Survey on Detecting Key-Needs in Crisis

Shailesh Gupta[1], Navin Joshi[2], Abhay Gupta[3], Sagar Kulkarni[4]
[1,2,3] *Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India*
[4] *Guide, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India*

***Abstract*- When a crisis occurs, the world springs into action to try and understand what is happening and what help is required. During these times, social media has become a key avenue through which to disseminate information. During this time, the system will classify Tweets into different Category such as volunteering, injury, death count etc. The Project is divided into multiple tasks as follows.**
Task 1: The system will need a large dataset for training and testing phase, so the data from different social media and blogs using API and web-scraping technique will be collected. After that data will be cleaned.
Task 2: After cleaning & authorization of data, Natural Language Processing Technique will be used to extract knowledge (Disaster event information such as earthquake, flood, terrorist attack etc) from data. In this section Text Processing Model such as BoW, TF-IDF, W2V will be used. The system will be trained and tested on data to create model with high accuracy which will predict the event.
Task 3 : After all this process the system will show the output using a suitable interface to the user.

***Index terms*- Crisis, Natural Language Processing, Tweets, Predict, Web scraping.**

I.INTRODUCTION

Whenever somewhere disaster occurs it's information needs to be distributed to the rest of the world to get help. Off late in such condition whenever most of the communication media fails internet remains to be a last resort to spread the information about the situations. People use social media channels to share updates regarding the event. But since this data keeps flowing a system is needed to extract information from it to make more use of data and make help available as quickly as possible. All the updates need to be classified into different types to fasten the process of help. An automated system which can classify the data of disasters into different categories in real time as data comes in without manual help needs to developed.

II. LITERATURE SURVEY

A literature review is an objective, critical summary of published research literature relevant to a topic under consideration for research. Thirteen published articles have been referred in order to create a firm base about the project.
Following is a brief overview of all the thirteen papers that have been referred.
Event Extraction from Newswires and Social Media Text in Indian Languages [1]
Author - Pattabhi RK Rao,Sobha Lalitha Devi, Year - 2018
Description - The task is to identify various events such as sport events, terrorist events, natural disasters, crime events, corporate events, political events, accidents etc in a given text. The texts can be in various sources such as Newswires, Blogs, and Micro blogs. In this technique we extract the data from twitter and then we create model using machine learning which provides required keyword events.
Information retrieval for microblogging during disaster [2]
Author-Moumita Basu, Saptarshi Ghosh, Kripabandhu Ghosh, Year - 2018
Description - Authors are making a system which is used to identify the disaster occur at some places using a different social media. They are also checking for the fact that news will be fake or genuine. They are using the binary class classification for identifying the fact message or non-fact message.
Entity extraction from social media indian language [3]
Author-Chintak Mandalia, Memon Mohammed Rahil, Manthan Raval, Sandip Modha
Description - This research is about information retrieval from the indian language social website. They are extracting the different entity such name, location ,organization name etc. Because in India

most of the people are comfortable with their own mother tongue.

Social Media Mining for Post-Disaster Management – A case study on Twitter and News [4]

Author- Banujan. K1, Banage T. G. S. Kumara, Incheon Paik, Year - 2018

Description - The Twitter posts and news posts from Twitter API and News API respectively,using predefined keywords relating to the disaster. Those posts were cleaned and the noise was reduced at the secisaster type and geolocation of the posts by usiond stage. Then in the third stage, they got the dng Named Entity Recognizer library API. As a final stage, they compared the Twitter data with news data to give a rating for the trueness of each Twitter post.

Sub-Event Detection from Tweets [5]

Author - Satya Katragadda , Ryan Benton , Vijay Raghavan, Year - 2017

Description - Authors are assuming a tweet as a single entity and treat it as such during the detection process.To capture the change in information over time, they extend an already existing Event Detection at Onset algorithm to study The evolution of an event over time. they introduce the concept of an event life cycle model that tracks various key events in the evolution of an event. The proposed unsupervised sub-event detection method uses a threshold-based approach to identify relationships between sub-events over time. These related events are mapped to an event life cycle to identify sub-events.

Information Retrieval from Legal Documents (IRLeD) [6]

Author - Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, Saptarshi Ghosh, Year - 2017

Description - The focus of 2017 IRLeD Track was Information Retrieval from legal documents. There were two tasks: (i) Catchphrase Extraction:Here task was to find important legal terms in case documents. Generally tokenization, POS tagging and Term frequency were used to determine keywords and then trained in different algorithms.(ii) Precedence Retrieval: Here task was to find out similar cases that happened prior. The solutions considering citation contexts performed well.

Detecting Key Needs in Crisis [7]

Author - Tulsee Doshi, Emma Marriott, Jay Patel, Year - 2017

Description - Authors have classified tweets into key categories like volunteer services, Displaced people and activations etc. The dataset consists of tweet IDs from 19 disasters representing 8 types of crisis in Spanish and English.After preprocessing the common Twitter slang was replaced with complete terminology.A model based on Feed Forward Classifier on single tweets outperformed LSTM on single tweet as well as sequence of tweets. The accuracy increased with Word2Vec vectors instead of predefined vector set like Glove.

From Clickbait to Fake News Detection [8]

Author - Peter Bourgonje, Julian Moreno Schneider, Georg Rehm, Year - 2017

Description - The authors have proposed solution for the detection of the stance of headlines with regard to their corresponding article bodies. The dataset available had articles classified as clickbait, discuss, agree, disagree. In solution lemmatised header & articles are compared to find matching n-grams. A score is calculated based on it which if above some threshold the pair is taken to be related.

A Comparison of Classification Models for Natural Disaster and Critical Event Detection from News [9]

Author - Tim Nugent, Fabio Petroni, Natraj Raman, Lucas Carstens and Jochen L. Leidner, Year - 2017

Description - Authors have presented contrastive study of document-level event classification of a range of seven different event types. They have compared classification approaches such as SVM, RF, CNN and HAN. Their baseline model uses matches to positive and negative gazetteer lists assembled for each event type. Also each match is weighted by inverse distancing from beginning of document. The results with predefined word vectors were better than those likes of TF-IDF.

A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets [10]

Author - Kunal Chakma, Amitava Das, Year - 2016

Description - Authors have given way to collect a corpus of Code-Mixed Indian Tweets via Twitter API by defining a bag of words containing Hindi words in Roman Script. They have also cited different problems faced by current IR systems in handling Code-Mixed data.

Detecting Paraphrases in Indian Languages (DPIL) [11]

Author - M.Anand Kumar, Shivkaran Singh, B. Kavirajan, and K. P. Soman, Year - 2016
Description - This project is also based upon the concept of information retrieval of indian language news sources and identifying their meaning.They divide the task into two parts 1.sentence from news paper classify them as paraphrases or non-paraphrases. 2. They are giving two sentences and task is to identify whether they are completely equivalent (E) or roughly equivalent (RE) or not equivalent (NE).
Detecting Real-time Events using Tweets [12]
Author - Koichi Sato , JunboWang, ZixueCheng, Year - 2016
Description - A scheme is proposed which can detect what happens in the real world in real time only by analyzing tweets as Big Data and let the user know the event. To this end,the following problems have to be solved. They are a) quantifying importance of words accurately and b) evaluating the quantified values dynamically. As the solutions for the problems, two new methods are proposed which are the Extended Hybrid TF-IDF and the Remarkable Word Detecting Method, and they are used in the proposed scheme.
Location Identification for Crime & Disaster Events by Geoparsing Twitter
[13]
Author - Nikhil Dhavase, Prof. A. M. Bagade, Year - 2014
Description - Authors are using a method of geoparsing through which they can find the location in the text this can be used at the time of emergency. Use of social media during such crisis events has been rapidly increasing all over the world, as well as in India. Extracting the location information to the level of streets & buildings will help to detect the exact location of the event; this is done with the help of NLP methods.

III. PROPOSED SYSTEM

3.1 Overview

The Section presents an Overview and Description of techniques used for the system. In this project the data will be collected from Twitter API or different social media platform using web scraping. After collecting the data it will be processed to convert into suitable form so that model can be trained on it. The model will be trained using different disaster data to classify different categories. After building model, test data will be given to model to test it's accuracy. Model performance and accuracy will be also visualized. The process which can be implemented during project are shown.

Existing System Architecture [1]
Existing solutions only classify tweets whether they are related to a disaster crisis or not. In Existing Architecture, they are not using as much as large amount data so that accuracy or predictability of will be less. [1]

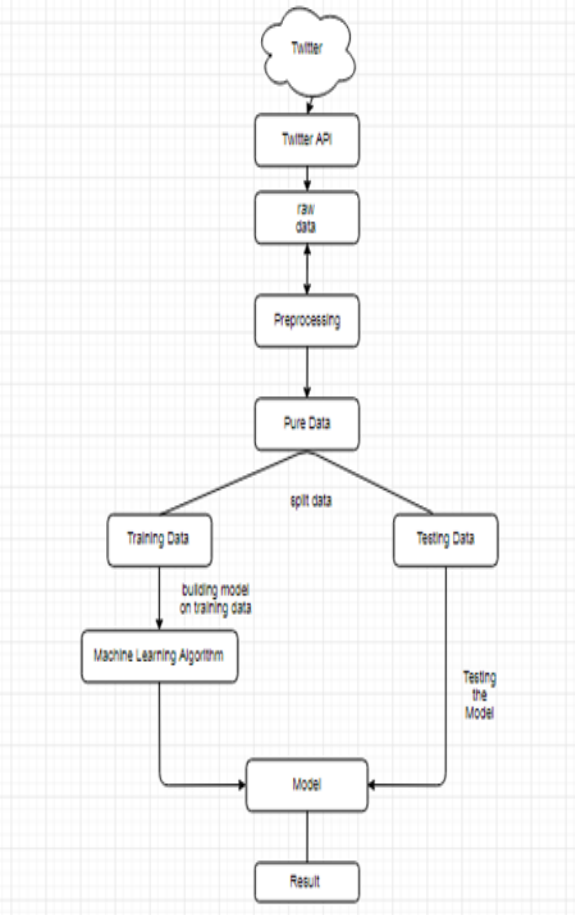

Figure 1 Existing system architecture

Proposed System Architecture
In our proposed system we try to extract key needs from the classified tweets. We will try different algorithm and find their accuracy and performance. The model which gives best result That will be implemented.

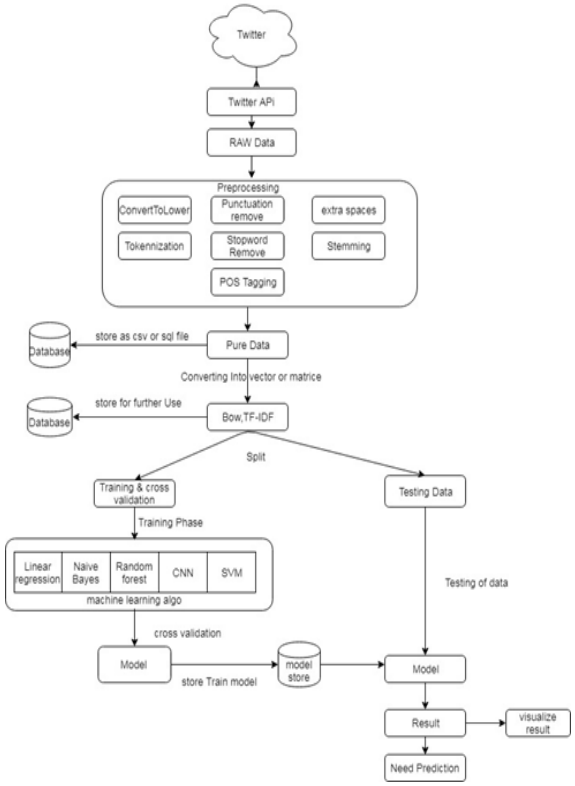

Figure 2 Proposed system architecture

3.2 Hardware and Software Specifications

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 1 and Table 2, respectively.

Table 1 Hardware details

Type	Minimum Requirement
Processor	1.8Ghz Intel/Amd
RAM	4GB
Graphics	512MB
Programming Language	Python 3.6, Html, Css
IDE	IDLE/jupyter
Operating system	Windows 7 and up.
Database	Mysql
Browser	Chrome

REFERENCES

[1] Pattabhi RK Rao,Sobha Lalitha Devi, "Event Extraction from Newswires and Social Media Text in Indian Languages", FIRE2018.

[2] Moumita Basu, Saptarshi Ghosh,Kripabandhu Ghosh, "Information retrieval for microblogging during disaster" , IRMiDis FIRE2018.

[3] Chintak Mandalia, Memon Mohammed Rahil,Manthan Raval,Sandip Modha, "Entity extraction from social media indian language", FIRE2018.

[4] Banujan. K1, Banage T. G. S. Kumara, Incheon Paik, "Social Media Mining for Post-Disaster Management – A case study on Twitter and News", International Research Conference on Smart Computing and Systems Engineering - 2018.

[5] Satya Katragadda, Ryan Benton, Vijay Raghavan, "Sub-Event Detection from Tweets", International Joint Conference on Neural Networks (IJCNN) 2017.

[6] Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, Saptarshi Ghosh, "Information Retrieval from Legal Documents (IRLeD)", FIRE-2017-IRLeD.

[7] Tulsee Doshi, Emma Marriott, Jay Patel, "Detecting Key Needs in Crisis" , Stanford Education 2017.

[8] Peter Bourgonje, Julian Moreno Schneider, Georg Rehm, "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles", DFKI GmbH, Language Technology Lab

[9] Tim Nugent, Fabio Petroni, Natraj Raman, Lucas Carstens and Jochen L. Leidner, "A Comparison of Classification Models for Natural Disaster and Critical Event Detection from News.", IEEE Big Data DSEM Workshop.

[10] Kunal Chakma, Amitava Das, "A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets", 2016.

[11] M. Anand Kumar, Shivkaran Singh, B. Kavirajan, and K. P. Soman, "Detecting Paraphrases in Indian Languages (DPIL)", FIRE 2016.

[12] Koichi Sato,Junbo Wang,Zixue Cheng, "Detecting Real-time Events using Tweets", IEEE Symposium Series on Computational Intelligence (SSCI) 2016.

[13] Nikhil Dhavase,Prof. A. M. Bagade, "Location Identification for Crime & Disaster Events by Geoparsing Twitter", International Conference for Convergence for Technology-2014.