Monitor and Assess Natural Disaster using Twitter Data in Data Analytics

Dr.S. Padmapriya[1] K. Sowmiya[2] P. Subhasini[3] T.S.Vijayalakshmi[4]
[1]Professor & Research Coordinator[2,3,4]B.E Student
[1,2,3,4]Department of Computer Science &Engineering
[1,2,3,4]PEC, Chennai, India

Abstract*—*Disaster management is a sequential and continuous process planning. A complex mixture of disasters, ranging from solar flares, cosmic explosions and meteorites, to earthquakes, tsunamis, landslides, floods, hurricanes, droughts, terrorism, wars, and to disasters due to technical failures or human operator faults imperil people, populations, civilization, and humankind. Defending against these threats requires various kinds of endeavors supported by varied tools and large technical and human capabilities. Social media is emerging as important information based communication tool for disaster management. People suffer from unexpected natural and man-made disaster due to lack of awareness. Using machine learning, human language is recognized as machine language and graph is generated for monitoring the tweets and places where affected more affected due to disaster is generated for prevention and this information can be shared in social media by a particular member.

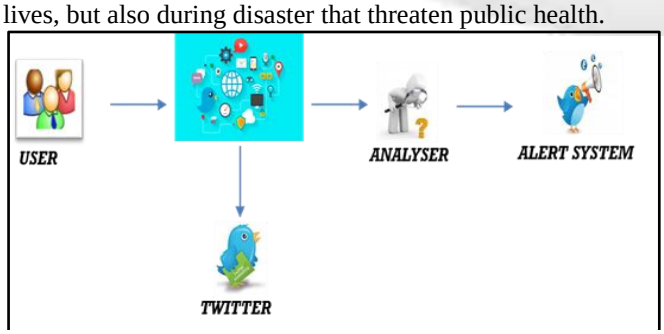
Keywords—Data Collection, Tokenizing, Data Preprocessing, Filtering, Parsing, Machine Learning

I. INTRODUCTION

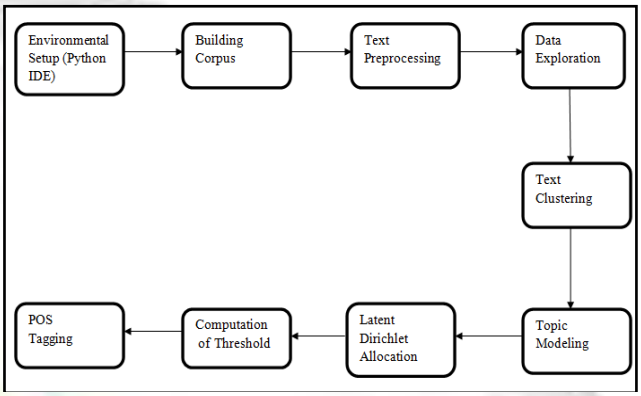
The service to monitor and assess disaster management using social network is based on data analysis that aims to assess the disasters and to spread information among people. In these days, though the sharing of information using social media has become so easy, the prediction of disasters cannot be accurate and this information cannot be shared too. To overcome such things it helps the people to be cautious by predicting the disasters. Then the information can be shared in social media by collecting twitter data sets and analyzing the most frequently used hash tags to produce a graph with a threshold value.

II. PROPOSED SYSTEM

In proposed system K Means clustering is implemented for accurate prediction. The user receives the tweets posted by his/her friends and it will be analyzed. This system helps to monitor and assess the disaster and can prevent loss by sharing the assessed information in social media by a particular member. Clearly, the social media are changing the way people communicate not only in their day to day lives, but also during disaster that threaten public health.



III. BLOCK DIAGRAM



IV. MODULE IMPLEMENTATION

A. Building Corpus

Here, twitter data sets (Chennai floods) are collected using anaconda prompt. Data sets are collected based on the date and time; these data's are stored in csv format. NLTK tool is installed to recognize the human language for applying statistical natural language processing.The hash tags are separated based on the common hash tags used at that particular period of time. Also the places affected by the natural disasters are mentioned in those hash tags. These hash tags are categorized based on the damage, time, area, geographical location.

B. Data Cleansing

The NLTK contains text processing libraries for,

1) TOKENIZING:

In python, NLTK is one of the leading platforms for working with human language data. This tool is used for natural language processing. Tokenizing means generally giving unique identity for each words to define the words. Example the number in the credit cards are given in a unique way to all the members. Not all the credit card have the same number since to show the uniqueness of it. Tokenization can be accomplished by using spaCy library. Using a method function generatetokens(), the tokenization can be achieved.

2) STEMMING:

Stemming is a sort of normalizing method. Many variations will be there in a sentence which carries the same meaning. When both the sentences have same meaning they can be normalized by using stemming method. Parts of speech tagging with NLTK module means labeling words in a sentence as nouns,adjectives,verbs,etc. Many POS Tagging list are NN means Nouns, NNS means noun Plural, JJ means adjective, etc.

3) Removal of Stop-Words:

Generally stop words are the words which we stop on. Many words are used in a sentence which may not be used anywhere. These words can be removed since it takes up more space in the database, so we call these words as stop words. This can be accessed by NLTK corpus withfrom nltk.corpus import stopwords

4) Removal of Punctuation:

Punctuation of words is the most important one in a sentence.Sometimes there may be punctuation marks in unnecessary places. This can be avoided by using this method function, where it removes the unwanted punctuations in the hash tags of the tweets.

5) Split Attached Words:

This is mainly used to split the hash tags in the tweets displayed so that the preprocessing of data is easy to achieve. One or more words are combined together in a hash tag these words can be splitted to arrive at a meaningful sentence and words of hash tags. This is done to recognize human language as a machine language by the NLTK tool.

C. Data Exploration

In this module the words are plotted again to find the most frequently used terms. A few simple words repeat more often than others: 'help', 'people', 'stay', 'safe', etc. Collocations are the words that are found together. They can be divided as bi-grams (two words together) or phrases like trigrams (3 words) or n-grams (n words). These depict the disastrous situation, like "stay safe".

Most frequently appearing Bigrams are:

[(u'pic', u'twitter'), (u'twitter', u'responsive'), (u'medium', u'twitter')]

D. Text Clustering

In this module lots of similar tweets are generated. They can be grouped together in clusters based on closeness or 'distance' amongst them. TF-IDF stands for "Term Frequency, Inverse Document Frequency." It's a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents.TF-IDF method is used to vectorize the tweets and then cosine distance is measured to assess the similarity.

Each tweet is pre-processed and added to a list. The list is fed to TFIDF Vectorizer to convert each tweet into a vector. Each value in the vector depends on how many times a word or a term appears in the tweet (TF) and on how rare it is amongst all tweets/documents (IDF).Before using the Vectorizer, the pre-processed tweets are added in the data frame so that each tweets association with other parameters like user is maintained.

E. Topic Model

Topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies.The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.After the text clustering process topic modeling is implemented for finding the relevant topics. For the topic modeling two procedures are implemented.They are Latent Dirichlet allocation (LDA) and Doc2Vec andK-means.

F. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA.Each document is assumed to be characterized by a particular set of topics LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. Given a dataset of documents, LDA backtracks and tries to figure out what topics would create those documents in the first place.LDA is a matrix factorization technique. In vector space, any corpus the top ten topic is collected from the bag of words. The topics are related words .for example in chennai floods the topics would be 'flood','save','help',food','water'.

G. DOC2VEC AND K-MEANS

Doc2vec is an unsupervised algorithm to generate vectors for sentence/paragraphs/documents. Distributed Representations of Sentences and Documents. The algorithm is an adaptation of word2vec which can generate vectors for words. The vectors generated by doc2vec can be used for tasks like finding similarity between sentences/paragraphs/documents. Here also topics are created according the related action/text. The top six topics are created just as LDA. To get a threshold value sum the score of all words from the retrieved topics and divide by length of score. This threshold value considered as assessed value for predicting natural disasters.

H. POS TAGGING

In corpus linguistics, part-of-speech tagging (POS tagging), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context. i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. POS tagging applied on clusters to getting nouns. Nouns are like places, names, etc. This will analyze and produce where losses occurred more and giving awareness by analyzing the disaster affected areas.

V. CONCLUSION

As social media has reached nook and corner of the earth. They are not only used for entertainment purpose but also it has involved in many sectors like education, business, marketing, creating awareness, etc., Hence it is a great platform for forecasting the disasters and prevent the loss. The proposed system combines the NLTK, clustering, LDA and a threshold value, places where affected more due to disaster is produced for the prediction of disaster. For further enhancement alert system can be implemented for the vast alerting to the community.

REFERENCES

[1] M.V.Sangameswar, Dr.M.Nagabhushana Rao and N.S.Murthy "Twitter Data Analysis on Natural Disaster Management System" International Journal of Engineering Trends and Technology (IJETT), March 2017.

[2] Suraj Singh Chouhan and Ravi Khatri "Data Mining based Technique for Natural Event Prediction and Disaster Management" International Journal of Computer Applications (0975 – 8887), April 2016

[3] Si Si Mar Win, Than NweAung "Target Oriented Tweets Monitoring System during Natural Disasters" Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th international conference, May 24-26, 2017

[4] Miguel Maldonado, Darwin Alulema "System for monitoring natural disasters using natural language processing in the social network Twitter" Security Technology (ICCST), 2016 IEEE International Carnahan Conference , 24-27 Oct. 2016

[5] Horia-Nicolai Teodorescu "Using analytics and social media for monitoring and mitigation of social disasters" Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech 2015