An Enhanced Model for Disaster Management using Social Network and Machine Learning Techniques

Thesis Submitted in Partial Fulfillment of the Requirements for the Award of

Degree of

Master of Technology

in

Information System

Under the guidance of :

Dr. Rahul Katarya

Associate Professor

Department of Computer Science & Engineering

Submitter by:

Divya Jain

Roll No.: 2K16/ISY/04

Department of Information System

Session : 2016-2018

Declaration

I hereby declare that the thesis work entitled An Enhanced Model for Disaster Management using Social Network and Machine Learning Techniques which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of degree of Master of Technology (Information Systems) is a bonafide report of thesis carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

_____

(Signature)

_____

(Divya Jain)

_____

(Roll No.)

Certificate

This is to certify that Divya Jain (2K16/ISY/04) has completed the thesis titled "An Enhanced Model for Disaster Management using Social Network and Machine Learning Techniques", under my supervision in partial fulfillment of the Master of Technology degree in Information Systems at Delhi Technological University.

Dr. Rahul Katarya
Associate Professor
Department of Computer Science and Engineering
Delhi Technological University
Delhi -110042t certificate

Acknowledgements

Firstly, I would like to thank Dr.Rahul Katarya, Associate Professor at Department of Computer Science & Engineering, Delhi Technological University, Delhi. This work would not have been possible without his constant support and research ideas. I would also like to thank all other faculty members of Department of Information System at DTU Delhi for their continuous review and feedback of the work and the motivation to keep pursuing the research interest.

I would also like to thank my parents and friends who have always motivated me to work hard and has been my constant support in all situations.

Divya Jain
2K16/ISY/04
Department of Information System
Delhi Technological University, Delhi

Abstract

Millions of posts in the form of text, images and videos are being posted everyday on social media and platforms like Twitter, Facebook plays a very critical role in spreading and sharing of the information across the world. Due to this, social media can have an active role to play in the case of natural disasters. During and after natural disasters, a large number of users posts information regarding the disaster and their situation on the social platforms. These platforms are also actively used by the government & other agencies to share critical information. According to a survey, social media ranks fourth most popular source for accessing emergency information. Although due to the lack of right information among the huge volume of the incoming tweets during any disaster, the resources are not properly mapped which leads to further loss of life & property which can be prevented to a certain extent.

Large volume of the data present on social media can be leveraged to access the disaster situations and prepare a management plan accordingly. In this work we study how the data present on social media can be helpful to manage the disaster situation. Efficient analysis of the data can help in preparing a plan for proper information propagation. Most critical part is to mine the right data in a quick time from the huge volume of data posted by the users. So in our first task we plan to classify the tweets obtained during the disaster situation and separate the tweets which are related to the ongoing situation. Doing this reduces the number of tweets to be analyzed further to get the important information.

After the identification of the disaster related tweets, it is important to understand the need and emotions of the user, thus we plan to do sentiment analysis over the tweets and gather the information about the needs of the users which can be used to provide assistance for their recovery.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social media has become an integral part of people's lives in this world where all big and small events are posted online by the people. Online platforms like Twitter, Facebook receives thousands of post every minute and even the minute details of the happening in user's life can be seen from these platforms. Therefore, online social media platforms, in today's world, are the biggest source of live data and can be possible solution for many problems that poses challenge because of the lack of the data.

In the recent years, several studies have been conducted to analyze the role of social media in tackling the disaster-related situation. In [1], the study is focused on analyzing technical, social aspects of the disasters situations. Several studies has focused on understanding the relationship between social media and the disaster situations. [2] explains how the tweets are posted just after the disaster situation and what socioeconomic factors are important in the prediction of disaster-related tweets.

In this work, we propose a system where in the disaster situations, when thousands of people post on twitter regrading the situation and the current information, this data can be used to identify the people who needs help and map the resources to the needy people at the right time. In our work we have proposed three different tasks that can be performed on the Twitter data to analyze and extract the information from the tweets. With the growing popularity of online platforms, even government agencies are using social media platforms to disseminate information and also people are using social platforms to post urgent request during the times of crisis, as analyzed in [3]. There are various other studies where the engagement of social media in the disaster situation has been studied in [4, 5, 6, 7, 8].

1.1 Overview

In this section, we discuss the overview of our proposed work. We have categorized our work in to three tasks.

1. In the first task, we have applied the machine learning models to identify the disaster-related tweets. In this task we separate the disaster-related tweets from the large volume of tweets. This task has been done using various features in the tweet.

2. In our second task, we have discussed a method based on the machine learning algorithms where we categorized the disaster-related tweets either subjective or objective. This task is important to understand the needs and emotions of the user. More details have been discussed later.

3. For the last task in our proposed model, we have proposed a system to identify the relevant named entities from the disaster-related tweets which will be helpful to map the available resources in the right manner in quick time.

1.2 Major Contribution

In this section, we describe the contribution made by us through this work. This work is important to extract the relevant information from the data posted on social media platforms. This work can be used to identify the entities and map the available relief resources in an efficient manner. Following are major contributions of our work:

- We proposed the method to differentiate the disaster-related data from the unrelated data on the social media during any disaster.

- In this work, we have proposed the method to do sentiment analysis over the disaster-related tweets.

- We categorize the tweets in to subjective and objective categories based on the sentiments expressed in the tweets by the user.

- We propose the method to extract the relevant entities from the available data.

- We recognize the entities like location, persons, organizations, etc.

In this work, we first discuss the similar work done by other researchers to extract the relevant data from the social media in crisis situations. Chapter 3 briefly describes the background terminology required for the better understanding of the proposed method. In chapter 4, we discuss our work with all the details of the model. We describes our results and ends this work with conclusion and future work in chapter 6.

Chapter 2

Related Work

This chapter briefly talks about the research work carried out in the field of disaster management using the social media data. This field is quite new in the field of data mining but still many studies have been introduced explaining the techniques that can be used to extract the relevant data from the data available from social media platforms.

In [9], the authors have presented a way to extract the data from the tweets on Twitter platforms. In their work they focus on extraction of disaster-relevant information from the tweets. Authors in [10] have developed a platform to detect, assess and summaries the data from the social media platform Twitter during any crisis situation. Their tool is developed to focus on crisis coordination and situation awareness and the tool was deployed for the trial in the real world. In [11], Beigi et al have provided an overview on the use of sentiment analysis in social media and how that can be helpful in the disaster-related situations. In their work, the relationship between the sentiment analysis and the social media has been discussed in detail and that relationship provides a chance to better understand the disaster situation through social media.

There have been various studies and surveys which focus on using data from social media platforms to enhance the strategies to counter the disaster-related situations and spread awareness. In [12], authors have focused to analyze the Twitter data posted during disaster situations and study that how natural language processing techniques can be used to formulate strategies. Some relevant approaches in their work were burst detection, which sends an alert when some unexpected event is detected and geo-tagging. Similarly in [13], automatic methods to extract the relevant

information has been discussed using the machine learning classifying models.

Recently in [14], author has done a thorough assessment of the role social media can play in disaster situations.

Interested readers can refer to [15, 16, 17] for more detailed surveys and studies on the role of social media in disaster situations.

Chapter 3

Background

In this chapter we will look at the topics needed to create the background to better understand our work. We will discuss the techniques to pre-process the data and then will briefly describe the different type of features used in our model for the classification tasks. We will also look at the different classifiers and the metrics to evaluate the performance of our proposed method.

3.1 Pre-Processing

This section describes the different techniques used to pre-process the data before passing it to the actual methods. These techniques are important because these techniques basically formats the data to match our needs.

3.1.1 Tokenization

In tokenization, the sentience or the string is broken down in to pieces with the help of some delimiters. The keywords, phrases and the elements received because of tokenization are known as *tokens*. Now, these tokens can further be used for further processes like data mining and many more. Tokenization is commonly performed for white spaces, punctuation marks or some special character.

3.1.2 Text Segmentation

Text segmentation is basically a process of dividing a string or sentence into the meaningful parts. The process of segmentation basically defines the position in text where the topic or the subject changes. So., through segmentation the aim to identify

those positions and divide the text based on the topics or subjects. In our work, we have used the NLTK Python library for the process of text segmentation.

3.2 Features

Bag of Words : It is simply just the representation of the text in form of the words in the text. It creates the vocabulary of the words and represent the document based on occurrences of the words. In this approach, individual or group of continuous words is being used as features, generally called as *n-grams*.

Term Frequency : This feature takes in to account the number of time any term is present in the document, not just only the presence of the term.

$$tf = \frac{Frequency\ of\ term\ in\ document}{Total\ number\ of\ terms\ in\ document} \tag{3.1}$$

Length Features : There are various features such as *average word length*, *average sentence length* and *average length of the review* which gives details about the writing style of the user. All these features are based on the length of the text in question.

n-gram: In the field of text mining and data science, *n-gram* is a sequence of n continuous items from any text. Most commonly used n-grams are unigrams, sequence of single tokens and bigrams, sequence of two tokens.

3.3 POS Tagging

In these type of features, words in the review are being tagged with the part of speech (grammatical tagging) based on its context. Different kind of POS tags such as nouns, verbs, adjectives, adverbs etc., are marked for the role of the words in the text. POS tagging is often done using a POS tagger which is basically a software that assigns the POS tags to every word in the text.

3.3.1 Twitter POS Tagger

For this work, we have not used the standard NLTK POS tagger but have used a special POS tagger implemented in [18]. In this work, the authors have developed a POS tagger specifically for the tweets from Twitter. POS tags for the hashtags,

URL links and also the frequently used abbreviations on the social media have been provided.

3.4 Classifiers

In this section, we discuss the various machine learning algorithms or classifiers used in our work.

3.4.1 Support Vector Machine (SVM)

SVM was first introduced in [19] in 1998. It is the most commonly used classifier and is often regarded as the simplest algorithm in the supervised machine learning algorithms.
Support Vector Machines are supervised learning models that analyzes and classifies the data using associated learning. In SVM, the input variables ie., the extracted features forms a space of n-dimensions, where n is the number of features acquired and then the points are plotted in n-dimensional space. A hyper plane is then selected which is a line that best divides the points according to their class. Though SVM is the most commonly used classifier and is very effective in high dimensional space, it does not works well with the large datasets because of very high training time. **Support vectors** :The vectors (cases) that define the hyper plane are the support vectors as shown in fig. 3.1.

SVM works really well if we plot the training data into multidimensional space and then by using hyper-plane, it tries to separate the data into two separate groups. If the classes are not immediately linearly separable in the multidimensional space then a new dimension will be added by the algorithm or by the model, so that they will be separated into two categories and the same process will be continued till the data is completely separated into two groups, A and B by the hyper-plane as shown in fig. 3.2.

Pros of SVM

- It superbly works with a clear margin of dissociation.
- It has been effective over the high dimensional plane.
- It has been effective if the count of dimensions is bigger than the count of samples.

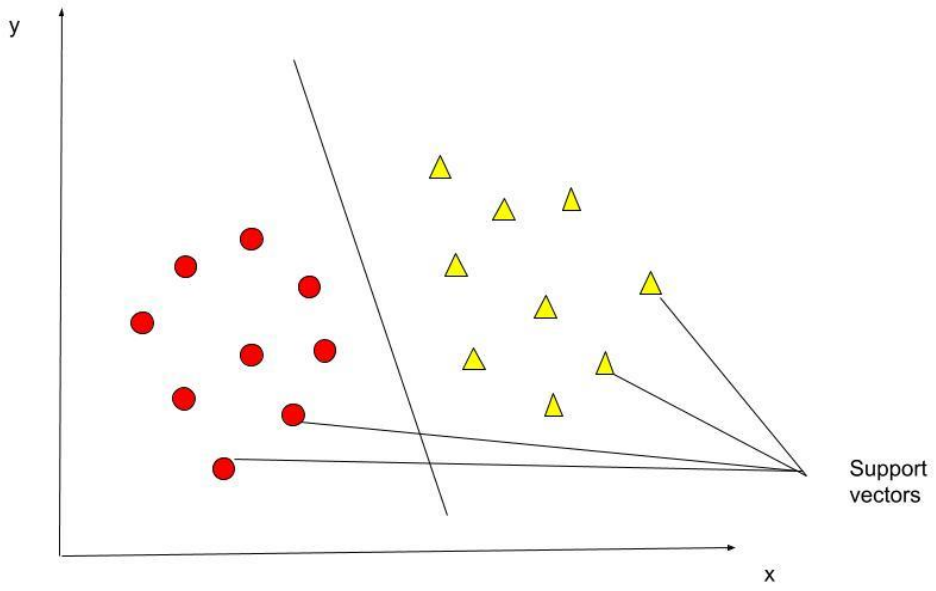

Figure 3.1: Support Vectors of SVM

- A subset of training points used in the decision function (i.e support vectors), that why it is memory efficient.

Cons of SVM

- Its performances are not good if the size of the dataset is large as the time needed to train the classifier will be long.
- Its performance is not really good if the dataset is more noisy .
- Probability estimates are not provided by SVM directly, but can be calculated by the cross-validation method which is quite expensive. It can be done using the Python scikit-learn library which has the Support Vector Classification (SVC) method.

3.4.2 Random Forest

Random forest algorithm was first introduced in 2002 in [20]. The classification through this algorithm is based on the idea of decision trees. More the number of trees in the forest, better results will be obtained. In this algorithm, root nodes and

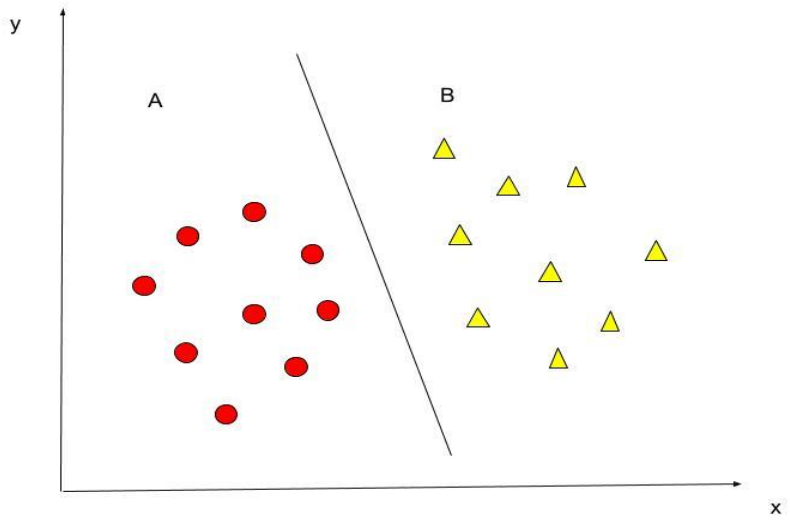

Figure 3.2: Classes in SVM

the decision nodes are decided randomly. Different trees are created by selecting different features from the total features randomly. Then among the selected features, best split node is selected and in order to build the forest, same process is iterated over the total number of trees.

Pros of Random Forest

- It takes the decision very accurately.
- It works very well over the large data set.
- It can be helpful to extract variable importance.
- feature engineering does not need (like scaling and normalization)

Cons of Random Forest

- If the data is noisy then over fitting is there.
- Not like in decision trees, it is difficult to interpret the result.
- To get high accuracy, good tuning is required with hyper parameters.

3.5 Performance Metrics

Before discussing the metrics, we will talk about the basic terminology for the performance metrics for better understanding and then will discuss the metrics to evaluate the performance of our results for different tasks explained in the next chapter.

- **Total Accuracy:** It is the ratio of the classifications which are done correctly (including both true positives and true negatives) to the total number of classifications.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.2}$$

where, TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

- **Positive Predictive Value (PPV):** It is the ratio of the correct positive classifications done by the classifier to the total positive classifications (including the false positives also).

$$PPV = \frac{TP}{TP + FP} \tag{3.3}$$

- **Negative Predictive Value (NPV):** It is the ratio of right classification which are negative to the total number of negative classifications (including the false negative predictions).

$$NPV = \frac{TN}{TN + FN} \tag{3.4}$$

Chapter 4

Proposed System

In our work, we have proposed the method to mine information out of the twitter data. We have categorized our work into three different tasks. The three tasks can be summarized as follows:

1. Classification of disaster-related tweets
2. Sentiment analysis of disaster-related tweets
3. Identifying the important identities from the twitter corpora

In this chapter, we will describe the techniques used and the proposed methods for each of these task.

4.1 Classification of disaster-related tweets

Millions of tweets are posted on the social media platform Twitter on daily basis. The nature of the tweet varies from people sharing their daily life experiences to the government agencies sharing important information with the people and also addressing grievances of the common people. During a disaster situation in any part of the world, a large number of tweets are posted on Twitter in which people provides information of the scale of the disaster, their well-being to their friends and family and also people look for help and try to communicate the relief agencies through their tweets.

Although a large number of tweets are posted related to the disaster situation, but at the same time a even larger number of tweets are also posted for other events happening around the world which makes the disaster related tweets not so visible to the agencies looking to help people. Though people uses hash tags in their tweets which

can be used to search the relevant tweets but we cannot expect people in distress to use the exact hash tags or to use the hash tags at all. Therefore, it is necessary to have some mechanism to identify the relevant disaster-related tweets from the huge twitter corpus.

The objective in this task is to identify and differentiate the tweets which are related to the disaster from the tweets which are not related to the disaster situation. It is important because a huge volume of tweets is posted in Twitter every second and it is almost impossible to go through all the possible tweets that are posted manually.

Therefore, we need a mechanism which can differentiate between these two categories and give back the tweets which are related to a disaster situation. Moreover it will be easier as also the size of the dataset will be considerably reduces thus a lot of time will be saved by only further processing the disaster-related tweets for extracting further information.

4.1.1 Dataset Information

The dataset contains 10,877 labelled tweets including both disaster related and not disaster-related tweets. Value 1 is assigned to the tweets which are disaster-related and value 0 to tweets which are not related to the disaster.
The dataset is provided by CrowdFlower, "Disasters on social media

4.1.2 Proposed Method

In this section, we propose a method of classification to identify the tweets which have information related to a disaster situation.

There are different components in the method proposed for the classification of the tweets as disaster-related or not related to disaster. Fig. 4.1 shows the flow of the proposed method for the identification of disaster-related tweets.

1. At first, we have a labelled dataset of tweets in which the disaster-related tweets are labelled with 1 and the other tweets are labelled as 0.

2. Now POS tagging and the feature extraction is done on the dataset available. Twitter POS tagger discussed in the last chapter has been used for the task.

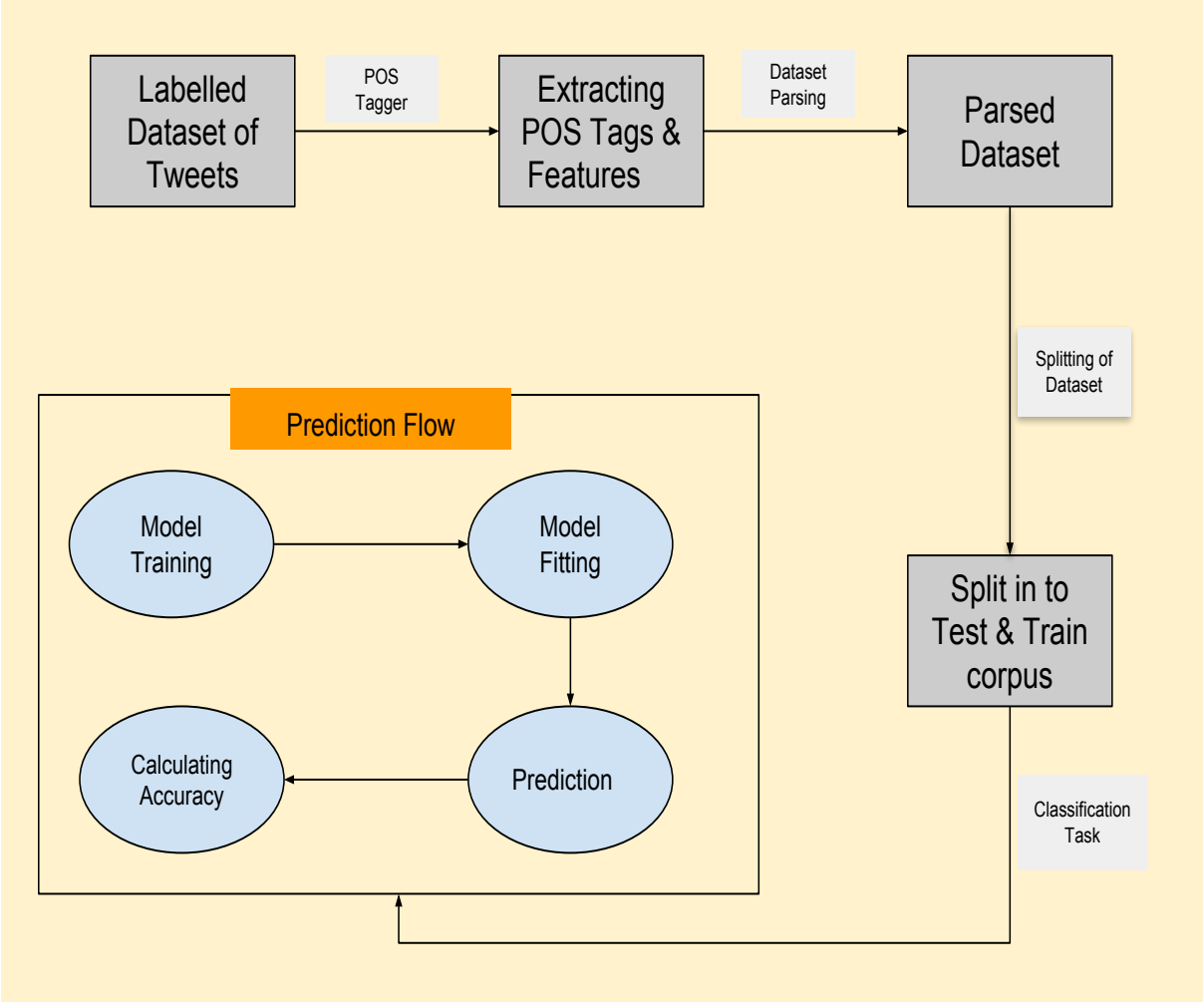

Figure 4.1: Identification of disaster-related tweets

Also several features are extracted for the classification task.The features extracted from the tweets are discussed further.

3. Now, the parsed dataset is split for the training and testing purposes. 90-10 rule is followed for the splitting of the dataset i.e, 90% of the dataset is used for training and rest 10% is used for the testing task.

4. After this the machine learning classification models are developed and fitted for the prediction task.

4.1.3 Features Used:

A variety of linguistic features are extracted and used for the classification task of the tweets into disaster-related and not disaster-related classes. We discuss here a list of all the features that are extracted from all the tweets in the dataset.

1. **POS tags** for tweets are extracted using Twitter POS Tagger, following tags are given to the words :
   - Adverbs
   - Nouns
   - Pronouns
   - Frequently used social media abbreviations
   - Numerals
2. Does tweet contain link and number of links
3. Does tweet contain hash tag and number of hash tags
4. Happy emojis in the tweets
5. **Unigram and bigrams** are extracted from the tweets for the prediction task. We have used a Python library for the tokenization task.

POS Tagging Example :
Tweet in dataset :

"anyone from miami who decided to go to tampa to evacuate #HurricaneIrma https://t.co/EAWvf0KbtX"
POS Tagging of the tweet :

anyone N (Noun) from P (Pre/Post - position)
miami ^(Proper Noun) who O (Pronoun)
decided V (Verb) to P (Pre/Post - Position)
go V (Verb) to P (Pre/Post - Position)
tampa ^(Proper Noun) to P (Pre/Post - Position)
evacuate V (Verb) #HurricaneIrma (Hashtag)
https://t.co/EAWvf0KbtX U (URL or Email)

We have used SVM and Random Forest machine learning classifiers for this task. Both the unigrams and bigrams are separately used for the classification task in both the models. Results of this task are discussed in the next chapter.

4.2 Sentiment analysis of disaster-related tweets

After the identification of the tweets which are related to the disaster situation, the next important and critical task is now to identify the need and emotions of the people who are affected by the disaster situation.

After the classification of the tweets in the dataset, we have successfully identified the disaster-related tweets out of all the tweets and significantly reduced the size of the data available. But still a large volume of data might be present and moreover we do not know which tweets can be helpful to understand the emotions and needs of the people the tweets.

Now understanding the emotions and the needs of the people through the tweets is not an easy task as it involves different way of writing tweets for different people. Also, there are a large number of tweets available which makes it further difficult to extract the needs and emotions manually from the Twitter corpora. It is therefore important to have a method which can basically differentiate the tweets that includes the needs or the information about the people in distress during the ongoing disaster situation. We propose a method through sentiment analysis of the text written in tweet by the user which can possibly help us to categorize the user's tweet as either containing fact or describing the emotions and needs during the disaster situation.

Majorly, there are two kinds of tweets, one which states the facts such as the tweets from the news agencies and the other type are the tweets from the users who

describes their needs or emotions over the disaster. Here, we describe the first kind of tweets as ***Objective*** and the second type of tweets as ***Subjective***.

For example the following is an Objective tweet,
FedEx no longer to transport bioterror germs in wake of anthrax lab mishaps #news #phone #apple #mobile ,as this tweet directly states a fact about FedEx, it is an objective tweet,
whereas the following tweet is subjective,
@denisleary Not sure how these folks rush into burning buildings but I'm grateful they do. #TrueHeroes , because this tweet expresses the opinion of the user.

4.2.1 Dataset Information

The dataset for this task is the collection of disaster-related tweets identified in the previous task. Manual annotation of the tweets is done for objective and subjective tweets. Around 80% of the tweets are labelled as objective while the rest are labelled as subjective. This may be because of the large number of tweets from news agencies.

4.2.2 Proposed Method

In this section, we describes the details of our proposed method through sentiment analysis to categorize the tweets as either **subjective** or **objective**. Our proposed mechanism is basically about extracting and using the right kind of features from the tweets which can be analyzed on the sentimental values and other tweet meta data and can be helpful for the task. We have run experiment for our model using different machine learning algorithms with different parameters and observe the change in the accuracy.
We here discusses our proposed method for the sentiment analysis of disaster-related tweets. Fig. 4.2 shows the steps involved for the sentiment analysis.

1. In the dataset for this task, we have the disaster-related tweets and the subsequent POS tags from those tweets as extracted in the previous tasks. The tweets in the dataset are pre-labelled as *subjective* and *objective*.

2. From the tweets in the dataset, features are extracted which are used for the sentiment analysis further.

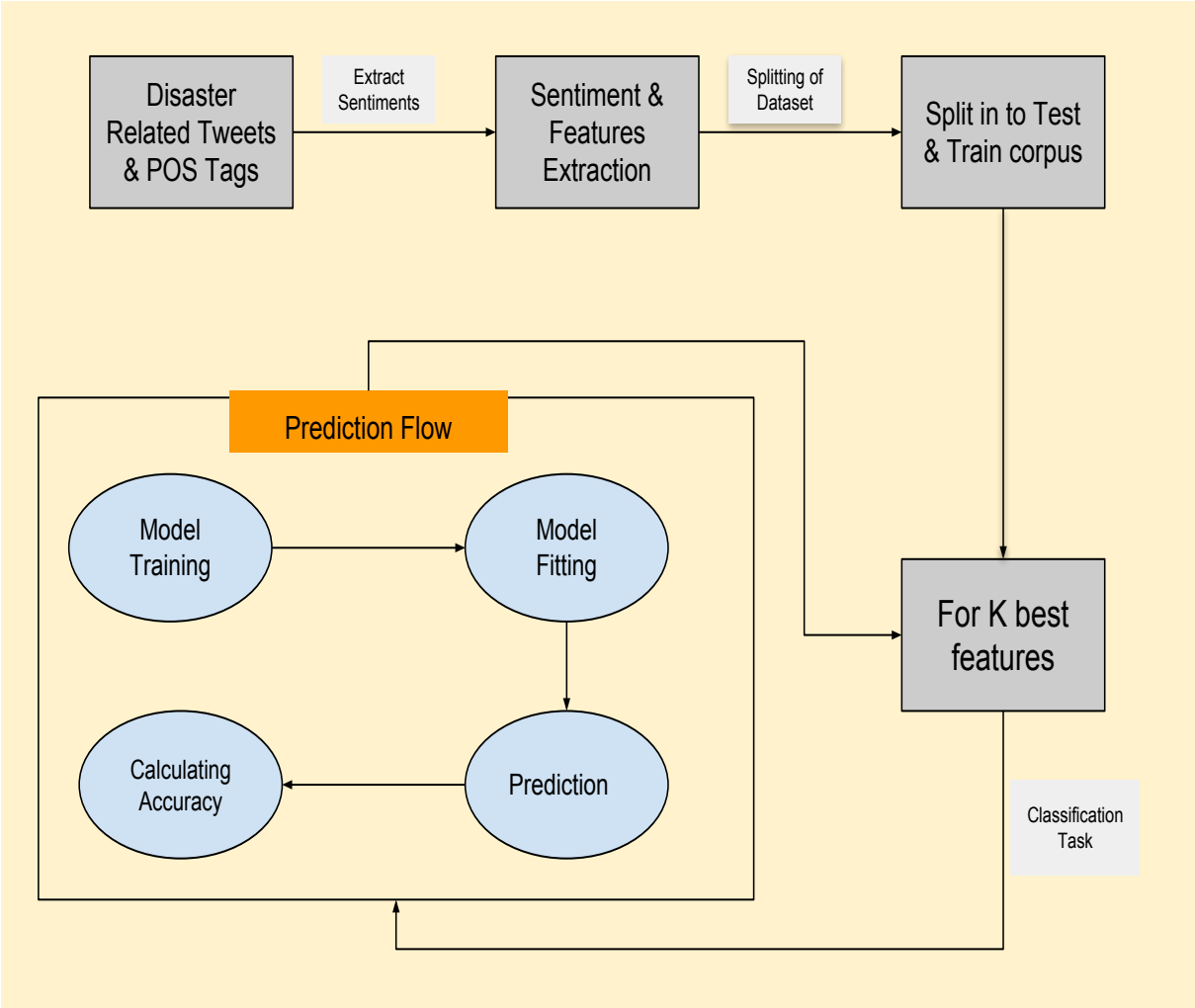

Figure 4.2: Sentiment analysis of disaster-related tweets

3. The dataset is further split into test and train corpus, same rule of 90:10 is followed for the splitting in this task.

4. After this, the machine learning models are built over training data and then the prediction is done over testing corpus.

5. Also, during the training and prediction in this task, all the extracted features are used as the experiment is done repetitively with the increasing number of features and then best feature set is also output in the result.

6. Also, we do not examine the vocabulary of the tweets for this task.

4.2.3 Sentiment extracted from tweets

1. Interesting POS tags which includes adjectives, verbs, proper noun, pronoun, social media abbreviations.

2. Exclamation presence & exclamation count

3. Question mark presence & question mark count

4. Emoticon presence

5. Digits count

6. Cap word count

7. Punctuation marks and symbols count

8. URL presence

Random forest and SVM machine learning models are used to analyze the sentiments from the tweets and categorize the tweets as **subjective** and **objective**. We discuss the results in the next chapter for this task.

4.3 Entity recognition from the tweets

After the identification of disaster-related tweets and also categorizing the tweets as subjective and objective, we still have large volume of the tweets to work with which can take a long time to process and identify the locations and people who need help. In a disaster situation, time is the most expensive entity, with every passing minute

the damage keeps increasing and the suffering of the people increases. Therefore, it is of utmost importance to have a mechanism which can help to extract the important and relevant information from the tweets or the data available.

In this section, we propose a method to extract the important entities from the disaster related tweets. Important entities can be consisted of locations, organizations, persons, time indicators, events, natural phenomenon and many more. In the tweets recorded during the disaster situations, it is observed that people looks for help and provide their whereabouts in the disaster situation. If this kind of information can be provided to the rescue teams in proper time frame, this can be helpful to save a huge amount of damage to life and property during disasters. Thus, we aim to extract the relevant entities from the tweets through Named Entity Recognition techniques discussed in [21]. In this work it is discussed that how the data from the tweets can be used for the efficient extraction of relevant information and identifying the important entities. We will discuss the technique in detail in the next section.

4.3.1 Entity Recognition Technique

Now, we describe the technique for entity recognition from the text. There are various features used for the purpose for recognizing entity from the tweets. All these features are extracted from the text.

Part of Speech Tagging

In natural language processing, POS tagging is very important and applicable to large number of applications. In POS tagging, every word of the raw text is assigned to the most frequent tag from the English part of speech categories. For the task of entity recognition, we perform the POS tagging using NLTK package of Python.

Chunking

It is another important process in the field of natural language processing where the phrases are phrases are identified from the raw text such as the noun phrases, preposition phrases from the text. NLTK Chunker has been used in our approach for the extraction of phrases from the tweet text.

Capitalization Issue

In most of the tweets, use of the capital words is not reliable because different kind of people have different approaches towards writing the tweets. To avoid this issue, we have converted all our text in to lowercase before processing the tweets. In [21], the authors have developed a capitalization classifier which determines whether the tweet is informative because of the capitalization used. Authors have manually annotated the dataset of 800 tweets either informative or not informative. The technique they followed is if the tweet begins with word which has first alphabet as capital then it is informative, otherwise not informative. Support vector machine is used for the classifier learning and features from the tweets in the form of fraction of capitalized words to the frequently used words in the dictionary is used for the classification task. The results show that the capitalization is always informative. But although the results show that the capitalization is informative, we may have to rely a lot on users to use the right capitalization while writing the tweets which is not always the case.

In the next section, we will discuss the details of how the named entities are recognized in our model.

Groningen Meaning Bank (GMB)

Before we discuss the actual process of extracting entities from the text, we will discuss about the Groningen Meaning Bank (GMB) which is a very large corpus of entities. So, for the purpose of identifying the entities from the tweets, we have used in our process, the GMB [22] which is very large corpus. This corpus contains thousands of texts in the raw form, tokenized form, it also have the POS tagging done for the text also. This is very useful for the process where the existing tags and entities can help in identifying the new entities by the process of classifying using any machine learning algorithms.

The corpus in GMB has been arranged properly in the form of categories in lexical order, and the following are the top level categories of entities present in GMB,

- Geographical tagged entities
- Organization, both government and private
- Person

- Geopolitical Entity, these are the important persons involved in politics most probably well known around the world, like Narendra Modi
- Time indicators, like if the time is mentioned for some important even in the texts
- Events
- Natural Phenomenon

Named Entity Recognition

First, we will discuss what does named entity recognition means. Named entity recognition is a process of identifying and marking the important entities from any text and to mark those entities with the suitable tags, entities here can be anything like a geographical location, some important person, any well known organization either government agencies or private agencies. Time indicators or any mention of certain event or natural phenomenon can also be a important entity.

Now, in our work also, it is of utmost important to identify these kind of entities in disaster related situations. There are various agencies during the disaster which keeps posting information about the current situation and warning for the upcoming situations. Also, there are people who are stuck in the disaster situation and post the tweets for help with their location or need but all these get lost in millions of tweets posted at that time. Thus, it is important to identify these kind of entities from the tweets so that the available help can reach to right people in a quick time.

Now, we will discuss process of extracting named entities from the tweets. Fig. 4.3 shows the architectural framework of the process of recognizing the named entities from the text or tweets in our work.

There are various steps involved to recognize the entities in the text which are discussed as follows:

1. At first the text or tweet is available in raw form from which the entities have to be identified.
2. In the next step, we are performing the text segmentation which the process of dividing the text in to meaningful parts. This is an important step which

basically divides the text according to the entities which will later be extracted from the same text. To perform this step, we have used the NLTK [23] library which is quite famous and often used in the field of natural language processing. Through this step, we aim to extract the meaning parts from the given text, so that it becomes easier in the upcoming steps to extract the entities.

3. After the text segmentation, we have performed the process of tokenization over the meaningful parts of the text which is the process of breaking the sentence or phrases in to words or tokens. This has been dome using the word tokenizer of NLTK library.

4. For the next step, we extract the POS tags with the tokenized form of the text. We basically tag the words or tokens with the tags that defines which part of speech that the word belongs to. Here, in this step for POS tagging, we have used NLTK POS tagger, not the Twitter POS tagger we have used in the last sections. The reason behind this is the large corpus of NLTK which improve the chances of identifying the tags more appropriately.

5. In the next step, the major task of identifying entities has been done using the POS tags which were assigned in the previous step and the GMB bank we have discussed previously which is the very large corpus of entities.

6. The identified entities are then given as the output which can further be used for the better dissemination of the information.

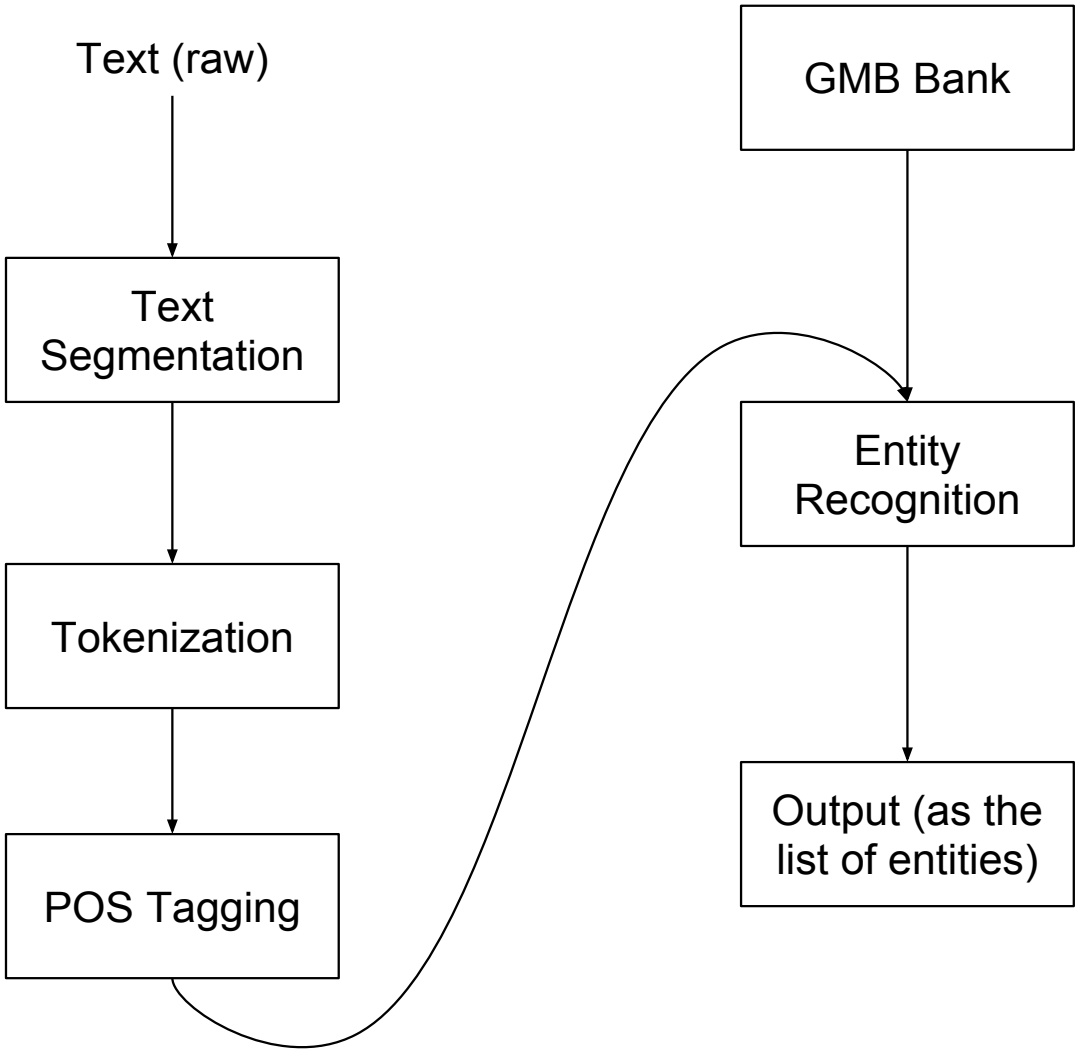

Figure 4.3: Process of recognizing entities in text

Chapter 5

Results

In this chapter, we describe and discuss the results we have obtained for the different tasks performed on the corpora of tweets. Like in the last chapter, here also we discuss the results section wise for the different tasks.

5.1 Identification of the disaster-related tweets

For this task we have Support Vector Machine (SVM) and Random Forest classifiers from the machine learning algorithms. For SVM we have tuned the penalty constant and for Random Forest, number of estimator has been tuned.

	SVM Unigrams	**SVM Bigrams**	**RF Unigrams**	**RF Bigrams**
accuracy	0.935	0.933	0.912	0.905
ppv	0.961	0.952	0.945	0.934
npv	0.935	0.926	0.892	0.883

Table 5.1: SVM and Random Forest results for identification of disaster-related tweets

As can be seen from the results, both the classifiers have performed reasonably well over the given dataset. Performance of SVM is marginally better than Random Forest. Also, in Random Forest it is clear from the results than unigrams have performed better than the bigrams. For SVM, performance of both unigrams and bigrams is almost similar.

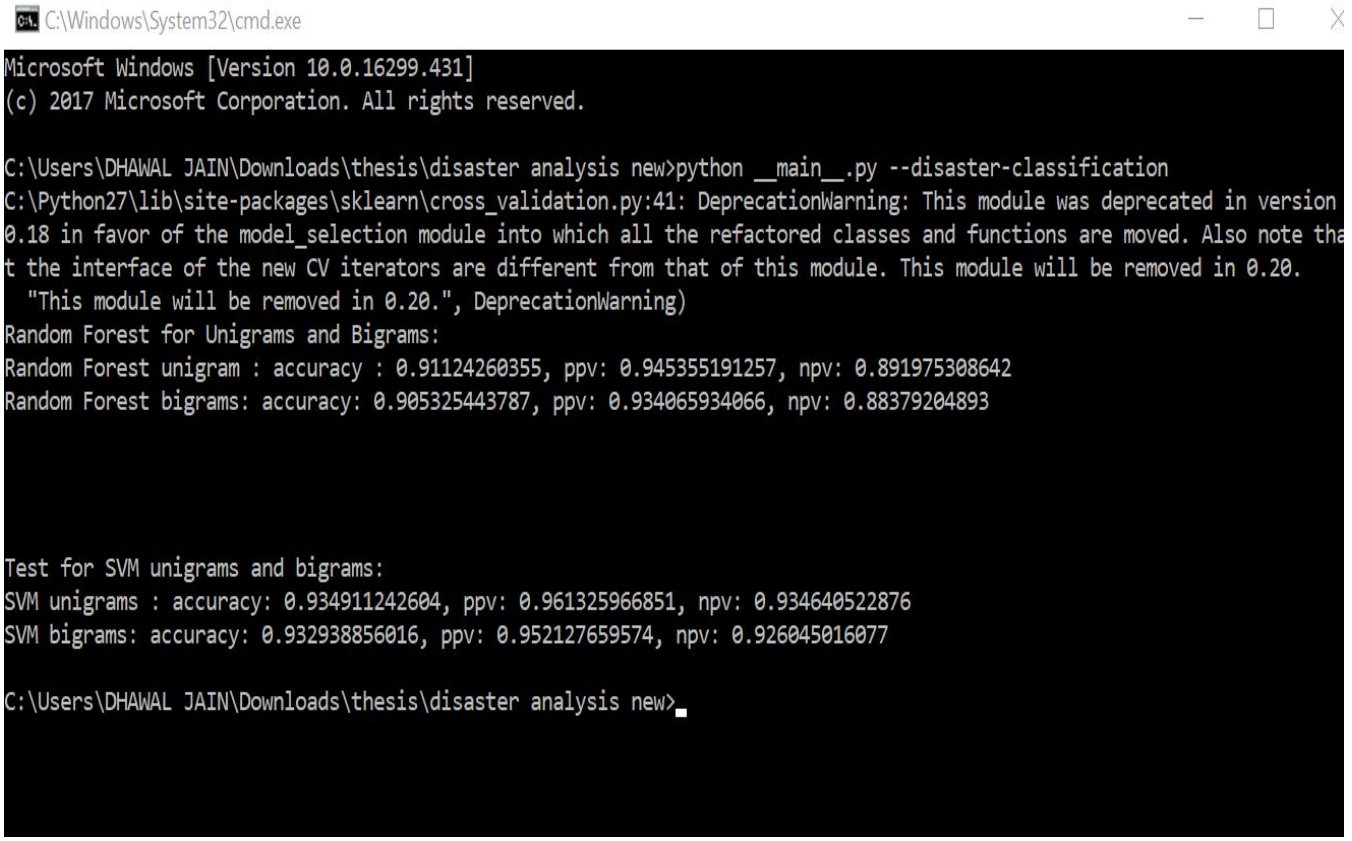

Figure 5.1: Task 1 Results

5.2 Sentiment analysis of disaster-related tweets

For this task also, we have performed the experiment using SVM and Random Forest classifiers from the machine learning algorithms. In this task, we have performed the experiment repetitively every time with the increasing number of features to estimate which features can perform the best out of all.

	SVM	**Random Forest**
accuracy	0.848	0.872
ppv	0.905	0.906
npv	0.667	0.769

Table 5.2: Sentiment Analysis results for disaster-related tweets

As clear from the table of results, Random Forest performed significantly better than the SVM machine learning algorithm.

Also, for the random forest the best results are obtained when 14 features are

used to train the model and prediction. The features that generated the best results for Random Forest are : exclamation count, exclamation presence, question mark presence, question mark count, url presence, emoticon presence, digits counts, capital words count, cap_letters_count, number of punctuation marks and symbols count, length. For SVM, only 4 features are selected for the best results, number of exclamation marks, presence of URL, presence of emoticon, number of punctuation marks and symbols.

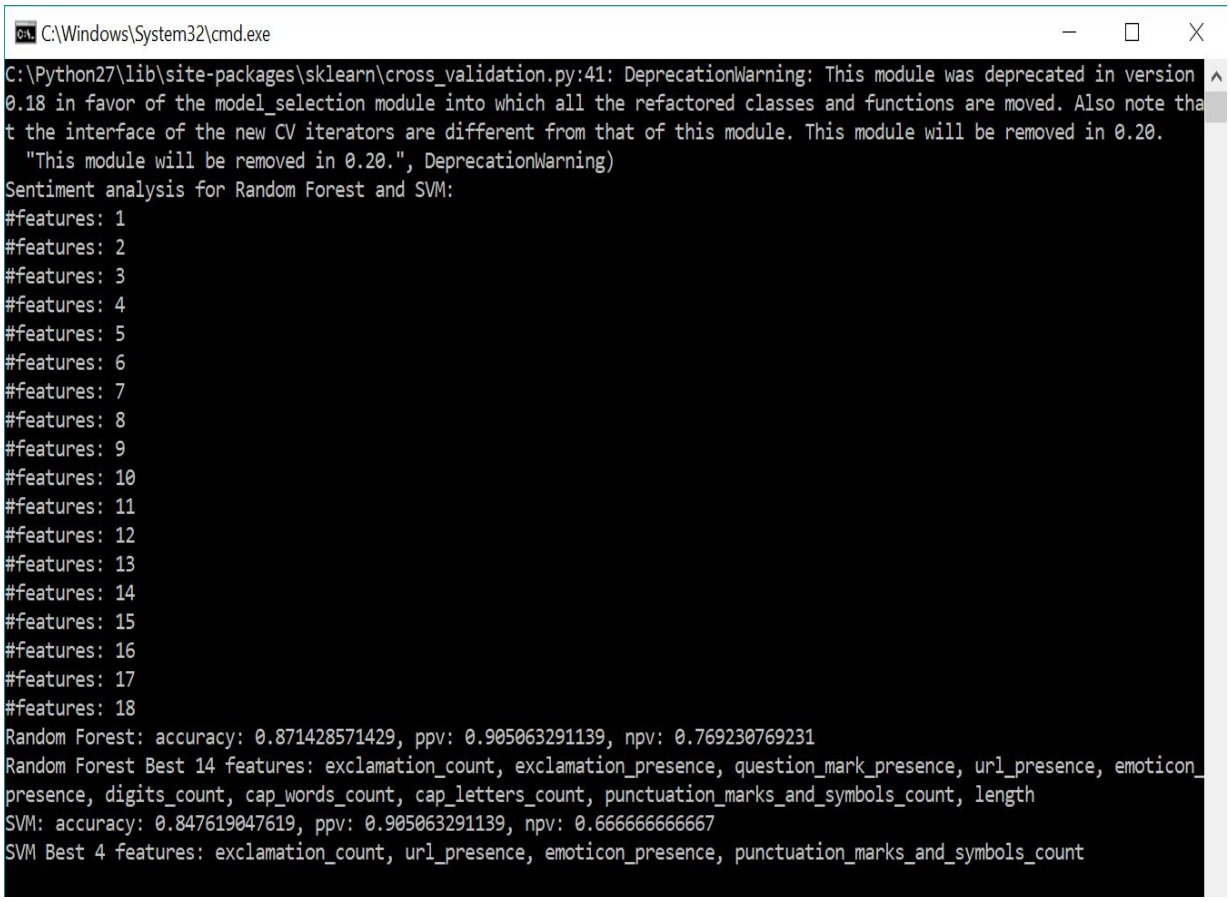

Figure 5.2: Task 2 Results

5.3 Entity recognition from the tweets

For the task of recognizing entities from the tweets, out of all the categories discussed in GMB corpus, in our model, only four of the categories are found to be relevant. Location, organization, persons, geopolitical entity are the important entities. The results that were produced in our dataset for these entities are the following :

- **Location :** From the disaster related tweets in our datasets, California and Hiroshima are identified as the important locations because of the disaster related tweets from Japan.
- **Organization:** In this category, news agencies are given as the output from the dataset because the dataset contains tweets from the news agencies for the information regarding disasters. ABC news and BBC are the main organizations in the output. Though in this technique, some user references are also given as the organizations which are the false positives.
- **People:** In this category, only Obama is produced as the relevant result.
- **Geopolitical Entity:** This section in the result is mostly comprises of the users nationalities.

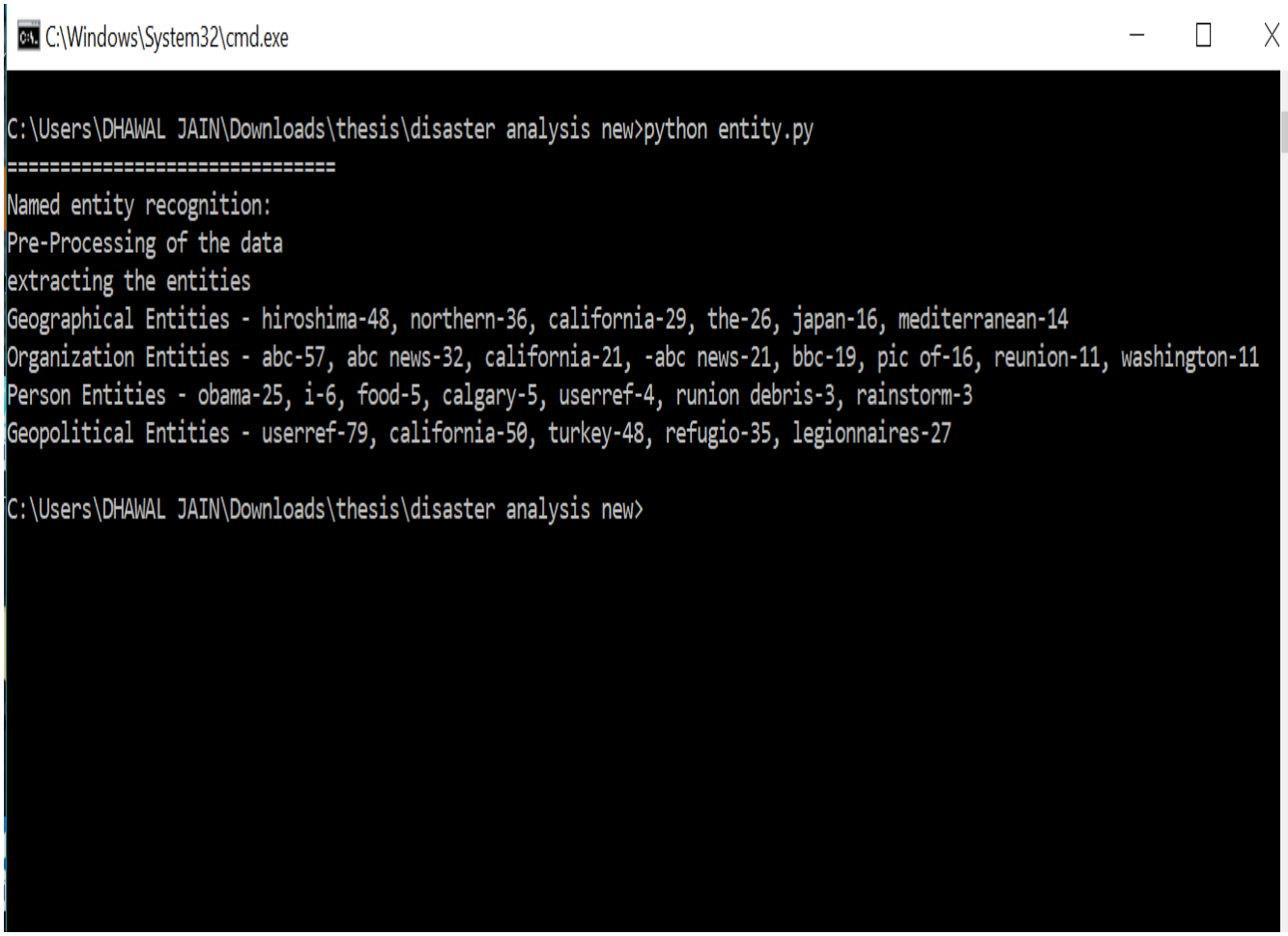

Figure 5.3: Task3 Results

Chapter 6

Conclusion

With the increasing popularity of the online social media platforms, everyone in using the online social platforms to share the important information and the event in their life. Even the government agencies are using these platforms to address the grievances of the people who post their problems online.

In this work, we have proposed a system where the Twitter data can be used to extract the relevant information at the time of disasters. We have identified the the disaster-related tweets from the large volume of the tweets posted at the same time. We then have proposed to do the sentimental analysis of those tweets to understand the needs and emotions of the user. We then have proposed a technique to identify the important entities from the tweets which is important map the relief resources in quick time. Results on our proposed technique shows that it is very much possible to better manage the disaster situations if we can use the data from the social media platforms in the correct manner.

In future, other datasets from the other disaster related situations can be analyzed over this model for the better analysis of the performance of the model. Also, it will be interesting to see how this model can be extended to other social media platforms like Facebook.

References

[1] L. Palen. Online social media in crisis events. *Educause Quarterly* 31, (2008) 76–78.

[2] Y. Xiao, Q. Huang, and K. Wu. Understanding social media data for disaster management. *Natural hazards* 79, (2015) 1663–1679.

[3] D. Feldman, S. Contreras, B. Karlin, V. Basolo, R. Matthew, B. Sanders, D. Houston, W. Cheung, K. Goodrich, A. Reyes et al. Communicating flood risk: Looking back and forward at traditional and social media outlets. *International Journal of Disaster Risk Reduction* 15, (2016) 43–51.

[4] J. Kim and M. Hastak. Social Network Analysis: The role of social media after a disaster. *In 10th anniversary homeland Defense/Security education summit* .

[5] S. E. Middleton, L. Middleton, and S. Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems* 29, (2014) 9–17.

[6] M. E. Poorazizi, A. J. Hunter, and S. Steiniger. A volunteered geographic information framework to enable bottom-up disaster management platforms. *ISPRS International Journal of Geo-Information* 4, (2015) 1389–1422.

[7] C. Reuter, O. Heger, and V. Pipek. Combining real and virtual volunteers through social media. In Iscram. 2013 .

[8] E. Yoo, W. Rand, M. Eftekhar, and E. Rabinovich. Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. *Journal of Operations Management* 45, (2016) 123–133.

[9] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Practical extraction of disaster-relevant information from social media. In Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013 1021–1024.

[10] M. A. Cameron, R. Power, B. Robinson, and J. Yin. Emergency situation awareness from twitter for crisis management. In Proceedings of the 21st International Conference on World Wide Web. ACM, 2012 695–698.

[11] G. Beigi, X. Hu, R. Maciejewski, and H. Liu. An overview of sentiment analysis in social media and its applications in disaster relief. In Sentiment analysis and ontology engineering, 313–340. Springer, 2016.

[12] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems* 27, (2012) 52–59.

[13] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Extracting information nuggets from disaster-related messages in social media. In Iscram. 2013 .

[14] P. W. Kinyua and P. Mbataru. An Assessment of the Effects of Social Media on Disaster Management in Kenya, Case of Nairobi City County .

[15] P. M. Landwehr and K. M. Carley. Social media in disaster relief. In Data mining and knowledge discovery for big data, 225–257. Springer, 2014.

[16] M.-A. Abbasi, S. Kumar, J. A. Andrade Filho, and H. Liu. Lessons learned in using social media for disaster relief-ASU crisis response game. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer, 2012 282–289.

[17] B. Takahashi, E. C. Tandoc Jr, and C. Carmichael. Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. *Computers in Human Behavior* 50, (2015) 392–398.

[18] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. 2013 198–206.

[19] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications* 13, (1998) 18–28.

[20] A. Liaw, M. Wiener et al. Classification and regression by randomForest. *R news* 2, (2002) 18–22.

[21] A. Ritter, S. Clark, O. Etzioni et al. Named entity recognition in tweets: an experimental study. In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011 1524–1534.

[22] J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva. The Groningen Meaning Bank. In Handbook of Linguistic Annotation, 463–496. Springer, 2017.

[23] S. Bird and E. Loper. NLTK: the natural language toolkit. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004 31.