National College of Ireland

Combination of topic modelling and deep learning techniques for disaster trends prediction

MSc Research Project
Data Analytics

Ankita Behera
Student ID: x18103090

School of Computing
National College of Ireland

Supervisor: Dr.Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Ankita Behera
Student ID:	x18103090
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr.Anu Sahni
Submission Due Date:	12/08/2019
Project Title:	Combination of topic modelling and deep learning techniques for disaster trends prediction
Word Count:	7358
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Combination of topic modelling and deep learning techniques for disaster trends prediction

Ankita Behera
x18103090

Abstract

A natural disaster, even though its small, results in great human and environmental loss. If the disaster is predicted at the earlier stages, then it will be helpful for the people and the government for helping and coordinating with the rescue team. The media platform like Twitter is a valuable communication medium which allows the user to update about the condition and situation, which directly aware people in the affected area and helps rescue team to co-ordinate more effectively. This paper illustrates a recently developed approach which classifies various disaster and later predicts the trend from their severity level.This approach was previously applied in the field of mortality prediction and recommendation system. The framework consists of the integration of LDA for classification and LSTM (artificial recurrent neural network) for predicting the trends using the tweets. The proposed model is created considering the tweets and the weather data over a period. The model has predicted the accuracy of 80.5% with the neural network which is proved to better than any other machine learning techniques. This new combination of the latent topic information and the semantic information by the neural network shows great promise for improving the disaster management activities and forecasting the catastrophic trends.

Keywords: Topic modelling, text analysis, Natural language processing, Text classification, Deep learning, long short-term memory, RNN, Predictive analysis.

1 Introduction

A natural disaster is an emergency event which results due to the natural processes of the earth and doesnt have a fixed time of its arrival. The damages include both the property damages and even lots of loss in life. In the year 2018, it was reported by EM-DAT (International Disaster Database) that there was total 215 natural disaster from which overall 68 million people were affected from the natural calamity. There was a total loss of US$131.7 billion in economic loss across the world (Yaghmaei; 2019). These losses would decrease if the upcoming disaster is known before its arrival. The disaster could not be stopped but the public health emergencies and the natural disaster management always find a proper way of communication in the different ways that are either by evacuation notice, early warning or risk information or by making people aware about the symptoms and medical treatment after the disaster(Reynolds and Seeger; 2005). For that proper communication and prediction is a must.

1.1 Motivation

From the past few years, many kinds of research and practices have been done and applied to know about the disaster beforehand and to handle the situation better. According to Ogie et al. (2018) we can get quick and exact information by gathering real-time relevant information by the advancement of the machine learning techniques and artificial intelligence. The main motto behind applying such methods is to provide proper disaster management, reduce the loss and damages. Protecting peoples life and property against the adverse weather condition is the main aim of the project.
Information and communication technology play a key factor in determining the forecast and helping people to generating and early warning and by arranging immediate help to the disaster-prone area (Lohumi and Roy; 2018). Many data mining and analytical techniques are used for gathering proper data for disaster management, classification and prediction from various sources like satellite, remote sensing, geographic observation or by social networking sites like Twitter, Reddit etc (Goswami et al.; 2018).
According to a report generated by the EM-DAT, the number of deaths caused by the disaster was decreased to 11,804 in the year 2018 compared to previous years(Yaghmaei; 2019). The best example of the early warning system was seen in the recent news about cyclone Fani which has hit Indias coastal region where the Indian Meteorological Department generated early warnings that helped the authorities to manage an evacuation plan and which in return reduce the loss of life from a powerful cyclone. 1.2 million people were evacuated and warned within 24 hours which leads to the biggest evacuations in the history [1].

1.2 Project Background

Social Media captures the emotions, trends, opinions, and behaviour of the people during a calamity situation. Either people seek help or enquire information regarding food or shelter or any transportation services. Because the social network is so widely spread that it allows individuals who are affected to rapidly connect with the needed resources (Velev and Zlateva; 2012). There are many social networking sites which are very popular during mass emergencies like Facebook, Twitter, Reddit, YouTube, myspace, etc. Among all the social media twitter has been considered in this project. According to a report generated by twitter[2], there are overall of 326 million monthly active users and 500 million total tweets are sent per day.Twitter is a huge and quick platform which collects a huge amount of data with few milliseconds. The live streaming data which is collected describes the emotions of the users during the disaster situation.
The weather data is even a key factor for predicting the disaster. Over the last century, Scientist has made a significant improvement in monitoring and predicting the climatic changes and water-related disaster. This is due to the proper meteorological satellites and data processing high-performance computers which monitor the weather system anytime any place. In this project, both the tweets and the meteorological weather data is considered to forecast the upcoming disaster trends whether still, the disaster is severe, or the effect is reduced[3], journal=World Meteorological Organization.

[1] https://timesofindia.indiatimes.com/india/un-agency-praises-indias-efforts-to-minimise-loss-of-life-from-cyclone-fani/articleshow/69179573.cms

[2] https://www.omnicoreagency.com/twitter-statistics/

[3] https://public.wmo.int/en/resources/bulletin/monitoring-predicting-and-managing-meteorological-disasters

According to an article published by Adam Ostrow on the earthquake which had occurred in Japan, Twitter is the first and quick means which had informed about the news of earthquake before any other social network [4]. A similar type of research was done by Shekhar and Setty (2015) where a sentiment classifier was built based on the twitter data where the reaction was categorized(negative or positive), various stress level was measured and geographic distribution of the tweets was measured by the K-nearest Neighbour algorithm. There are many works done which tell about the usage of Social media as an effective way and uncensored medium for disaster analysis.
In this research paper, a model has been implemented which is the integration of deep learning techniques (Long short term memory) and topic modelling is done by Latent Dirichlet Allocation (LDA). The latent topic information is derived by LDA whereas the sematic information of the model is done by LSTM which is an artificial recurrent neural network. The demerits of LDA is that, the twitter data being very short the data sparsity effects the efficiency of the model as compared to the large text (Hajjem and Latiri; 2017). Whereas RNN is the first kind of art which can remember previous input in the memory, and it works better in a large sequential dataset with better accuracy than others. Also, RNN can model sequence the data where each sample is dependent on the previous one or the time series which is very helpful in predicting the trends in disaster [5].
The model which was created classify the tweets by a classifier which combines NLP and machine learning techniques. The first step before implementing is data extraction which is done by twitter API which acts as the source of data. Then the data which was extracted was analysed and transformed which reduces the size of the text, which leads to classifying similar pattern data together into a single group. After classifying, the data was predicted, and a proper conclusion was drawn from it.

1.3 Research Question

RQ: How efficiently integration of topic modelling and deep learning methods would help in classifying and predicting disaster trends in the area?

Sub RQ: Whether the proposed model is more accurate than the other machine learning techniques?

To solve the research, question the following objectives are specified and implemented

1.4 Research Objective

Obj1: A critical review of literature on Extraction, classification, and prediction of disaster (2004-2019).
Obj2: Extraction of the twitter data from Twitter API.
Obj3: Implement, evaluate and result of the LDA.
Obj4: Implement, evaluate and results of Recurrent Neural Network.
Obj5: Implement, evaluate and results of Long short term memory.
Obj6: Implement, evaluate and results of Support Vector Machine.
Obj7: Implement, evaluate and results of Random Forest and Naive Bayes.
Obj8: Predicting disaster trends.

[4]https://mashable.com/2009/08/12/japan-earthquake/?europe=true

[5]https://medium.com/@purnasaigudikandula/recurrent-neural-networks-and-lstm-explained-7f51c7f6bbb9

Obj9: Comparison of the developed models.

The remaining paper is arranged in a sequence described below. In the next section, it mainly investigates the existing literature in the classification and prediction of the disasters.Section 3 describes the scientific methodology and approach(data extraction, pre-processing and data mining techniques).Chapter 4 tells us about the implementation, evaluation and the results. Section 5 finally concludes the paper and has the future work to be done.

2 Related Work

The literature review investigates about the classification and prediction of the disaster using topic modelling and deep learning techniques. The section is segregated into many subsections i.e. 1) Extraction and classification using social media 2) Disaster prediction and monitoring 3) Literature survey on deep learning and topic modelling on different fields 4) Comparison and Conclusion.

2.1 Extraction and Classification using social media

There are many papers which talks about the extraction of data from social networks. According to Zahra and Purves (2017) twitter being an independent platform that has numerous amounts of ideas, views and information regarding various events. Here the Naive Bayes is used for classification of the tweets from the users from Asia and Europe which reveals the granularity of geographical information and various aspect of credibility (User-based features). The program was run by R which captures real-time tweets based on keywords related to disaster. The classification was done based on informational or non-informational tweets which had provided a satisfying result.
According to Cuesta et al. (2014), the social network is a platform which is used for knowledge extraction and Twitter is an open environment where people share information, opinions and which is the best source for trends discovery. Here an open framework is proposed which collect and analyze data from the twitters public streams which emerges as a repository of knowledge and information for many types of research. The framework even provided with a complemented language-agnostic sentiment analysis module which is helpful for sentiment analysis of the collected tweets. It classifies the tweets into positive, negative and neutral.
According to Bouazizi and Ohtsuki (2016) whose paper mainly focuses on automatic sentiment analysis on the tweets collected and further it has classified the tweets into positive and negative. A pattern-based approach was proposed which classify the tweets into seven different classes (Happiness, sadness, anger, love, etc) and detect the hidden emotions from the post. Here the data mining was difficult to do due to the short text which resulted in lower performance. 4 different types of features were extracted after which random forest is used for the classification. For binary classification i.e. for positive and negative the accuracy reaches up to 56.9% than ternary or multi-class classification. Hence, it is proved that the training dataset should be more optimized than the present. Sriram et al. (2010) has even proposed very similar way where they have used a domain-specific feature extraction which classifies the tweets into few generic classes (opinions, deals, events, news, and other categories) which were a disadvantage for other approaches like a bag of words. Tweets patterns are different for different authors, so the information

related to the authors plays a very important role. Here Naive Bayes classifier using 5-fold cross-validation is used for tweets classification. This approach performs better than other traditional approaches (Bag of words). But the flaw of this approach is the data has some noise contents which decreases the performance of the model.
Benny and Philip (2015) proposes an approach where the tweets are collected using a distinct keyword and then topics are summarized from the related keywords. By using clusters of frequent patterns, topic detection is done. The topics are found out by using new pattern-based topic detection techniques which overcomes the wrong correlation of patterns which was seen by pattern-based clustering method. It uses two novel algorithms i.e. TDA (Topic Detection using AGF) and TCTR (Topic Clustering and Tweet Retrieval) to extract topics. It is concluded that the result obtained is better than other approaches irrespective to the size of the dataset.
Kireyev et al. (2014) has provided a study on the topic modelling from the tweets collected. As topic modelling uses the bag-of-word approach, multinomial probability distributions and the topic infer the latent relationship between elements in the data, so it is easy to go for topic modelling in large document corpora. From twitter data, they can retrieve the date of the event and the location. The term weighting scheme has been used here which means that specific words have higher weight age. They were able to segregate the tweets into two different categories. First one was based on an event (Tsunami or earthquake) and the second one by searching messages with keywords to specific categories.Qualitative analysis was done for evaluation.
Verma et al. (2011) studied on an approach which creates situational awareness from the tweets generated from during mass emergencies. A classifier was built which uses the automatically extracted linguistic features and classify the tweets which are helpful for the people who need information during an emergency. Natural language processing (NLP) and machine learning techniques are used to classify the tweets based on personal or impersonal style or subjective based or formal or informal messages. The subjective based tweets are people who need help or sympathy related tweets or infrastructure damages etc. In the experimentation stage Maximum Entropy and Naive Bayes is used where the former gives a better result. It is concluded that the link between situational awareness and other characteristics improves the result and gives a good prediction. The drawback of the classifier is that as twitter contains much redundant information written in a different style which is misclassified.
A close work as (Verma et al.; 2011) was done by Stowe (2016) where the classification was not only related to the situational awareness, but it was more general or multi labelled annotation (reporting, action, sentiment, information or evacuation) was done and the categorization was done more finely. The schema was built considering the attitudes, sources and decision-making behaviour of the tweets. Here for classification three techniques i.e. Naive Bayes, Maximum Entropy and support vector machines were used. The conclusion which was got that relevant tweets were identified automatically with high precision, but the fine-grained classification was difficult to classify due to lack of positive examples.
According to Ashktorab et al. (2014) a twitter mining tool called tweedr was made which extracted information for the people who worked in the disaster relief department. It mainly has three parts i.e. classification, clustering and extraction. Classification is done on the basis of whether it is infrastructure damage or people need for help. Various machine learning algorithms are used like sLDA, logistic regression and SVM for classification of the tweets. The similar tweets were merged together in the clustering phase.

And in the last extraction stage the tokens and specific information related to the tweets were extracted.

2.2 Disaster prediction and monitoring

Many researchers have worked on the forecasting the natural disaster which helps in generating the early warning and awaking people and government to take necessary steps before the arrival of the disaster. Even though there will be a loss of many properties still millions of lives can be saved by this prediction. The prediction is mainly done using machine learning algorithms or by using neural network.

Bande and Shete (2017) have predicted the flood by using an embedded system hardware and wireless communication network (IOT) based flood monitoring system and artificial neural network. It was done to improve the reliability and scalability of the flood disaster management system. The data is captured by the wireless system and then the data is sent to the computing device for analysis. ANN uses the prediction algorithm to analyse the data for forecasting of the flood. As ANN which consist of three internal hidden layers, uses the neurons in the neural network which helps in generating the output. Parameters like pressure, rainfall, water level, humidity etc are considered as temporal correlative information. From here we can conclude that result of ANN is better than any other real-time flood prediction.

A similar kind of approach was used by Bhatia et al. (2018) where the earthquake was predicted to decrease the death count and to prevent from any economic loss on the affected region. Artificial neural network was used in creating an earthquake forecasting system. The accuracy was achieved by using different hyper parameters and input. Number of earthquake cycle is calculated to get the current progression of the earthquake cycle. Due to less number of attributes the neural network was failed which was replaced by LSTM network. But as the input data was very unevenly distributed so LSTM was not successful.

Kansal et al. (2015) had proposed wireless sensor network (WSN) for generating warning about the upcoming disasters using the meteorological data. The focus was mainly on how fast and accurate the result will be so that much damages will not be caused due to this. WSN are mainly sensors which helps in monitoring ands ending the result to the machine learning models. Detecting accuracy was enhanced by using the machine learning techniques. Many machine learning techniques are used in this paper, but regression proved to give the best result than any other techniques. It achieves a low root mean square error and high R-squared value.

One of the sandstorm prediction websites was designed by Ahmed Shaiba et al. (2018) where they have used source as the historical weather data by using weatherData package from R programming which retrieve data from location if valid date and time is given. The forecasting was done 24 hours before in real-time by using machine learning methods. The experiment is done on three methods i.e. Logistic Regression, Naive bayes and CART decision tree out of which CART decision trees gives better result than others. The ROC Curve for CART decision tree performs best than others.

An algorithm for proposed by Singh et al. (2017) which collects the twitter post related to the flood and find out the people who are asking for help from the effected place. Tweets are given as an input to the system and then according to the priorities it is classified into low and high. If the location of the user having higher priority is not known then using Markov model the past location of the user is calculated and help was send accordingly.

In this paper 3 classification algorithm is used i.e. random forest, SVM and gradient boosting. The performance was calculated by considering the F1-score, accuracy, ROC curve and precision.
Sakaki et al. (2013) have used the real-time twitter data to develop an earthquake reporting system to monitor the tweets and to detect the targeted event. A twitter classifier is made based on the keywords, context and the number of words in the tweets. To forecast the upcoming earthquake in Japan and to warn the people this model was created which sends notification to all the users around the location. Twitter is treated as a sensor and apply filtering to locate the exact place of the occurrence. Through SVM the tweets are classified as positive and negative and then by probabilistic spatiotemporal model the location is detected. IT was reported that the model detected 93% high probability than Japan meteorological agency.

2.3 Literature survey on integration of deep learning and topic modelling on different fields

In past few years many researchers have worked on the merger of topic modelling and the neural network. The combination of both helps in better classification of topics and in prediction. In one of the papers where the same approach of combination of LDA and LSTM has been used for classification of the reviews and understanding the context related recommendation system (JinXin et al.; 2018). Also, it has been in mortality prediction in healthcare field(Jo et al.; 2017). According to Zaheer et al. (2017) who proposed a LLA model which is Latent LSTM Allocation which bridges the gap between sequential RNN and non-sequential LDA. The model which is created is highly concise and highly interpretable. This new technique of combining topic modelling with the deep neural network is used to predict the disaster by twitter data.

2.4 Conclusion

Based on the many researches which are discussed above we can conclude that there are many machine learning algorithms done to identify and classify the disaster occurring at a place. There are few papers related to the prediction of the natural calamity before its occurrence which is helpful for the warning of the people. The combination of the topic modelling and the neural network has been proved to give a better result in many fields. In order to get a better prediction and classification the integration of the topic modelling and the deep learning techniques is implemented. In the next section, the approaches and the methodology are discussed in order to achieved the prediction model.

3 Methodology

Out of many data mining methodologies, KDD fits the best for the research,. According to Saltz et al. (2017) it works easily in the large dataset and extracts pattern and information using machine learning techniques. As KDD steps focus mainly on steps of execution rather than project management approach so it is best suited for the classification and prediction. From figure 1 we can clearly see 6 different stages which complete the model. The methodology of the KDD is modified according to the project need and this modified

KDD is used in the research. As reported by[6], the data is collected from internet or huge repositories or large database or data warehouse and then a specific algorithm is applied to automate the whole process. The limitation of the model is as there the data which is collected is mainly from twitter data so there are lots of noise present in the data which later affects the accuracy of the prediction. Also, as the deep neural network is used for prediction so there are chances of overfitting and underfitting which should be taken care of. The process of the research work is divided into following stages:

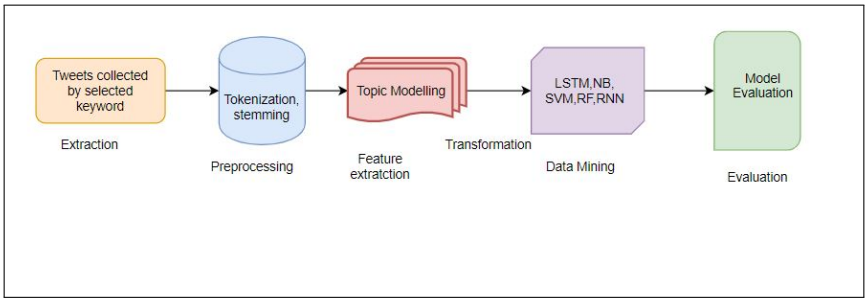


Figure 1: Methodology for disaster prediction

- Data Collection: The real time data is extracted from the twitter search API using selected and specific keywords(Shekhar and Setty; 2015). In this project keywords like natural disaster, earthquake, flood, hurricane, natural calamity etc are used. All the tweets are captured and transferred into csv file.

- Pre-processing: The next stage is the pre-processing stage where using R programming the cleaning the data captured. As the tweets are short and informal and has lot of noise so its necessary to clean it before using it in model. We have even considered the weather data for a specific period. The weather data is downloaded from online portal which gives past historical weather data[7]. Few pre-processing steps like removing URLs and hashtags, removing NAs, removing re-tweets, removing non-English words, removing special characters, removing stop words, changing to lowercase, removing numbers and stemming. .

- Feature Extraction: The data which is collected is unsupervised data. And to make it supervised data, topic modelling is used which extracts different types of disasters which is occurring at that time period when the tweets are collected.

- Transformation: Before using the model into any of the machine learning techniques few factors need to be transformed and the feature need to be extracted which is done by the sklearn package in python. Label encoder of the sklearn is used to convert the text into the numerical and to vectorize it for the models.

- Data Mining:In this stage different data mining techniques are used to create different models. Packages like sklearn and keras are used to create different models using SVM, Naive Bayes, Random forest, RNN and LSTM.

[6]http://www2.cs.uregina.ca/ dbd/cs831/notes/kdd/1$_k$*dd.html*
[7]https://www.worldweatheronline.com/developer/premium-api-explorer.aspx

- Evaluation: At the last stage, which is evaluation, taking accuracy, sensitivity, specificity,ROC, precision as measures are found out of different models and compared and tested.The learning curves and confusion matrix are evn plotted. 10-fold validation test is done for verification of models.

4 Design Specification

The prediction model which is seen in figure 2 has two section. The first section is where the classification of the tweets is done, and the second section is where the prediction of the disaster trend is seen. The raw tweets are collected and classified accordingly by checking for the dominant topics. Once the disaster is available, it is combined with the weather data to be feed into the neural network model to forecast the upcoming disaster trends.

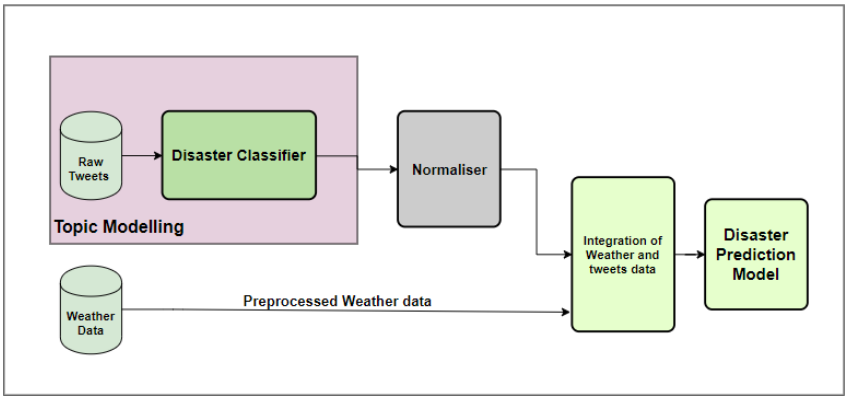


Figure 2: Project Design

- Topic Modelling : It is one of the essential algorithms where different disaster are extracted. LDA is used which is one of the generative probabilistic models for the collection of discrete data. Each document is considered as the mixture of topics and every topic is considered as a probability distribution of the words.In the twitter data the weightage of the specific word is considered in whole document. The dominant topic is considered by calculating the word which has higher frequency. A certain document has multiple topics. Each topic uses the bag-of-word technique which is the occurrence of words in the document. A dictionary is created from the bag of words which consist of several words and occurrence of each words, which is later used to create LDA model.
  (Yang and Zhang; 2018) According to LDA each word comes from a topic and each topic is selected from document distribution over topics. So the probability of a word given a document is

$$\sum p(w|t,d)p(t|d) \tag{1}$$

  Where,
  td = P(t—d) which is the probability distribution of topics in documents.
  wt = P(w—t) which is the probability distribution of words in topics.
  T is the total number of topics.
  Later TF-IDF is used to visualize the topics. To improve LDA results, TF-IDF is used. TF-IDF is used to show how important the word is in the document. Now

the topic modelling is used to identify the topic which are hidden in the tweets. In this case k = 2 (number of topics) and the number of frequent words considered is 7.

- Prediction: Here the prediction of the upcoming events is mainly done by deep learning techniques because the context information is maintained by the text. Even though neural network is one of the powerful and most used techniques still the no memory is being associated with the designed model which cannot be used for sequential data. This disadvantage of neural network is solved by RNN which has a feedback loop which acts as a memory for the model. The sequence model i.e. mainly RNN is considered mainly because RNN traverse through the data from one end to the other and maintains the details about the input variable for the future use in its internal memory. LSTM is an advance form of RNN which has both long-term and short-term memory component. As the meaning of a word is linked with the word preceding it so LSTM is best used for sequential data. As we are forecasting for the disaster, so LSTM is very much helpful in our case.

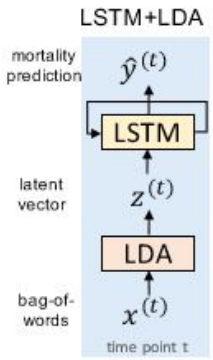


Figure 3: LSTM and LDA (Jo et al.; 2017)

From figure 3 we can see that, LSTM inputs are vector topic distribution which are represented in vectors: Z = (z(1),...,z(T)), and the output of LSTM is prediction y(t) at every point t which is also trained to decrease the prediction loss $H(\vec{y}^{(}t), y)$ time to time. where y is different types of disaster which is either 0 0r 1 in this case. (Jo et al.; 2017)
The integration of sequential LSTM and non-sequential LDA is ideal because it requires less input and has accuracy in terms of prediction than rest all methods. LDA has strong interpretability quality whereas LSTM captures the temporal information. LDA alone has data sparsity issues which effect the efficiency of the model and it ignores the structure of the word(Hajjem and Latiri; 2017). On the other hand, topic modelling has the information about the occurrence of the words which balances the loss of information by deep learning. The bag of word ignores the sequence information ($\sum logp(wi|model)$. The temporal information can be preserved by using some RNN model to it

$$(\sum logp(wi|w1, ..., wi1, model)) \tag{2}$$

Due to the less gradient value the RNN is not best suited for the model which is removed by using LSTM which is designed to handles the vanishing gradient

problem. This ensembled model the sparsity and interpretability is taken from LDA whereas the accuracy is got by LSTM together makes a perfect model for prediction.(Zaheer et al.; 2017)

5 Implementation

5.1 Introduction

In this section we have discussed different type of methods which we have used to implement the model. A brief description about the dataset is even done in the section. Methods are discussed how the disaster was classified and later how different models are implemented for prediction. The programming language used for this research is R and python. Different type of models has been implemented to compare the result and find out the highest accurate model.

5.2 Dataset

The data for the research is collected from the twitter streaming API which collects the real-time tweets by the people during the calamity situation. The SQL database was set up to connect with the Twitter API to collect the feeds. With proper authentication from the twitter considering the GDPR issues the tweets are collected with proper keywords. In this case the tweets are collected for one month (27th May 2019 to 30th June 2019) considering all the disaster occurred all over the world for that duration. The second dataset which is considered is the weather data of few places where the disaster is mainly seen i.e. in USA and Gujarat (India) from online historical weather data[8]. Both the csv file was merged using the location in R programming.

5.3 Implementing LDA (Topic modelling)

To extract the specific topics from a document we use LDA which is one of the examples of topic models. As the pre-processed data is unlabelled, so its very important to first make it labelled data by extracting different topics. Topic per document or words per topics is built from the LDA. But before that we must clean and make the data noise free. This is done by R programming. This is the fundamental process of NLP. The R programming libraries which are used for pre-processing are "tidytext, dplyr, DBI, textcat, tidyverse, tokenizer". Removal of the Nas, tokenization, retweets and non-English tweets are done. It is very important to keep only distinct tweets instead of repeated. The tweets are then converted to corpus for further processing which is done by using "tm, SnowballC and NLP" libraries. After which the text is converted to lowercase, punctuations are removed, removal of URLs/links, numbers and the stop words are done. Stemming is even one of the important steps for text mining where all the words are converted to its root words.

[8]https://www.worldweatheronline.com/developer/premium-api-explorer.aspx

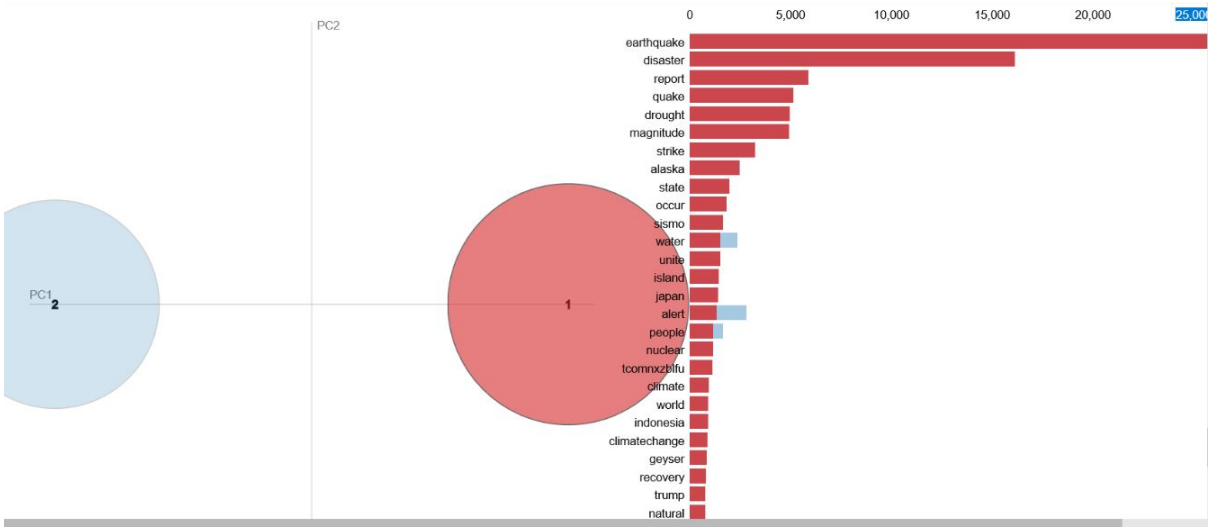


Figure 4: PyLDAvis output

For creating the topic models, we have used python library i.e. "gensim" and "nltk" libraries. Nltk package is used for lemmatization which uses spacy model. The next step is to create a dictionary from the data which is processed which contains the number of times the word appears in the training set. The dictionary is created by gensim Dictionary which contains total words and its occurrence. Later the dictionary is helpful in creating bag of words by "gensim doc2bow".Then the frequency of each word is calculated by using TfidfModel and transformation is applied on that model. The LDA model is generated by genism.models.ldamodel where we mention how many topics we need to extract. Here 2 dominant topics are taken into consideration.The visualization of the topics is properly done by pyLDAvis python library which helps the user to visually interpret the topic model. Figure 4 clearly visualize the dominant topics by pyLDAvis. ALso, we can visualize it by WordCloud seen in figure 5, of individual topics which shows us the dominant topics. It is implemented by using WordCloud() in WordCloud library in python.The dominant topics are then assigned to each of the rows in the document.

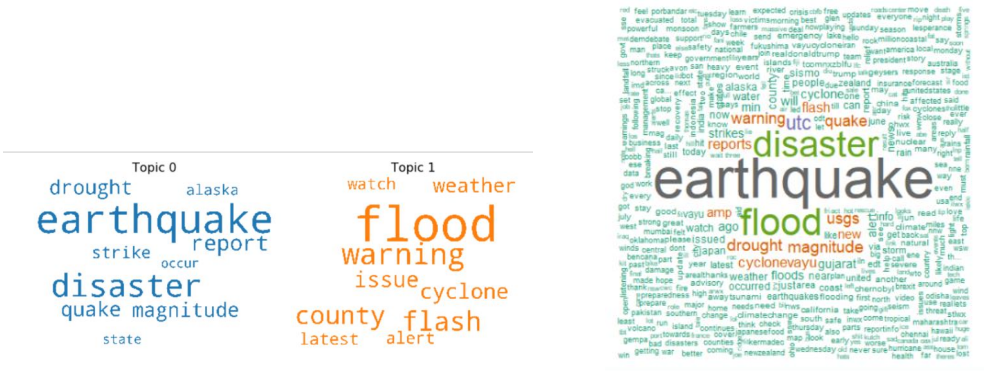


Figure 5: Word cloud for topics

5.4 Implementing different models for Prediction:

In this case, the discussion will be on different models which we have applied for prediction of the disaster trends. As in the previous section we have already created our labelled our

data by two topics i.e. earthquake and flood. Here the prediction is done on the trend of the disaster whether the disaster will remain, or it will be less compared to the previous day. The weather data has a column which contains the description of the weather is combined with the topics to predict for the fore coming disasters. Scikit-learn python library is used which is very efficient tool for data analysis. Before getting a predictive model, we must prepare the data for the model. In a predictive model it is very necessary to convert a categorical data or text into numbers which is done by one hot encoder and label encoder[9]. Here the dominant topic column and the column which has weather description is label encoder using LabelEncoder class from scikit-learn library which is fitted and transformed the text. The whole dataset is split into 75%train dataset and 25% test dataset. Below is the implementation of different models which is later used for the comparison:

- Naive Bayes: It is one of the simplest supervised machine learning algorithms which works on conditional probability theorem. Since here for prediction we must calculate the frequency of the word and along with we are using the discreate weather data, so we have considered using Multinomial naive Bayes. It is executed by splitting the data into 75% train and 25% test data. It is developed by sklearn library where MultinomialNB() is used.

- Random Forest:It is one of the ensemble tree classifiers. It is implemented by using RandomForestClassifier() which is present in the sklearn library.It is trained on the training data where 75% is considered as train data and 25% as test data. The accuracy is obtained by taking number of estimators as 200.

- Support Vector Machine: SVM is one of the supervised learning algorithm which is used as a non-linear probabilistic binary classifier.It helps in finding the plane that separates the classes accurately and efficiently. It is implemented and executed by splitting the data set into 75% train data and 25% test data. SVC() of the sklearn library is used for the implementation.

- Recurrent Neural Network: It is one of the deep neural networks which is implemented by using Keras python library. Sequential () is used to build it where 3 layers are used, in which 2 activation layer uses real and the third layer which is the output layer uses sigmoid. The other hyper-parameter which is used is loss = binary_crossentropy, optimizer = adam and metrics is accuracy. We have considered a drop out of 0.5 in the case to avoid the case of over-fitting and under-fitting. In case of training, the network is re-adjusted and neurons which are newly added are dropped out but in case of the testing, the weights are multiplied with the probability of the dropout associated with the units.

- LSTM:It is one of the artificial recurrent neural networks which is implemented by using Keras and TensorFlow. Sequential () is implemented here by using the embedding layer, 2 LSTM units and last is the activation layer where SoftMax is used as the activation function. There is a dropout (0.5 for the hidden layer and 0.2 for the input layer) included in each layer to reduce overfitting and improve the performance of the model. The other hyper-parameter used is loss

[9]https://medium.com/@contactsunny/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621

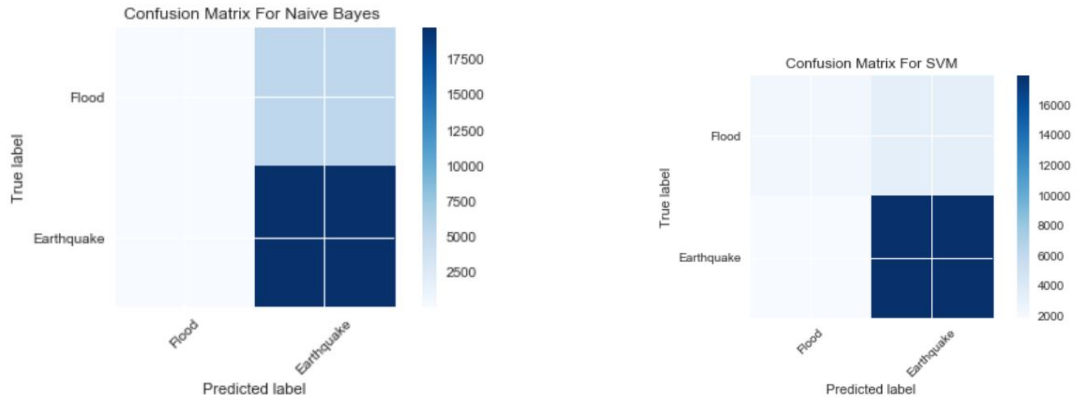
= sparse_categorical_crossentropy, optimizer =adam and metrics is accuracy. For faster execution, google colab (free jupyter notebook environment) is used which runs on the cloud with GPU accelerator.

6 Evaluation

There are many methods in which we can evaluate our model. As it was mentioned earlier that the dataset was divided into training and testing so for validation of the model the testing dataset is used. Below are the evaluation methods which is used in the models:

6.1 Evaluation of Naive Bayes

In all this model we are calculating the accuracy, precision, sensitivity, and specificity of the model. It is done by calculating the prediction value of the testing data and then evaluating the parameters or the metrics of the model. The accuracy of the model is found out to be 78.2%, precision is 78.3%, whereas the sensitivity and the specificity is 99.78% and 1.81%. Also, the confusion matrix which is seen is figure 6(a) is plotted which tells about the performance of the models where the x-axis denotes the predicted label and y-axis is the true label. We can even conclude that prediction related to the earthquake is done correctly than flood which has seen to perform very poorly.



(a) Naive Bayes (b) SVM

Figure 6: Confusion Matrix for NB and SVM

6.2 Evaluation of Random Forest

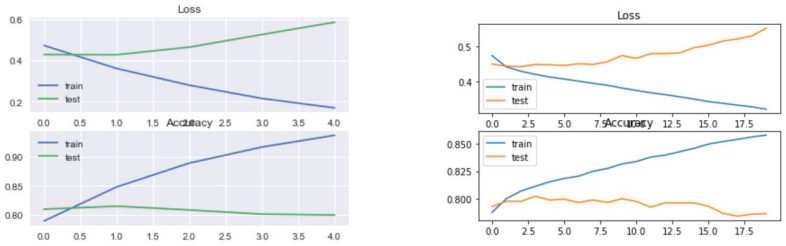
The accuracy of the model is 79.3% and the precision is seen as 81%. The true positive rate or sensitivity of the model is 97% and the specificity of the model is 19.2% which means that actual value predicts negative very less percentage.

6.3 Evaluation of SVM

In the case of SVM, our model gives 80% accuracy. The model is 84.7% precise when predicting positive instances. The model is 41.9% selective in predicting positive instances. The positive rate or recall is 90.7% for the model. The confusion matrix for SVM is seen in the figure 6(b), from which we can see that prediction of flood and earthquake is good enough than previous models.

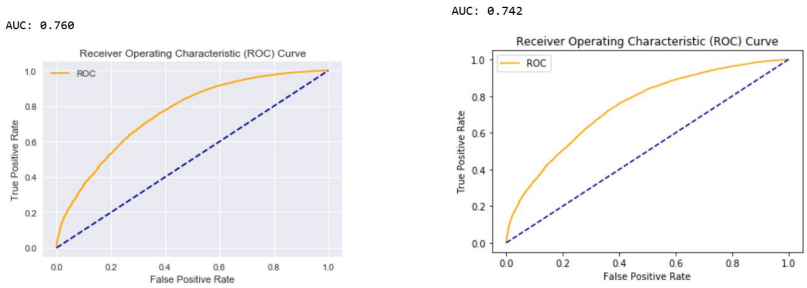
6.4 Experiment done on RNN

The test has been calculated multiple times by varying different hyperparameters. The epochs were varied from 5 to 1000 and tested and the accuracy varied from 79.8% to 81%. It is even evaluated by doing 10-fold cross validation test. StratifiedKFold () from scikit-learn is used to split the dataset to 10 folds. The number of instances in every class in a fold is balanced. The accuracy which we got is 80.6% which is different from accuracy without k-fold by 0.5%.
Figure 7(a) describes the learning curve which shows the loss curve and accuracy curve for training and testing data. It tells about the performance of the model each epoch. From loss curve, we can see that it has a good learning rate and from accuracy curve, we can conclude that it has the presence of some over fitting as the graph between train and test is high.
A bar graph is plotted between the actual value and the predicted value. Even the evaluation metrics are calculated for the model where the sensitivity is 55% and precision is 47.3%. The ROC curve(Figure 8(a)) is plotted for the model which is seen is a figure 10 where the x-axis is (1- specificity) and the y-axis is sensitivity. The AUC is found out to be 0.76 which means it falls under the fair band of the model.



(a) RNN (b) LSTM

Figure 7: Learning Curve for RNN and LSTM



(a) RNN (b) LSTM

Figure 8: ROC Curve for RNN and LSTM

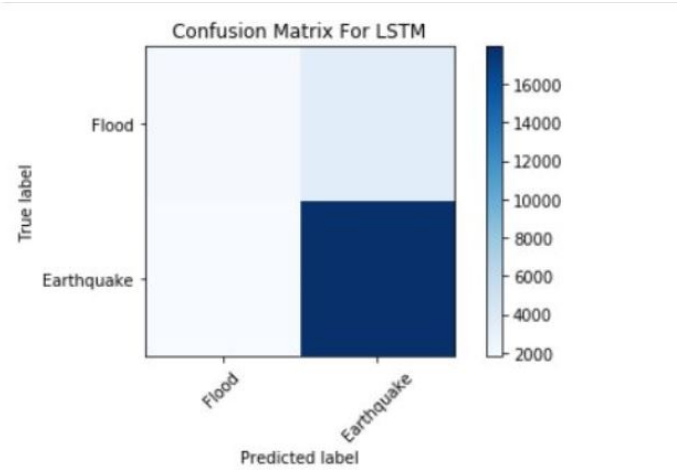


Figure 9: Confusion Matrix of LSTM

6.5 Experiment done on LSTM

For evaluation of LSTM two hidden LSTM layers are used which is ideal of the model. It is tested several times with different hyper parameters and the final model is created with an accuracy of 80%. 10-Foldcross validation test is even done to verify the accuracy with 5 epochs which resulted to 79.8%. In figure 7(b) we can see the learning curve for loss and accuracy. The evaluation metrics were calculated where 90% positive cases were predicted correctly which determines the sensitivity and the specificity were found to be 35%. The AUC is 0.74 which is even a good fit for the model and ROC curve is plotted in figure 8(b). The confusion matrix for the model is even plotted Figure 9.

6.6 Comparison between different models

In this section comparison between all the models (Naive Bayes, Random Forest, SVM, RNN, and LSTM) is done in figure 9.From the diagram below we can see the specificity of RNN is more than any other model which proves that it is more specific to negative values. As RNN, LSTM and SVM have the same accuracy. Therefore, the neural network is better than others if the hyperparameters are tuned properly.

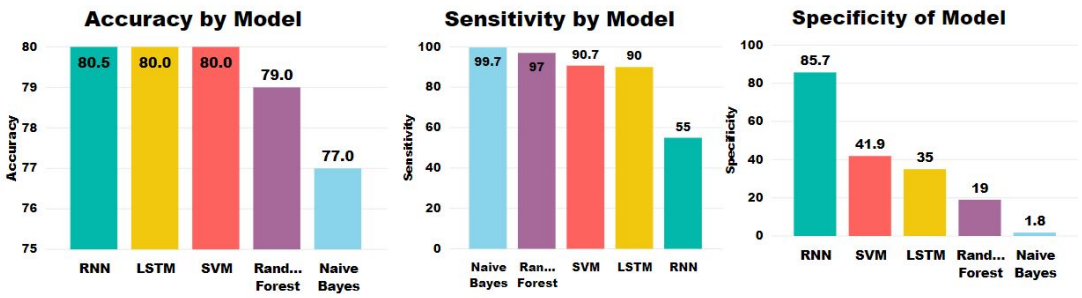
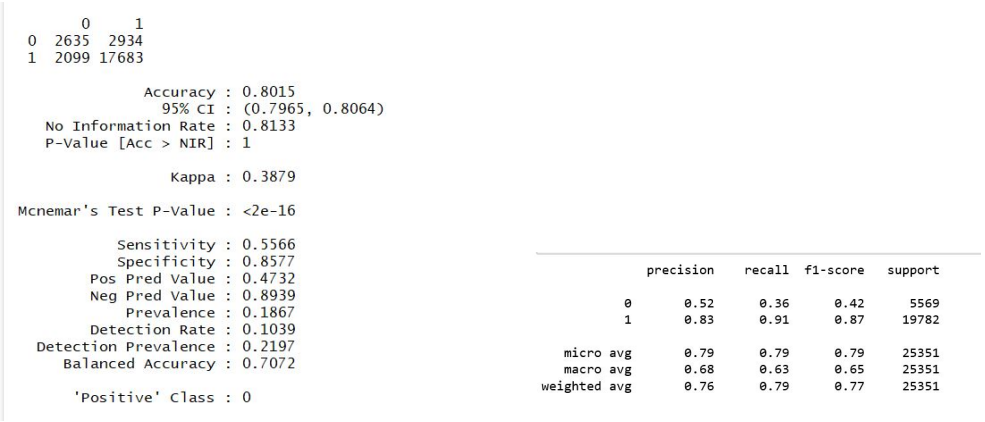


Figure 10: Evaluation Metrics

6.7 Discussion

The research which is done in this paper to collect the tweets and predict disaster, effectively uses the social network. Several experiments were conducted on the model which was created from neural network with different hyper parameters. Choosing a right parameter is very much necessary to get a fit model. In case of LSTM 2 to 3 hidden layer is ideal for prediction which is implemented in proposed model. The dense layer after every hidden layer reduces the over fitting issues. Still, from the learning curve it is seen that the training accuracy and validation accuracy are not near each other, which makes the model to have chances of over-fitting. It should be removed by increasing the number of epochs or iterations[10].



```
      0     1
0  2635  2934
1  2099 17683

               Accuracy : 0.8015
                 95% CI : (0.7965, 0.8064)
    No Information Rate : 0.8133
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3879

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5566
            Specificity : 0.8577
         Pos Pred Value : 0.4732
         Neg Pred Value : 0.8939
             Prevalence : 0.1867
         Detection Rate : 0.1039
   Detection Prevalence : 0.2197
      Balanced Accuracy : 0.7072

       'Positive' Class : 0
```

(a) RNN

	precision	recall	f1-score	support
0	0.52	0.36	0.42	5569
1	0.83	0.91	0.87	19782
micro avg	0.79	0.79	0.79	25351
macro avg	0.68	0.63	0.65	25351
weighted avg	0.76	0.79	0.77	25351

(b) LSTM

Figure 11: Evaluation metrics of RNN and LSTM

The objective of the research was to check how efficient the integration of topic modelling and deep neural network is, which was achieved as the accuracy of RNN is higher than other techniques. Specificity of the model is higher than other model which means that it predicts the negative values correctly. Also the AUC is 0.76 which makes it a best fit. In the proposed model the system could find an accuracy of 80.5% with a combination of LDA and recurrent neural network which is better than any other ML techniques. Previously Singh and Saxena (2016) had implemented a hybrid model using KNN and HMM where prediction was done on web-based data and accuracy was found to be 90%. As the data was collected from twitter so there is more percentage of noise in it which decreases the performance of the model. Also it is seen that Bouazizi and Ohtsuki (2016) had done the same twitter classification using random forest and got an accuracy of 56.9% whereas from our model we got it as 79%. A similar type of approach for disaster classification was done by Verma et al. (2011) where Naive Bayes was used for prediction of situation based event from tweets disaster where the accuracy was found out to be 80% where as in the proposed model it is found out to be 78%.
The framework which is created makes social media as a tool for the help of the general people. The classification result (figure 10) of the models with neural network is even calculated which proves how efficient the proposed model is. Various evaluation test are conducted like cross-validation and k-fold cross validation. The result of the test are seen

[10]http://pmarcelino.com/how-to-read-learning-curves-and-why-do-we-need-them-in-deep-learning/

in figure . We can that the mean of the test is highest and the performance metrics each time(k) are very close to each other. Hence, the model is not overfit and suitable.

7 Conclusion and Future Work

In the research paper, the implemented model efficiently integrates the topic modelling and deep learning methods for classifying and predicting the trends in disaster affected area. The model will surely increase the knowledge and improvement in the field of early prediction. RNN and LSTM are best suited for prediction which gives an accuracy of 80.5% and 79.9% respectively. The sensitivity and specificity value being maximum proves that the true positive rate and true negative rate is good than any other model which has less specificity value. By increasing a greater number of hidden layers and increasing the epochs we can even more improve the accuracy. The implemented model can be helpful in improving the current disaster management system. The main objective of the paper is to help people generating early warning during disaster situation which will save millions of lives. Most difficult step was in extracting the features from a noisy dataset collected from twitter and identifying the correctly classified disaster. The limitation of the project is the unsupervised data being very noisy and contains many unnecessary information which is not helpful. Also, while using the neural network the system has much chance of going for over fitting or under-fitting condition for which proper hyper-parameters should be used.

7.1 Future Work

The research can be continued by extracting more fine-tuned features(needs of the people during calamity situation) from the twitter by different advanced techniques.Also for prediction few new techniques i.e. temporal convolution network(TCN) can be used , where the architecture is framed by taking the best features from CNN and RNN. TCN takes less memory for training as compared to any other neural network which is very useful.Previously CNN was implemented for prediction(Nguyen et al.; 2017). Also the research can be extended by use of iot and by collecting data from many different data sources.

8 Acknowledgement

I would like to acknowledge everyone who helped me in my academic career, specially my mentor and guide Dr. Anu Sahni for her support, supervision and guidance through out my research work.Her advice and suggestion helped me a lot in completing this thesis.

References

Ahmed Shaiba, H., Sulaiman Alaashoub, N. and Ahmed Alzahrani, A. (2018). Applying machine learning methods for predicting sand storms, pp. 1–5.

Ashktorab, Z., Brown, C., Nandi, M. and Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response, *ISCRAM*.

Bande, S. and Shete, V. V. (2017). Smart flood disaster prediction system using iot amp;amp; neural networks, *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pp. 189–194.

Benny, A. and Philip, M. (2015). Keyword based tweet extraction and detection of related topics, *Procedia Computer Science* **46**: 364–371.

Bhatia, A., Pasari, S. and Mehta, A. (2018). Earthquake forecasting using artificial neural networks, *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-5**: 823–827.

Bouazizi, M. and Ohtsuki, T. (2016). Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter, *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6.

Cuesta, A., Barrero, D. and R-Moreno, M. (2014). A framework for massive twitter data extraction and analysis, *Malaysian Journal of Computer Science* **27**: 50–67.

Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A. and Chakraborty, B. (2018). A review on application of data mining techniques to combat natural disasters, *Ain Shams Engineering Journal* **9**(3): 365–378.

Hajjem, M. and Latiri, C. (2017). Combining ir and lda topic modeling for filtering microblogs, *Procedia Computer Science* **112**: 761–770.

JinXin, M., Huiling, L., Hankz, Z. and Zhuo, H. (2018). Combining deep learning and topic modeling for review understanding in context-aware recommendation, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* **1**.

Jo, Y., Lee, L. and Palaskar, S. (2017). Combining lstm and latent topic modeling for mortality prediction.

Kansal, A., Singh, Y., Kumar, N. and Mohindru, V. (2015). Detection of forest fires using machine learning technique: A perspective, *2015 Third International Conference on Image Information Processing (ICIIP)*, pp. 241–245.

Kireyev, K., Palen, L. and Anderson, K. (2014). Applications of topics models to analysis of disaster-related twitter data.

Lohumi, K. and Roy, S. (2018). Automatic detection of flood severity level from flood videos using deep learning models, *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)Sendai, Japan,* pp. 1–7.

Nguyen, V. Q., Yang, H., Kim, K. and Oh, A. (2017). Real-time earthquake detection using convolutional neural network and social data, *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pp. 154–157.

Ogie, R. I., Rho, J. C. and Clarke, R. J. (2018). Artificial intelligence in disaster risk communication: A systematic literature review, *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pp. 1–8.

Reynolds, B. and Seeger, M. (2005). Crisis and emergency risk communication as an integrative model, *Journal of health communication* **10**: 43–55.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development, *IEEE Transactions on Knowledge and Data Engineering* **25**(4): 919–931.

Saltz, J., Shamshurin, I. and Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects, *Journal of the Association for Information Science and Technology* **68**(12): 2720–2728.

Shekhar, H. and Setty, S. (2015). Disaster analysis through tweets, *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* **0**(0): 1719–1723.

Singh, A. and Saxena, A. (2016). A hybrid data model for prediction of disaster using data mining approaches, *International Journal of Engineering Trends and Technology (IJETT)* **41**(7).

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A. and Kapoor, K. K. (2017). Event classification and location prediction from tweets during disasters, *Annals of Operations Research* .

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. (2010). Short text classification in twitter to improve information filtering, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, ACM, pp. 841–842.

Stowe, Kevin J. Paul, M. . P. M. . P. L. . A. K. (2016). Identifying and categorizing disaster-related tweets, *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media* pp. 1–6.

Velev, D. and Zlateva, P. (2012). Use of social media in natural disaster management.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., H. Martin, J., Martha Palmer and M. Anderson, A. S. . K. (2011). Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency., *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain,* .

Yaghmaei, N. (2019). Disasters 2018: Year in review (cred crunch issue: 54).

Yang, S. and Zhang, H. (2018). Mining of twitter data using a latent dirichlet allocation topic model and sentiment analysis.

Zaheer, M., Ahmed, A. and Smola, A. (2017). Latent lstm allocation: Joint clustering and non-linear dynamic modeling of sequence data.

Zahra, K. and Purves, R. (2017). Analysing tweets describing during natural disasters in europe and asia.