using SVM algorithm. Then they applied it to classify remaining tweets in two categories: tweets about actual patients, general information regarding flu. The high accuracy of results show that their method can predict the widespread flu earlier than traditional methods.

In natural disasters like earthquake, minutes are crucial to rescue affected people. Studies over the experience of Japan Tsunami [26] and China earthquake [27] indicate that Twitter was one of the most helpful communication channels to get real-time information.

Li et al [27] researched tweets posted about China earthquake of 2008. They studied the quality of Twitter in terms of accuracy, accessibility and completeness of tweets in critical situations. Their conclusion emphasized on the importance of using Twitter as an effective tool with a low bandwidth in time of disasters.

Acar et al [26] sent out a survey about using Twitter to a random sample of people who experienced the Japan Tsunami in 2011. The reliability of tweets was mentioned as the biggest flaw. Nevertheless, many survivors claimed that Twitter played a significant role for communication during the disaster (in some cases the only accessible tool).

Mendoza et al [28] addressed the reliability issue of tweets in disasters in their researches too. They analyzed Chilean users' behavior on Twitter after earthquake through learning users' engagement and activities, the propagation of news, and the most frequent words in each day. They found out there was a low variance of words. They also investigated on characteristics of several fake vs true news that were broadcasted on Twitter after the event. They found out the propagation of truth was almost twice of fake news and there were a significant number of user's who denied or questioned the fake news on Twitter. They suggested to use the different reaction of users as an informative tool to consider the probability of fake news before the confirmation of official sources.

Twitter can also be used for studying natural events like landslides when there are no other sources of information. Natural events like this cannot be fully detected by sensor data; besides, collected sensor data is usually not enough. Musaev et al [29] designed a system

to filter the tweets that were irrelevant to landslides as a natural disaster. In other words, they designed a classification model for landslide related to natural disaster because the term of landslide has multiple meanings. First, they used a list of words and phrases which could categorize some of their collected tweets. Then, they created several training sets from Wikipedia using bootstrapping method. Later, they trained SVM models. Finally, SVM models formed a voting system to classify unseen data.

Tweets labeling is one of the common issues in studies. Manually labeling tweets for training phase can be very difficult and time consuming. The other consideration is that a disaster could be completely new with no labeled data available. Li et al [30] trained classifiers using a set of tweets related to Hurricane Sandy, and applied it on Boston Marathon bombing tweets. They found out that when disasters are similar they can apply the old model to classify tweets on a new event.

Also, Cobo et al [31] developed an auto classification of natural disaster messages on Twitter in the purpose of using it as an information channel for public.

There are also researches on effectiveness use of social media for disaster management of heatwave events. Heatwave that is named “a silent killer” threats residents’ health especially among children and seniors. Furthermore, there are environmental and infrastructural damages that raise the residents’ security risk. Watson et al [8] studied the use of social media during UK heatwave in 2013. Their findings indicated that there were small number of users on social media that effected by heatwave when it is not in severe level. Therefore, still traditional way to spread the warning news was proper. They believed users involvement in propagating the information happened when it was related to them. They suggested to encourage and address the help of young audience in posting crisis

information. The other suggestion for organizations was posting relevant updates on social media.

In a PhD thesis regarding to the impact of heat in social media [32], author collected tweets in 7 weeks of July and August of 2012 in. He used three keywords for data collection: weather, heat and heatwave. Tweets are collected from 6 cities in the USA including Chicago. The result of analyzing data showed that heatwave and heat were used respectively in 5% and 3% of all tweets while in Chicago this number increased impressively (27% and 66%). This happened while information showed that the actual heat severity was not higher than other areas. Therefore, Austin concluded it is the result of the trauma of deadly heatwave events in the past of Chicago.

In this project, we collect tweets using #heatwave. Heatwave is used with different meanings on social media. Therefore, we need a semantic classification to filter out the ones that are related to heatwave as a natural weather disaster. To obtain our goal, we apply eight famous classifiers to build models. Based on the results, we make an ensemble classifier to classify unlabeled tweets. Then we would analyze the most frequent words in each week.

CHAPTER 4

4. DATA ANALYSIS

In this chapter, we explain the steps taken for collecting and analyzing the tweets related to #heatwave. The chapter is divided to three sections corresponding to the main typical steps of a data analysis: data collection, data preparation, and finally training and testing predictive models. Significant points and challenges of each step are discussed in more detail in several organized subsections.

4.1. Data Collection

As the first step for collecting data, we developed a Python program using Tweepy (version 3.3.0). Tweepy is a Python library for accessing the Twitter API [33]. For this research, we collected all Tweets in English with #heatwave as their keyword posted between 9/5/2017 and 11/7/2017. We are interested in original tweets, not retweets. Tweets are stored as Comma Separated Value (CSV) files with 6 features:

- tweet creation date
- the number of retweets
- twitter username
- number of followers
- user location
- the actual text of the tweet

The total number of collected tweets is 6384.

4.1.1. Obstacles and Restrictions

4.1.1.1. API

We had to overcome several challenges with respect to the Twitter API. There are not many documentations about Twitter API [34] to learn its entire usability and limitations. Another restriction of the API for data collection is its limit on keyword search. The free keyword search is limited to 10 previous days, and getting access to the older tweets costs money. Moreover, the Twitter API has a limitation of 180 request within a 15 minutes interval per authenticated user [35]. To overcome these two restrictions in this project, tweets are collected weekly, and a 15-minute sleep thread is used when the number of requests exceeds the quota.

4.1.1.2. User Location

Because of the optional nature of user location, this information might not always be populated correctly by Twitter users.

4.1.1.3. Tweet Content

One of the most challenging steps is structural complexity of tweet contents and difficulties of text mining in order to determine the related tweets. For example, heatwave is a polysemous word. That is, it can be used for unrelated concepts.

The maximum length of a tweet is 140 characters. This limit has motivated more concise phrases and more usage of abbreviations in comparison to other social networks. Consequently, some tweets are too short in the number of words.

Also, many tweets contain a URL (Uniform Resource Locator) address or a picture to convey their ideas. A URL in a tweet is changed to a 23-character link. Hence, it is not easy to gain information from a link. Similarly, retrieving information of a picture is challenging and not the focus of our project.

Taking into account the complexity of text classification of tweets, over 1000 collected tweets in our project were observed manually and categorized into related or unrelated tweets. As it was already explained, sometimes there are not many words to recognize the concept of a tweet and a notable number of tweets are too ambiguous to be categorized by high confidence. Therefore, they are left aside from training set.

4.1.2. Data Overview

As part of data analysis, it is good to have an insight on different characters of the collected data. Figure 5 displays the weekly distribution of the collected tweets in 9 weeks from September 5 to November 7.

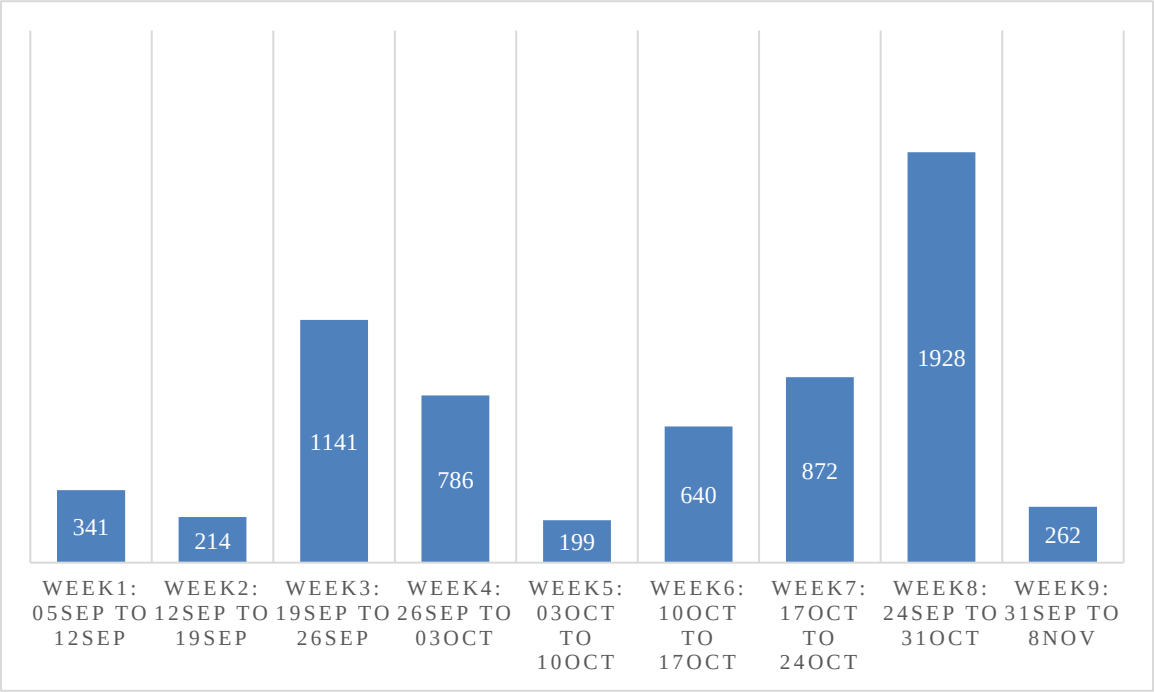

Figure 5: Weekly distribution of the collected tweets in 9 weeks

In comparison, the daily distribution of tweets with #heatwave is illustrated in Figure 6. While the average number of tweets per day is 99.73 and its median is 37.5, the highest number is on 24-Oct with 1392 tweets.

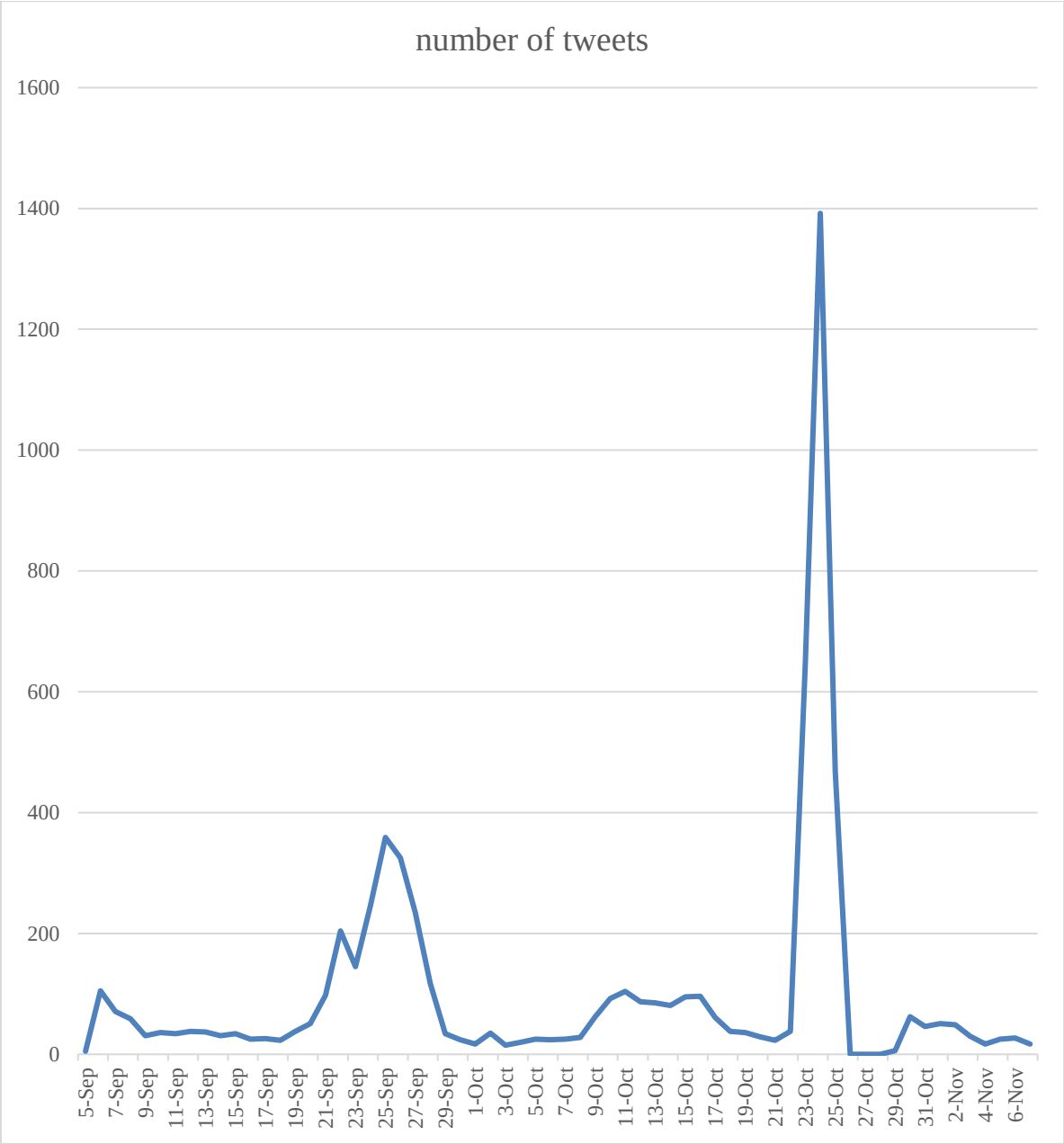

Figure 6: Daily distribution of Tweets

With respect to labeling the tweets to related and unrelated, we randomly picked 1100 tweets, in 11 batches of 100 tweets, to be manually labeled by a group of volunteers. We asked them to write T as a label for tweets that have used heatwave as a climate change and write F as a label for those that referring to other meanings like music bands, games

etc. Then labels are verified by one person. The result is 745 T, 182 F and 173 “N” label when the text is not clear enough for classification. After several passes of try-and-error, we decided to treat the tweets labeled N as noise and discard them from our training data. We asked the volunteers to provide 173 tweets with F label as replacement. Finally training set has 1100 instances which contains 745 T and 355 F labels.

4.2. Data preparation for analysis

There are several preparation actions that should be performed before we can feed the data to machine learning algorithms. Here, we summarize all these preparation procedures that has transformed the raw text of the tweets to a suitable data set for further processing.

For data cleaning, all URLs in tweet texts are replaced with place holders “URL”. Also, UTF-8 code of emojis are removed from the texts. Moreover, double quotations and commas were removed. Finally, the left text of each tweet is placed in a pair of double quotes.

4.2.1. Data Cleaning

We use Weka [36] (a data mining free software) as our text mining tool. Twitter’s actual text and gold labeling are used as two attributes for text mining. We transform data to an ARFF (Attribute-Relation File Format) file and add the proper header at its beginning (Figure 7). Now the ARFF file is qualified to be used by Weka.

```
@RELATION heatwave

@ATTRIBUTE   usertweet STRING
@ATTRIBUTE   relatedtweet {T,F}

@DATA
```

Figure 7: ARFF header

4.2.2. Attribute Extraction

We developed code for text mining by using Weka library.

In the first step, we need to break strings into words to use them as attributes. Unsupervised string-to-word-vector algorithm in Weka extracts attributes from tweets.

In preparing filters to apply string-to-word-vector, we select TF-IDF. They stand for Term Frequency and Inverse Document Frequency. They are useful to indicate the importance of words in given documents [37]. Lovins is used as stemmer. It is an algorithmic stemmer that removes suffixes from words to gain its root [38]. We compare the attribute result of several stop-words. We finally use Rainbow stop-word handler to remove stop-words. It is a list of 526 words [39]. Output-word-counts is set True. We apply two different tokenizers: N-Gram and Alphabetic Tokenizer to find the one that gains better result.

After applying the filters, we have a list of words/phrase. Then, we use feature selection by applying CFS-subset as evaluator and Best First search method. CFS (Correlation-based Feature Selection) uses a search algorithm to evaluate the worth of each attribute in classification [40].

4.3. Model Training & Test

All the previously introduced classifiers are used to train and build models. We use 10-fold cross-validation as a reliable method to train and test data.

4.3.1. Comparison of eight Classifiers

We compare the result of classifiers by using precision, recall, accuracy and f-score [41]. Table 3 represents the result of training and testing 1100 tweets for each classifier algorithm that includes Bayesian Network, Naïve Bayesian, Multinomial Naïve Bayesian, J48, Random Forest, IBk, Logistic and SMO. All classifiers are briefly explained in Chapter 2. Consider that SMO (Sequential Minimal Optimization) in Weka is the equivalent of SVM method, tree classifier J48 is an implementation of C4 and KNN is called IBk (Instance Based learner) in Weka.

4.3.1.1. Alphabetic Tokenizer

Alphabetic tokenizer splits words to meaningful segmentation by removing any non-alphabetic character from words [42]. The result is 59 selected attributes. Then, we train and test tweets with different classifiers. A visual comparison is provided in Figure 8.

Table 3: The Result of Alphabetic Tokenizer

Classifiers	%Precision	%Recall	%Accuracy	F-Score
Bayesian Network	78.66	99.46	81.36	87.85
Naïve Bayesian	**95.59**	66.98	75.55	78.77
Multinomial Naïve Bayesian	86.31	**99.87**	**89.18**	**92.59**
J48	72.30	**99.87**	74.00	83.88
Random Forest	85.19	99.60	88.00	91.83
IBk	85.71	**99.87**	88.64	92.25
Logistic	85.80	99.73	88.64	92.24
SMO	85.62	99.87	88.55	92.19

Comparing all applied classifiers, Multinomial-Naïve-Bayesian has the best accuracy with 89.18%, and the highest f-score with 92.59 among other classifiers. Note that the f-score over 80 is considered very high in terms of semantic classifications of tweets [43]. In terms of Precision, Naïve Bayes has the best result with 95.59% but it is a poor classifier in accuracy (as low as 75.55%) and recall (66.98%).

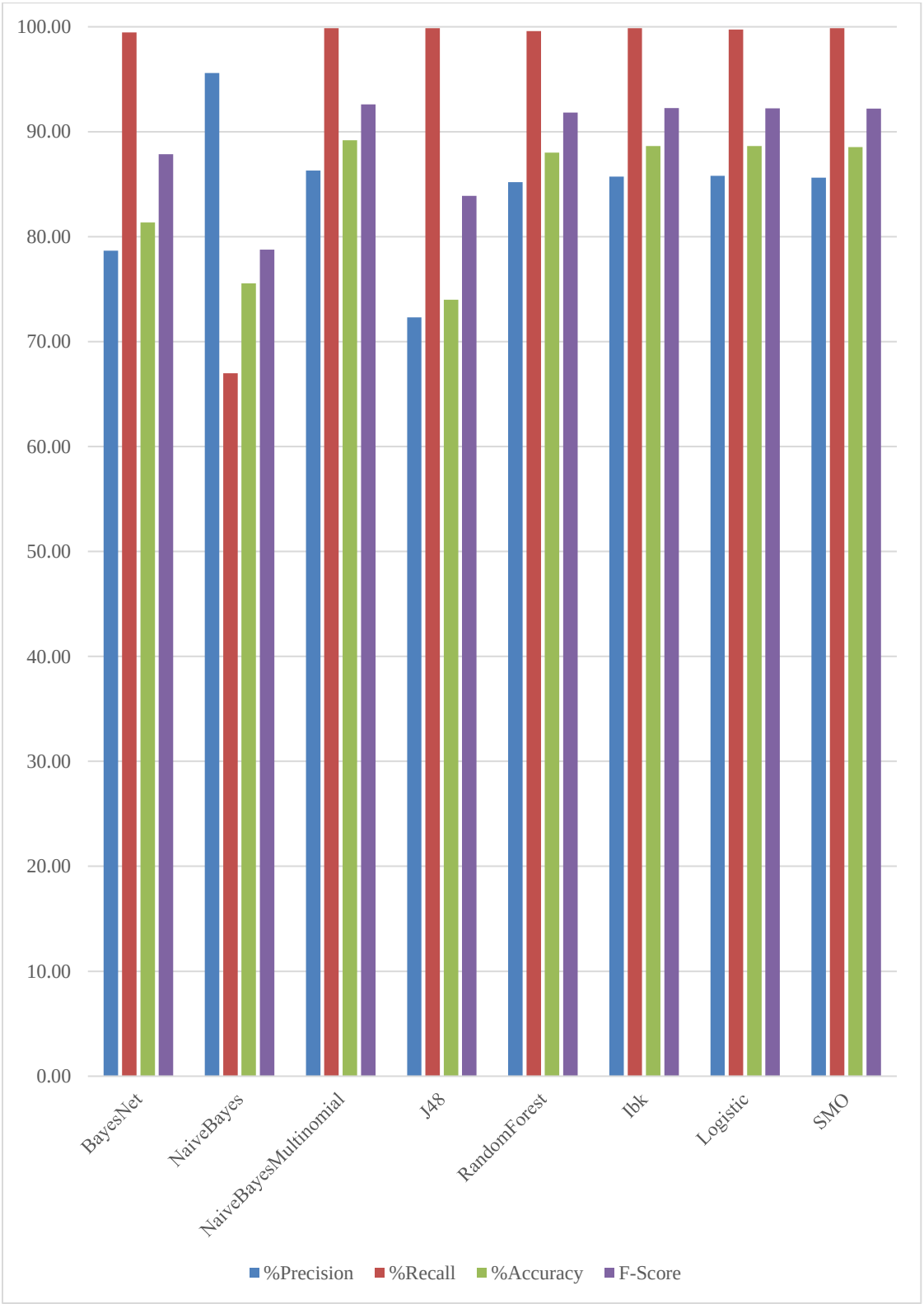

Figure 8: The Comparison of Classifiers with Alphabetic Tokenizer

4.3.1.2. N-Gram Tokenizer

N-gram Tokenizer splits string into words or a sequence of words with size N.

We set the minimum value of N as1 and maximum as 3. Finally, 69 attributes are selected as features. Table 4 represents the result of train and test. Figure 9 is the visual comparison of results.

Table 4: The Result of N-Gram Tokenizer

Classifiers	%Precision	%Recall	%Accuracy	F-Score
BayesNet	78.69	99.60	81.45	87.91
NaiveBayes	80.48	99.60	83.36	89.02
NaiveBayesMultinomial	**87.40**	99.60	**90.00**	**93.10**
J48	69.95	**100.00**	70.91	82.32
RandomForest	85.27	99.46	88.00	91.82
Ibk	86.38	99.60	89.09	92.52
Logistic	86.87	99.46	89.45	92.74
SMO	86.65	99.33	89.18	92.56

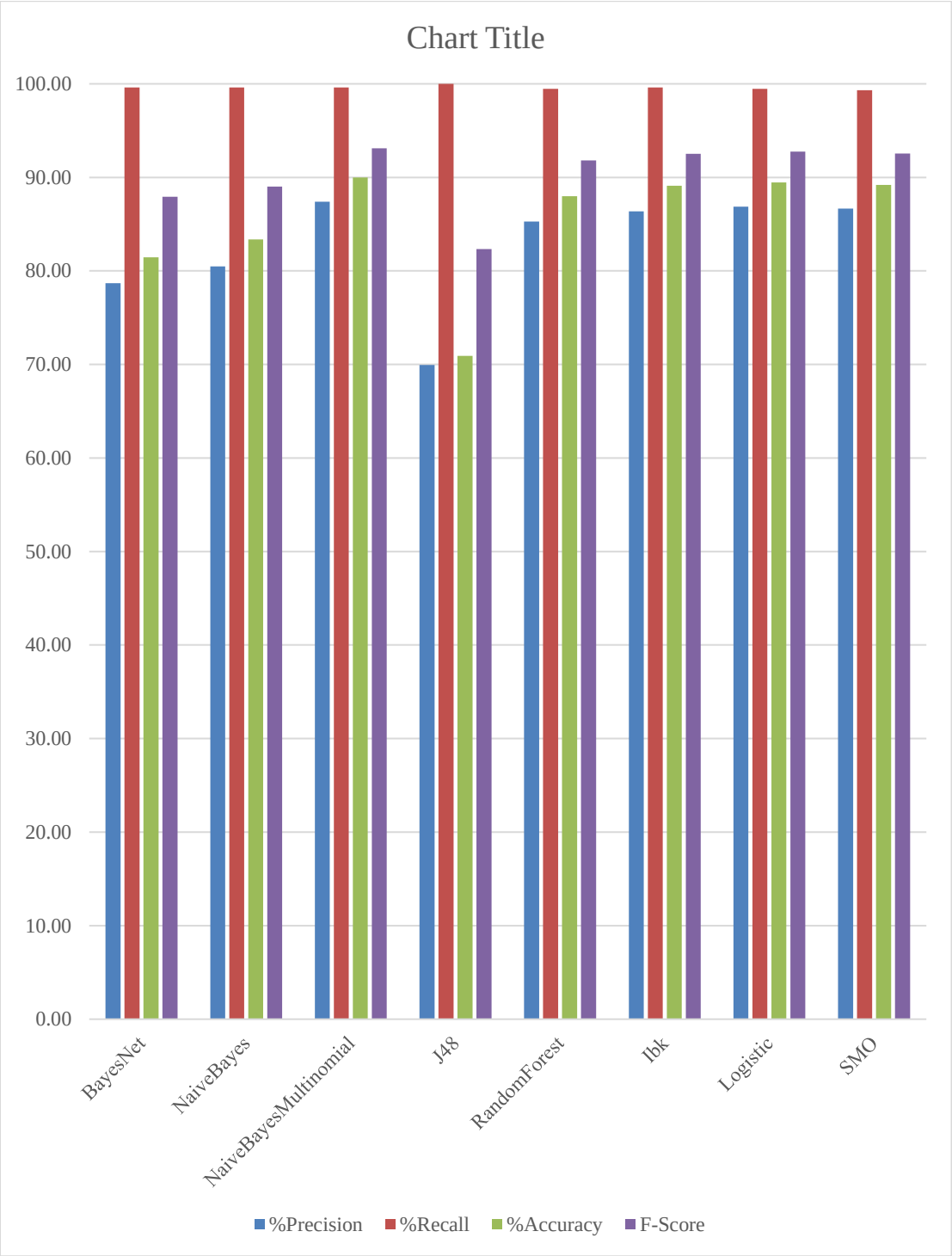

Figure 9: The Comparison of Classifiers with N-Gram Tokenizer

The result is slightly different from using Alphabetic Tokenizer. Recall for J48 classifier is improved to 100%. Multinomial-Naïve-Bayesian with 90.00% has the best accuracy and 93.10 is the highest f-score among other classifiers. In terms of Precision, Naïve Bayes is dropped to 80% and the best result with 87.40% belongs to Multinomial-Naïve-Bayesian.

4.3.2. Classification of unlabeled data

We calculate the average of three best accuracy in two groups of using Alphabetic and N-Gram Tokenizer. The three best accuracy in first group are from Multinomial-Naïve-Bayesian, Logistic and IBk. The average is 88.89%. The average of three best accuracy in second group (Multinomial-Naïve-Bayesian, Logistic and SMO) is 89.55%. Therefore, we choose N-Gram over Alphabetic Tokenizer for classification.

We decided to pick three classifiers with the best results in precision and three bests in recall. Then we use the aggregation of their votes to design a system for classifying unlabeled data.

Multinomial-Naïve-Bayes is on top of the list in term of precision. Then we selected Logistic and SMO. These three forms our “T” labeling system.

J48, Multinomial-Naïve-Bayes and Naïve Bayes algorithms formed the “F” labeling system.

Each of these has one vote and a new tweet can be categorized based on the vote of majority. An unlabeled tweet would be received in “F” labeling system. Each of three selected classifiers have one vote to classify it as “F” or not. If the majority vote on “F” then tweet would be labeled “F”. Otherwise, tweet would remain unlabeled and would be sent to “T” labeling system. Figure 10 illustrates the mechanism of “F” labeling system.

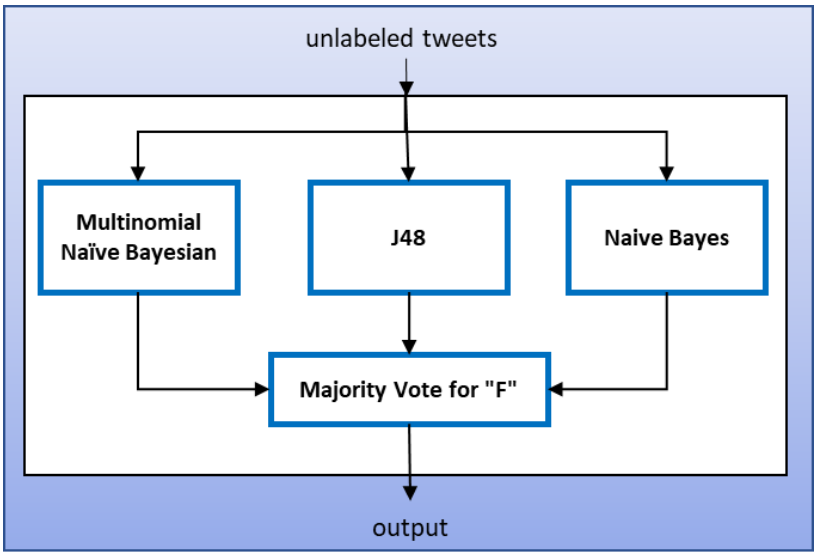

Figure 10: Decision Mechanism for "F" labeling

As explained, the output of “F” decision is F-labeled or unlabeled tweets. The unlabeled tweets are sent to decision system for “T” labeling. Similar to the previous phase, each of three classifiers has one vote. The vote of majority would determine whether a tweet is in “T” category or not. Therefore, the output of this step is T-labeled or unlabeled tweets. The complete process of labeling is illustrated in Figure 11.

It is proper to add all labeled tweets to training set and repeat the process. However, the result of our labeling shows that all tweets are labeled in one run.

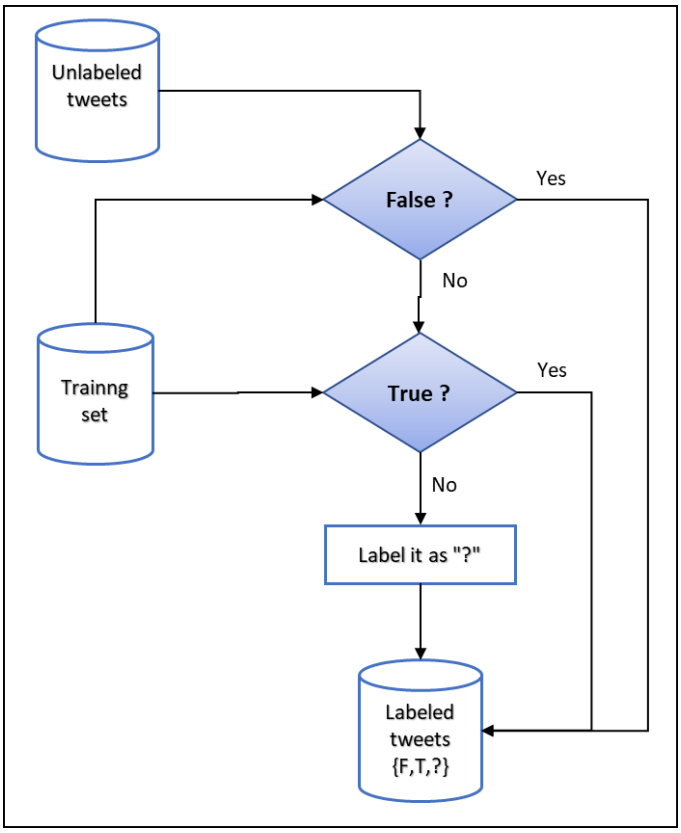

Figure 11: The Labeling Process

Then, we randomly selected 100 T and 100 F off classified tweets. We masked labels and asked a volunteer to read and label them. Table 5 shows the result. Based on this table, the calculated value for Precision is 68.22%, Recall is 73.00%, Accuracy is 69.50% and F-Score is 70.53%.

Table 5: Output Evaluation

	True (Hand labeled)	False (Hand labeled)
True (System output)	73	34
False (System output)	27	66

4.4. Final Results

All collected tweets are now classified into related or unrelated. The weekly frequencies of categorized tweets are illustrated in Figure 12. Related labeled tweets (Series1) is majority which means we can trust when there is a huge increment in number of tweets with keyword #heatwave, users are tweeting about weather.

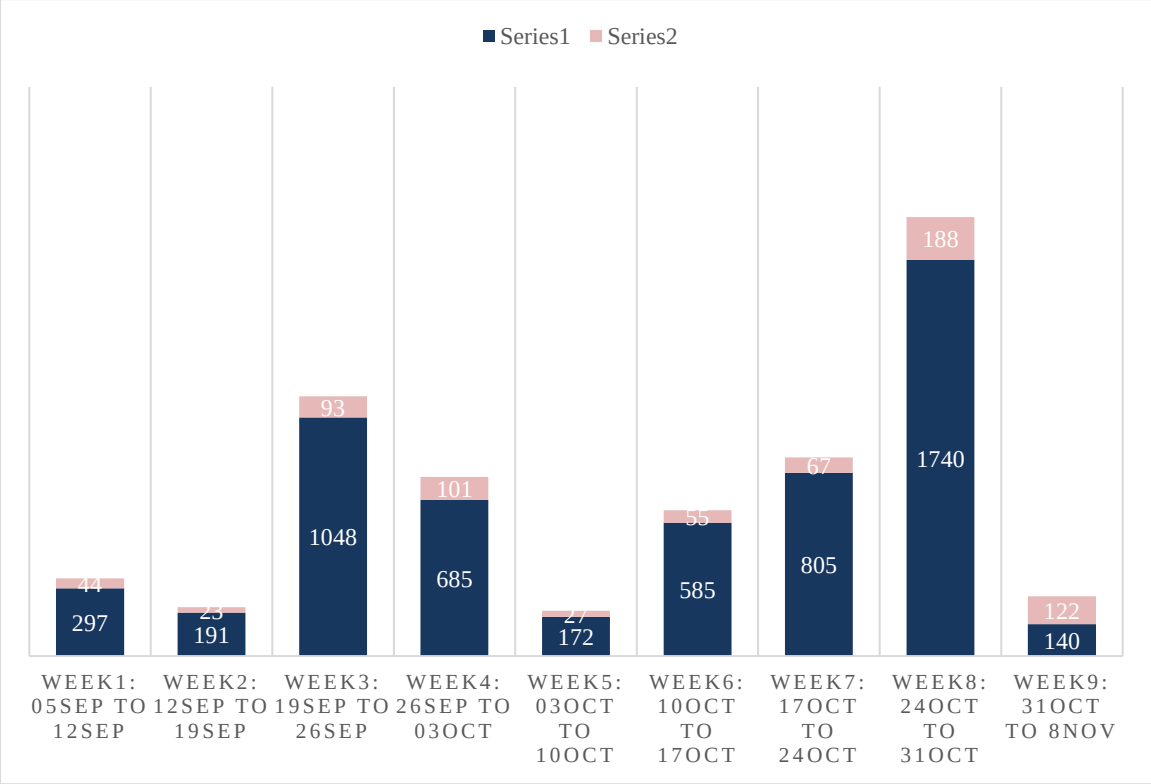

Figure 12: Series1= related Series2= unrelated

Excluding the stop-words, following is the 20 most frequent (over 125 times repeated) words in related tweet category:

{california, cool, day, degree, fall, heat, heatwave, hot, la, losangeles, october, september, socal, stay, summer, today, toronto, weather, week, wildfire}

The word frequencies indicate that there were heatwaves in Toronto and California, more specifically in South California and Los Angeles in September and October. The other interesting information which can be gained through word frequency is that many of tweets are related to wildfire where heatwave was the cause.

Figure 13 shows the daily distribution of related tweets. There is a pick on 24th October with 1268 tweets and a local maximum in 25th September with 334 tweets.

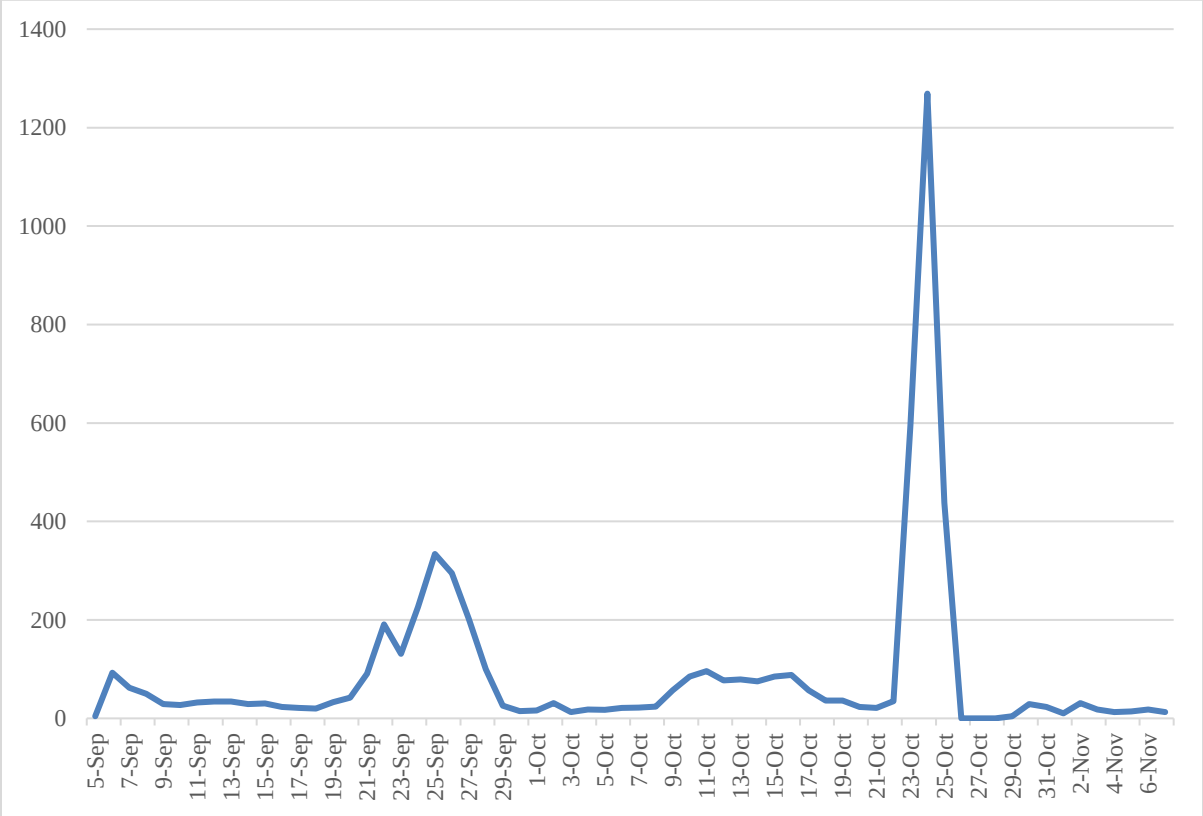

Figure 13:Daily distribution of Related Tweets

Following is th10th most frequent words on 23-25 of October:

{california, degree, fall, heat, heatwave, hot, la, october, today, weather}

Referring back to Table 1: Heatwave events in the US during our research, the word frequently shows the expected result that most tweets are about the extreme heatwave event in California. It is also concluded that when the level of heatwave is not severe there is no significant increase in number of tweets.

There is no heatwave event recorded around 25th September in Storm Dataset. Therefore, we find the 10th most frequent words on 24-26 of September among 855 tweets to gain more information of this increase:

{cool, day, fall, heat, heatwave, hot, september, summer, today, toronto}

Toronto shows up in this list. We found out The Weather Network reported heatwave in Toronto few days earlier than 25th September. Therefore, there is a high probability that the local pick is related to Canadian region.

CHAPTER 5

5. CONCLUSION

Heatwave is one of the natural disasters with different severity that harm public health specifically among elderly and children. Twitter as a powerful social media is able to help in disaster events from fast broadcasting warnings to make two-way communication service. Although, heatwave is damaging and taking many lives and it is named as “silence killer”, many people do not take it seriously. As you saw in Table 1, heatwaves caused 4 mortalities in the USA in September while they were not even “excessive heat”.

It would be helpful to do more researches on the impact of official warning announcement in social media networks like Twitter.

Therefore, we selected Twitter as our platform to gain information on effects of heatwave on tweets. We used Tweepy API to collect tweets with #heatwave keyword. We collected over 6000 tweets in 9 weeks. Three states of the US and some provinces in Canada faced heatwave during this time.

Since heatwave has several meanings with different usages in tweets, a labeling process looked necessary. There are always large volume of data, but having labeled data is important. Therefore, about 1000 of tweets were labeled manually. We used it to train models to make an automatic classification for unlabeled tweets. We applied eight popular classification algorithms that includes Bayesian Network, Naïve Bayesian, Multinomial Naïve Bayesian, C4, Random Forest, KNN, Logistic, and SVM. Also, we used 10-fold cross validation technique to assess our models. Moreover, we calculated recall, precision, f-score, and accuracy for each trained model. Then, we came up with a mechanism for

classifying the rest of our unlabeled data. We used three classifiers with highest value of recall to detect unrelated tweets and three classifiers with the best result in precision were selected to determine related tweets. Tweets after labeling can be added to training data to train new models to classify the remained unlabeled ones.

After labeling all the collected tweets, we examined the related ones. We illustrated the daily frequency of heatwave tweets in a chart. Looking at the official reports on heatwave events in the US, we found a correlation between severe heatwave event and the number of posted tweets classified as relevant. We found the 10 most frequent words during picks; they could clearly relate to the official reports about heatwave. It can be concluded that Twitter reflects the heatwave disaster in real world.

For the future work, it helps to collect the location of tweets instead of the location of users which provides more accurate information.

As a future work, users' emotions on Twitter in response of heatwave can be also studied. The data of comments and likes of posts can be also collected for further research. We collected tweets during the fall, when many people assumed there would be no heatwave. The project can be extended by collecting tweets during other seasons with more keywords to assess our current classification system.

The practiced techniques in this work and its results can be used in many other disciplines from emergency crisis management and public-health management to psychology and sociology as well.

REFERENCES

[1] MNK Boulos, B Resch, DN Crowley, JG Breslin, G Sohn, R Burtner, WA Pike, E Jezierski, KYS Chuang, "Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. International journal of health geographics," *International Journal of Health Geographics,* 2011.

[2] See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., ... & Liu, H. Y., "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. ISPRS International Journal of Geo-Information," 2016.

[3] Mavakala, B., Mulaji, C., Mpiana, P., Elongo, V., Otamonga, J. P., Biey, E., ... & Giuliani, G., "Citizen sensing of solid waste disposals: crowdsourcing as tool supporting waste management in a developing country.," 2017.

[4] Carolin Gerlitz, Eleonora Grotto, Patty Jansen, Lisa Madlberger, David Moats, Irina Papazu, Simeona Petkova, Antonin Segault, Anna K. Skarpelis, Rebekah Tromble, and Arnoud Zwemmer., "Mapping the JD Archive: Fukushima, Twitter and the Politics of Disaster Communications.," wiki.digitalmethods.net. Digital Methods Initiative, 2014. [Online].

[5] Laura R. Walton, Skye C. Cooley, and John Nicholson, ""A Great Day for Oiled Pelicans:" BP, Twitter, and the Deep Water Horizon Crisis Response," *the Fifteenth Annual International Public Relations Research Conference,* 2012.

[6] Cassa CA, Chunara R, Mandl K, Brownstein JS., "Twitter as a Sentinel in Emergency Situations: Lessons from the Boston Marathon Explosions," 2013.

[7] Bruno Takahashi, Edson C. Tandoc Jr., Christine Carmichael, "Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines," *ELSVIER Computers in Human Behavior,* vol. 50, pp. 392-398, 2015.

[8] Watson, H., & Finn, R. L., "Social media and the 2013 UK heat wave: Opportunities and challenges for future events," 2014. [Online].

[9] "NOAA's National Centers for Environmental Information (NCEI)," [Online]. Available: https://www.ncdc.noaa.gov/stormevents/.

[10] N Arceneaux, A Schmitz Weiss, "Seems stupid until you try it: Press coverage of Twitter, 2006-9," SAGE, 2010.

[11] "Statistica," 2018. [Online]. Available: https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/.

[12] "Social Media Update 2016," Pew Research Center, 2016.

[13] Kwak, Haewoon; Lee, Changhyun; Park, Hosung; Moon, Sue B., "What is Twitter, a social network or a news media?," *World Wide Web,* pp. 591-600, 2010.

[14] ROSENSTIEL, TOM; SONDERMAN, JEFF; LOKER, KEVIN; IVANCIN, MARIA; KJARVAL, NINA, "American Press Institute," 2015. [Online]. Available: https://www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-in-general.

[15] Smoyer-Tomic, K.E., Kuhn, R. & Hudson, A. Natural Hazards, "Heat Wave Hazards: An Overview of Heat Wave Impacts in Canada," *Natural Hazards*, p. 465–486, 2003.

[16] Habeeb, Dana; Vargo, Jason; Jr, Brian Stone, "Rising heat wave trends in large US cities," *springer,* pp. 1651-1665, 2015.

[17] "Ready," [Online]. Available: https://www.ready.gov/heat.

[18] Babak J.Fard, Hanieh Hassanzadeh, Mary Elizabeth Warner, Udit Bhatia, Auroop Ganguly, "Mitigation and Adaptation Strategies for Public Health Impacts of Heatwaves for Town of Brookline, MA," 2016.

[19] Martin., Daniel Jurafsky & James H., Speech and Language Processing, 2017.

[20] Mark Dredze, Renyuan Cheng, Michael J. Paul, David Broniatowski, "HealthTweets. org: a platform for public health surveillance using Twitter," 2014.

[21] Elizabeth M. Glowacki, Allison J. Lazard, Gary B.Wilcox, Michael Mackert, Jay M. Bernhardt, "Identifying the public's concerns and the Centers for Disease Control and Prevention's reactions during a health crisis: An analysis of a Zika live Twitter chat," *American Journal of Infection Control*, 2016.

[22] Stefanidis A, Vraga E, Lamprianidis G, Radzikowski J, Delamater PL, Jacobsen KH, Pfoser D, Croitoru A, Crooks A, "Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts," JMIR Public Health Surveill, 2017.

[23] Michael J. Paula, Mark Dredzea, David A. Broniatowskib, Nicholas Generous, "Worldwide Influenza Surveillance through Twitter," 2015.

[24] Sangeeta Grover, Gagangeet Singh Aujla, "Prediction Model for Influenza Epidemic Based on Twitter Data," 2014.

[25] Eiji ARAMAKI, Sachiko MASKAWA, Mizuki MORITA, "Twitter Catches the Flue: Detecting Influenza Epidemics using Twitter," 2011.

[26] Acar, A. and Muraki, Y., "Twitter for crisis communication: lessons learned from Japan's tsunami disaster," *Int. J. Web Based Communities*, vol. 7, pp. 392-402, 2011.

[27] Jessica Li, H.R. Rao, "TWITTER AS A RAPID RESPONSE NEWS SERVICE:AN EXPLORATION IN THE CONTEXT OF THE 2008 CHINA EARTHQUAKE," *The Electronic Journal on Information Systems in Developing Countries*, 2010.

[28] Marcelo Mendoza, Barbara Poblete, Carlos Castillo, "Twitter Under Crisis: Can we trust what we RT?," *1st Workshop on Social Media Analytics (SOMA '10),* 2010.

[29] Aibek Musaev, De Wang, Jiateng Xie and Calton Pu, "REX: Rapid Ensemble Classification System for Landslide Detection using Social Media," 2017.

[30] Hongmin Li, Kishore Neppalli, Nic Herndon, Anna Squicciarini, "Twitter Mining for Disaster Response: A Domain Adaptation Approach," in *ISCRAM*, Kristiansand, 2015.

[31] Alfredo Cobo, Denis Parra, Jaime Navon, "Identifying Relevant Messages in a Twitter-based Citizen Channel for Natural Disaster Situations," in *International World Wide Web Conference Committee (IW3C2)*, Florance, Italy, 2015.

[32] B. J. Austin, "Perspectives of weather and sensitivities to heat: Social media applications for cultural climatology," 2014. [Online]. Available: https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ETD_SUBID:95791.

[33] "Tweepy," [Online]. Available: http://www.tweepy.org.

[34] Katrin Weller, Katharina E. Kinder-Kurlanda, "Uncovering the Challenges in Collection, Sharing and Documentation:," *Standards and Practices in Large-Scale Social Media Research:*, 2015.

[35] "Twitter developer," [Online]. Available: https://developer.twitter.com/en/docs/basics/rate-limits.

[36] [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/.

[37] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in *Proceedings of the first instructional conference on machine learning*, 2003.

[38] Lovins, "Development of a Stemming Algorithm," 1968.

[39] shuson, "rainow," p. https://gist.github.com/shuson/b3051fae05b312360a18.

[40] Mark A. Hall, Lloyd A. Smith, "Practical Feature Subset Selection for Machine Learning".

[41] CD Manning, P Raghavan, H Schütze, Introduction to information retrieval, Cambridge University Press, 2008.

[42] George Forman, Evan Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," in *17th ACM conference on Information and knowledge management*, 2008.

[43] Aibek Musaev, De Wang, Jiateng Xie, Calton Pu, "REX: Rapid Ensemble Classification System for Landslide Detection using Social Media," *IEEE 37th Confrence on Distributed Computing System,* 2017.

[44] William B. Cavnar, John M. Trenkle, "N-Gram-Based Text Categorization," in *3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[45] Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu1 Ryo Hayasaka, Ryuzo Isochi, Zian Wang, "Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter," in *9th International ISCRAM Conference*, 2012.