FloodKnow – Information Graph

By

Nitin Chaube

Justin Nadar

Aniruddha Chaudhari

(Under the guidance of Prof. Supriya Khamoji)

Abstract

This research is conducted with objective of creating and presenting a system called FloodKnow, which is data extraction and visualization framework with capability of filtering, classifying and visualizing online social media data, to provide accurate information about ongoing events for decision making in disaster scenarios such as flood. Now a days social media platforms such as Twitter have received a great deal of attention as human sensors, revealing the information regarding the ongoing events such as natural disaster. Filtering and Analyzing disaster related situation updates from a large pool of data is still a challenge and has attracted a lot of research attention. Research works have been conducted in filtering and analyzing the online social media content for natural disaster but a complete system to provide a real-time a summarized information to the decision makers

is still at its infant stage. In this work a complete framework is proposed which uses Twitter as a sensor to gather the information of ongoing disaster event such as flood, information is filtered and processed using natural language processing, Deep learning for classifying priority tweets and knowledge graph for real-time visual representation of gathered information for the decision makers. At the end of this process, the data is presented graphically which narrate the local and global storyline of the affected stakeholders and hotspot maps to depict the area in which natural disaster are concentrated. Floodknow has proved to be on par with state of art disaster extraction systems, which helps in deep understanding the disaster development by visual approach. Among the disaster domain, we represent the knowledge graph for flooding event.

Introduction

Flood management (Disaster) involves three phases: Preparedness, Response, and Recovery. The preparedness phase comes into place when an emergency or a disaster is likely to take place. It corresponds to preparatory activities prior to a disaster to save lives and help response and rescue operations, such as stocking food and water, posting emergency contacts, and preparing evacuations. With plans and strategies developed beforehand, the response phase mainly puts them into action. Response activities happen during a disaster, usually involving evacuating threatened areas, search and rescue efforts, shelter management, and humanitarian assistance. After a disaster, the recovery phase refers to repair and reconstruction efforts to return to a normal or even better functionality level. Recovery actions usually include debris clean-up, precise damage assessment, and infrastructure reconstruction, as well as financial assistance from government agencies and insurance companies.

In social, political and even when tragedy happens, social media has played an important role in recent years, and it has also become the most influential contact media. In most situations, traditional communication fails during a disaster. During the tragedy, social media such as Facebook, Tweeter, Google, and so on made communication easier in this sort of situation. Gives the real-time information about the pre and post disaster/flood to avoid the losses. Messages and posts in the social media give more information to the people who are living in that area. This real time nature of social media makes it an attractive tool for disaster management, as both victims and officials can put their problems and solutions at the same place in real time. It provides the location and severity of the disaster so that the resource and rescue teams can be deployed to the stakeholders. Researchers have started to explore Twitter as a tool for disaster management.

The large amount of data generated during a short span of time during disaster. Even when there are resources available for help, many requests could go unnoticed as they are not rooted through proper channels. By collecting and analyzing such

generated requests will help the government authorities and NGOs to deploy the rescue teams.

During the time of disaster, people often post their real-time experiences and local news on social media to inform others. Many rescue agencies monitor this type of posts regularly to locate disasters and reduce the risk of lives. However, it is impossible for humans to manually check the large number of posts and identify disasters in real-time. For this purpose, many research works have been proposed to automate the filtering of huge mass of messages.

Twitter is a micro blog where crowd approved knowledge such as text messages, photographs and audio clips are sent. Since users write small messages, they regularly send it and check for retweets from others. Twitter updates for disastrous events includes storms, heavy rainfall, earthquakes, fire, traffic jams etc. A lot of work has been done to detect events, both social as well as disastrous from Twitter messages. Most work for disastrous event detection systems are confined to detect whether a tweet is related to the disaster or not, based on textual and visual content. The disaster related tweets are further used to warn and inform people about precautionary measures. These tweets are viewed not only as an awareness platform, but a place where people can ask for help during disaster. The tweets asking for help need to be separated from other tweets related to the disaster. These tweets then can be used to guide the rescue personnel.

Many studies show that huge amount of accurate and real time data is obtained from Twitter than collected through traditional data collection methods. Despite of its privileges, this real-time information provides interesting challenges, due to very high volume of messages, lack of metadata, verifying the veracity of the posted information, and missing geolocations. Disaster respondents do not have enough bandwidth and time to manually monitor huge amount posts collected on twitter during the situations like disaster and take appropriate action. Understanding the purpose of the posts such as is it about requesting for rescue operation, reporting about the situation, calling for donation, infrastructure damage, praying for affected people and understanding the exact location of the people who need help is the biggest challenge. This huge amount of information gathered from twitter has to be classified into relevant and non-relevant information. Relevant class of information can be further categorized into high priority information such as injured and help required and low priority such as praying for affected.

Social media such as Twitter provides three location information fields for sharing a user's location:

(1) User location;

(2) Place name and

(3) Geo-coordinate.

The user location field has 280-character in which the user can write their home location information while creating their profile. This field is compulsory to the user and the user can write any arbitrary words or leave it blank. In many instances, user does not want to reveal their location so they write meaningless words that might not refer to any location name. It is analyzed that 34% of users do not want to reveal their "user location" Information . Users use city level or below city level location names in their user location field. However, this field cannot be treated as the current location of the user as it is entered at the time of creating their profile as it is most of the time not updated by the users regularly. The second field is for the "place name," which can be attached to a tweet message when it is posted. The place name is represented by a location name with an array of the latitude-longitude pair in the form of the location's boundary coordinates. But these place names are predefined on the Twitter database, but it does not provide granular location information . Research found that only 47.33% of tweets contain place names. However, 12% of those place names are not correct in terms of their spatiotemporal information . The third field provided by Twitter is for the "geo-coordinates" (geographical footprints of latitude and longitude) which can be attached at the time of posting a tweet using a GPS- (Global Positioning System) enabled device. Most of the researchers have considered geo-coordinates as the most explicit information [10][11][12]. However, most users do not tweet with geo-coordinate information. Some researchers determined that only 0.42%, 3.17%, and 7.90%of tweets respectively are geo-tagged [6][7][13][14]. Researcher further reported that although geo-coordinates are the most precise location information, they are not always authentic in terms of their spatiotemporal information if the tweet is posted from some other location . Hence, all three location information fields, available in twitter and user profiles, have their own constraints and cannot be completely relied on.

Now a days many users provide location information in tweet texts which is vitally important and authentic source of geographic evidence as it represents the location information of any event or user during emergencies. Most commonly used methods for location extraction are gazetteer-based approach and the Named Entity Recognition (NER) based approach Gazetteer is a location names corpus (e.g., GeoNames, http://geonames.org). In the gazetteer-based approach, the words of tweets are searched upon in the gazetteer to find the location names. However, there are some problems with this approach:

(i) the unavailability of gazetteers for all the locations and

(ii) a location name mentioned in the tweet may have some other non-geographic meaning in the context of a text e.g., the word "Pentagon" may refer to a location name in US or it may also be used in another context. The other problem with this approach is the geo-ambiguity (distinct locations have the same name, e.g., Chamba in India is a town in Himachal Pradesh and Uttarakhand also. The second approach is Named Entity Recognition (NER). The NER technique generally tokenizes the tweet into different entities using language-specific part-of-speech tagging.

There are numerous cooperation teams which are prepared to help the people at the time of disaster, but the cooperation among different teams is weak. Besides, literature emphasizes the establishment of centralized cooperative emergency system, which can serve every needy request during disaster. For this a graphical representation which can visualize every detail of disaster information so that no request can go unnoticed and can have cooperation among all the rescue teams . Some study shows that disaster include heterogeneous data and lack interoperability. In particular, the case of social media data related to disasters, there are several issues, where the source and format of data are different because huge amount of data are collected by different organizations. This study proposes a visualization tool, knowledge graph to resolve the heterogeneity among various disaster data. Knowledge graph is used to assist, solve, and manage disaster problems .

Knowledge graph can be designed for a data extraction system with the capability of filtering social media data, to improve community resilience and decision-making in disaster scenarios. knowledge graphs (KG) can be used to connect insights, possible to generate real-time visual information about such disasters and affected stakeholders, to better the crisis management process .

Aim

To better the crisis management process, by disseminating such information to both relevant authorities and population alike, this generated Information Graph can provide real-time visual information about disasters like flood, earthquake and wildfire.

Objective

- To identify a tweet whether the tweet is related to disaster or not.
- To do a multiclassification on a dataset related to different classes.
- To use different models and find the best one giving high accuracies.
- To make a model for NER to extract useful information from a tweet e.g., locations .
- To extract tweets from twitter related to disasters from Indian regions.
- To Create Information Graph for better visualization of all the information.

Problem Statement

Problem Statement is to create a Information Graph to represent the different natural or man-made disasters occurring in India in a dynamic graphical representation so that government bodies or any social bodies could get all the information about the disaster and provide necessary relief to the affected people. We are making our system model using tweets of twitter on the disasters.

Our Problem statement consists of making of three deep learning model for classifying the tweets as informative or not, to do the To Create Information Graph for better visualization of all the information multiclass classification on the tweets and to extract the important entities from tweets using NER model

After processing the tweets and extracting all the necessary information, this information is automatically fed to Neo4j graph using python's py2neo library which then can be queried as per our needs to get the information that we need.

Proposed System

The block diagram of the proposed framework is depicted in figure 1. Each block is briefly described next:

Our main contributions are summarized as follows:

(1) Using the keywords such as Flood, Water logging, disaster, calamity etc. scraped the Indian Flood tweets.

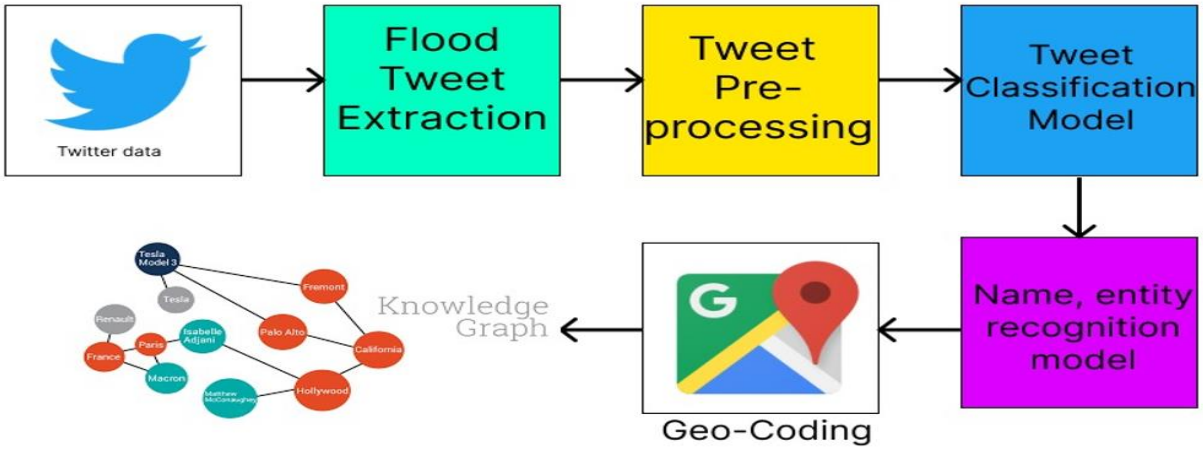
(2) A text preprocessing module to remove noisy textual features.

(3) Processed Tweets are classified into Binary (relevant and non-relevant) and multiclass classification (affected individual, caution and advice, donation and volunteering, Structure and utility damage, sympathy and support and non-humanitarian)

(4) An NER-based geoparsing strategy (toponym extractor) where location of the tweets is extracted.

(5) A Geocoder to query Google Maps API with each toponym, thus presenting results in latitude and longitude values.

(6) A Information Graph is used to graphically depict the overall local and global storyline from processed tweets.



The system consists of the following modules, which have been described in the subsequent sections.

– Data collection
– Data pre-processing
– Event classification (Binary and Multi class)
– Location estimation
– Information Graph

Data collection:

We have used the consolidate eight human-annotated datasets and provide 166.1k and 141.5k tweets for informativeness and humanitarian classification tasks, respectively. These consolidated datasets will help train more sophisticated models . Eight datasets that were labeled for different disaster response classification tasks and whose labels can are informativeness and humanitarian information type classification. Brief overview of dataset used for consolidation are:

CrisisLex is one of the largest publicly available datasets, which consists of two subsets, i.e., CrisisLexT26 and CrisisLexT6

CrisisLexT26 comprises data from 26 different crisis events that took place in 2012 and 2013 with annotations for informative vs. not informative as well as humanitarian categories (six classes) classification tasks among others. CrisisLexT6, on the other hand, contains data from six crisis events that occurred between October 2012 and July 2013 with annotations for related vs. not-related binary classification task.

CrisisNLP is another large-scale dataset collected during 19 different disaster events that happened between 2013 and 2015, and annotated according to different schemes including classes from humanitarian disaster response and some classes related to health emergencies (Imran, Mitra, and Castillo 2016).

SWDM2013, SCRAM2013, Disaster Response Data (DRD), Disasters on social media (DSM), CrisisMMD and AIDR are the additional five more datasets that are used for the consolidation. These datasets size is not as big as first two datasets .

Data pre-processing:

Tweets scraped directly from twitter using the keywords contain different types of noise and redundancies, such as emoticons, user mentions, Internet links etc. An efficient data pre-processing is needed to use these tweets for any meaningful purpose.

Event classification:

Hash tags and keywords in tweets help to extract tweets related to a target event. Consolidated Dataset by compiling 8 different datasets is prepared however, some of these tweets may be referring to general information such as “Floods have become a regular occurrence in Kerala”. The above tweet refers to natural disaster such as floods, which may be a target event, but it does convey real time update of the event all the time. Hence, tweet classification system is required to filter out the type of tweet related to the event. The compiled dataset

is classified into two categories Binary class (informative and non-informative) and multiclass category (6 classes).

Binary and Multiclass Classification:

When tweets are collected from a huge pool of twitter data using keywords of related flood still, we get tweets which are not related to the event hence the binary classification of informative and non-informative is required.

Dataset is also classified into multiple classes as given below:

Affected individual: Message regarding the victims who are affected. E.g., 10 people are struck in water logging near Kalina.

Caution and advice: Warning given about a related incident e.g. Flooded neighbours in Bandra and its approaching near high tide

Donation and volunteering: Tweets which are offering help. E.g., Shelter and food are arranged at shiva Hall near highway

Infrastructure and utility damage: Information regarding the damage caused to the buildings and resources. Building collapsed in Mumbra near station

1. Help and support: Required help to the victims. E.g., @Maharastra Government please send volunteers to the Kalina west some people are stuck.
2. Non-humanitarian: Messages not related to flood e.g. My heart is flooded with lots of love.

Models and Architectures used for classification:

Deep neural networks (DNNs) are ideally suited for classifying a crisis-related tweets. They are usually trained with large pool of data and have the elasticity to learn and modify from new batches of labeled data without requiring retraining from the beginning. Due to their distributed word representation, they generalize well and make better use of the previously labeled data from other events to speed up the learning process. DNNs prevent the need of manually crafting features and automatically learn hidden features as distributed dense vectors, which have shown to advantage various NLP tasks.

Convolutional Neural Network:

This classification model is based on the CNN architecture. We used similar architecture as proposed by (Nguyen et al. 2017). Convolutional Neural Network Figure 2 shows CNN model for classifying tweets into binary class and multi class for a crisis event. The architecture of our model is similar to the one proposed in (Nguyen et al. 2017). For distributed representation of words, we first construct a vocabulary V from the training set by selecting T most frequent words. Each word in the vocabulary is then represented by aD dimensional vector in a shared look-up table L ∈ R |V |×D, which is considered a model parameter to learn. We can initialize L randomly or using pretrained word embedding vectors like word2vec. Given an input tweet s = (w1, ・・・, wT), we first transform it into a feature sequence by mapping each word token wt ∈ s to an index in L. The look-up layer then creates an input vector xt ∈ RD for each token wt, which are passed through a sequence of convolution and pooling operations to learn high-level feature representations.

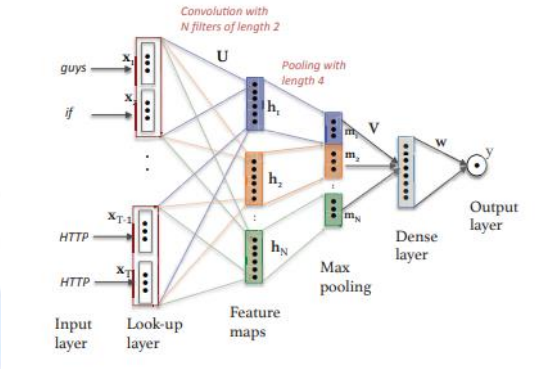


Figure 2: Convolutional neural network on a tweet: "guys if know any medical emergency around balaju area you can reach umesh HTTP doctor at HTTP"

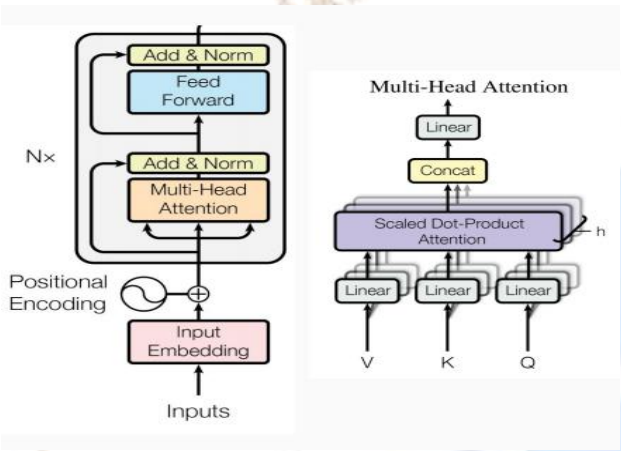
The convolutional layer is the core building block of a CNN, and it is where the majority of computation occurs. It requires a few components, which are input data, a filter, and a feature map.

Pooling layers, also known as down sampling, conducts dimensionality reduction, reducing the number of parameters in the input. Similar to the convolutional layer, the pooling operation sweeps a filter across the entire input, but the difference is that this filter does not have any weights. Instead, the kernel applies an aggregation function to the values within the receptive field, populating the output array.

Fully-Connected Layer performs the task of classification based on the features extracted through the previous layers and their different filters. While convolutional and pooling layers tend to use ReLu functions, FC layers usually leverage a SoftMax activation function to classify inputs appropriately, producing a probability from 0 to 1.

Transformer models: pre-trained model as shown in figure 3 have achieved state-of-the-art performance on natural language processing tasks and have been adopted as feature extractors for solving down-stream tasks such as question answering, and sentiment analysis. Though the pre-trained models are mainly trained on non-Twitter text, we hypothesize that their rich contextualized embeddings would be beneficial for the disaster domain.

BERT is basically an Encoder stack of transformer architecture. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side. BERTBASE has 12 layers in the Encoder stack while BERTLARGE has 24 layers in the Encoder stack. These are more than the Transformer architecture described in the original paper (6 encoder layers). BERT architectures (BASE and LARGE) also have larger feedforward-networks (768 and 1024 hidden units respectively), and more attention heads (12 and 16 respectively) than the Transformer architecture suggested in the original paper. It contains 512 hidden units and 8 attention heads. BERTBASE contains 110M parameters while BERTLARGE has 340M parameters.



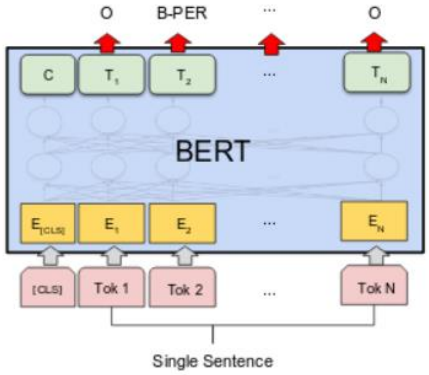
Pretrained Transformer Encoder

Location estimation:

In this work, we are particularly interested in the locations described in the content of tweets.

While both the news and literature told us that people used Twitter and other social media plat-forms to request for help and share information, we still do not know how specifically people describe locations in social media messages during this natural disaster. Manually analyzing the more than 100k tweets is tweets is practically impossible. We randomly select 1,00 tweets for testing.

BERT NER is a fine-tuned BERT model as shown in figure 4 that is ready to use for Name Entity Recognition and achieves good for the NER task. It has been trained to recognize four types of entities location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC).



NER BERT MODEL

When several tweets are collected, are fed into the classification model to obtain a series of predicted entities ˆy. Prior to the conversion of predictions into useful toponyms, words classified with O and PER tags are discarded (their presence is required in the training stage to capture entity tag transitions at the CRF output layer, but they are not needed for toponym identification). Furthermore, those classified as LOC and ORG are identified and joined to form a sentence, consequently creating a toponym, which is used to request a Google API location. Responses from Google are geocoded in JSON format to form an address with geographic coordinates.

Information Graph:

The knowledge base consists of a dynamic Information Graph [37]. This allows us to structure entities as a Information Graph in which new nodes and relationships are constantly being added as new feed is detected, to expand and upgrade its knowledge in real-time. This knowledge is disaster-centered, as disaster occurrences are represented by specific nodes which then connect to locations, dates, people, organizations, etc. In addition to having several node types, such as the date and locations, each node also has several attributes. This allows us for expanding in real time when dealing with various types of data associated with entities. As for its management, our knowledge base uses the Neo4j (https://neo4j.com/) graph database management system .

Lastly, the classified tweets must be included in our Information Graph. This step involves all validations regarding the previous existence of our gathered data, as well as the creation of new nodes and relationships between them, to express the identity of

the extracted disasters in a way that facilitate human consumption and future system integrations. To further deal with social media unreliability, all disaster nodes have an attribute which is incremented when the same disaster is

detected from different tweets. This attribute can then be used as a detection threshold, to allow for better filtering of relevant tweets. If new information regarding an already existing disaster is extracted from the previous steps, it is also validated and then connected with that event. This mechanic allows for our system to be continually learning and updating, expanding its knowledge, and bettering the knowledge it already has.

Results and Conclusion

Datasets

Python's snscrape module is being used to create datasets by extracting tweets from twitter. Each dataset contains anywhere around 5000-30000 tweets related to the particular event that had occurred in the past. These datasets will be used to train the deep learning model. The datasets contain the following columns- Username, Location, Tweet, Date, Followers count, Co-ordinates, Friends count, Protected status, verified status, Reply count, Like count, Retweet count

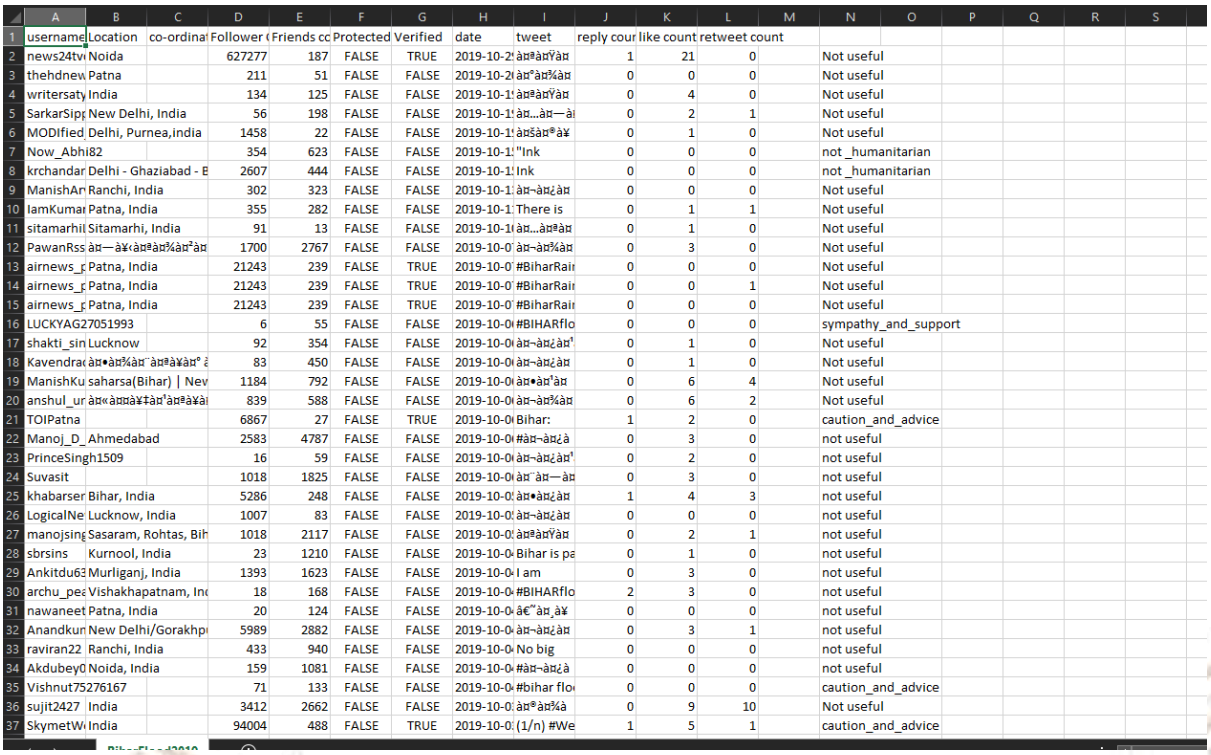
This is the code used for scraping the tweets-

```
import snscrape.modules.twitter as sntwitter
import csv
maxTweets = 30000

loc = ' 19.663280, 75.300293, 500km'
csvFile = open('Flood.csv', 'a', newline='', encoding='utf8')

#Use csv writer

csvWriter = csv.writer(csvFile)
csvWriter.writerow(['username','Location','co-
ordinates','Follower Count','Friends
count','Protected','Verified','date','tweet','reply count','like
count','retweet count',''])

for i,tweet in
enumerate(sntwitter.TwitterSearchScraper('(location name)
+ since:2016-06-01 until:2016-08-01 -filter:links -
filter:replies').get_items()):
        if i > maxTweets :
            break

csvWriter.writerow([tweet.user.username,tweet.user.location
,tweet.coordinates,tweet.user.followersCount,tweet.user.frien
dsCount,tweet.user.protected,tweet.user.verified, tweet.date,
tweet.content,tweet.replyCount,tweet.likeCount,tweet.retwee
tCount])
csvFile.close()
```

The following datasets were extracted using snscrape module-

- Assam Flood 2017
- Bihar Flood 2019
- Gujrat Flood 2017
- Maharashtra Flood 2021
- Odisha Flood 2019
- Srinagar Flood 2014
- Chennai Flood 2015

Sample Eg : *Assam Flood 2017*



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	username	Location	co-ordina	Follower (	Friends co	Protected	Verified	date	tweet	reply cour	like count	retweet count							
2	news24tv	Noida		627277	187	FALSE	TRUE	2019-10-2	Ã¤ªÃ¤Ÿà¤	1	21	0		Not useful					
3	thehdnew	Patna		211	51	FALSE	FALSE	2019-10-2	Ã¤ª Ã¤¾Ã¤	0	0	0		Not useful					
4	writersaty	India		134	125	FALSE	FALSE	2019-10-1	Ã¤ªÃ¤Ÿà¤	0	4	0		Not useful					
5	SarkarSipp	New Delhi, India		56	198	FALSE	FALSE	2019-10-1	Ã¤…à¤—à¤	0	2	1		Not useful					
6	MODIfied	Delhi, Purnea,india		1458	22	FALSE	FALSE	2019-10-1	Ã¤šÃ¤°à¥	0	1	0		Not useful					
7	Now_Abhi	82		354	623	FALSE	FALSE	2019-10-1	"Ink	0	0	0		not _humanitarian					
8	krchandar	Delhi - Ghaziabad - E		2607	444	FALSE	FALSE	2019-10-1	Ink	0	0	0		not _humanitarian					
9	ManishAn	Ranchi, India		302	323	FALSE	FALSE	2019-10-1	Ã¤¬Ã¤¿Ã¤	0	0	0		Not useful					
10	IamKumar	Patna, India		355	282	FALSE	FALSE	2019-10-1	There is	0	1	1		Not useful					
11	sitamarhil	Sitamarhi, India		91	13	FALSE	FALSE	2019-10-1	Ã¤…Ã¤ªÃ¤	0	1	0		Not useful					
12	PawanRss	Ã¤—Ã¥‹Ã¤ªÃ¤¾Ã¤²Ã¤		1700	2767	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¾Ã¤	0	3	0		Not useful					
13	airnews_p	Patna, India		21243	239	FALSE	TRUE	2019-10-0	#BiharRair	0	0	0		Not useful					
14	airnews_p	Patna, India		21243	239	FALSE	TRUE	2019-10-0	#BiharRair	0	0	1		Not useful					
15	airnews_p	Patna, India		21243	239	FALSE	TRUE	2019-10-0	#BiharRair	0	0	0		Not useful					
16	LUCKYAG	27051993		6	55	FALSE	FALSE	2019-10-0	#BIHARflo	0	0	0		sympathy_and_support					
17	shakti_sin	Lucknow		92	354	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¿Ã¤°	0	1	0		Not useful					
18	Kavendra	Ã¤•Ã¤¾Ã¤¨Ã¤ªÃ¥Ã¤° Ã		83	450	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¿Ã¤	0	1	0		Not useful					
19	ManishKu	saharsa(Bihar) | Nev		1184	792	FALSE	FALSE	2019-10-0	Ã¤•Ã¤°Ã¤	0	6	4		Not useful					
20	anshul_ur	Ã¤¤Ã¤¤Ã¥‡Ã¤°Ã¤¥Ã¤		839	588	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¾Ã¤	0	6	2		Not useful					
21	TOIPatna			6867	27	FALSE	TRUE	2019-10-0	Bihar:	1	2	0		caution_and_advice					
22	Manoj_D_	Ahmedabad		2583	4787	FALSE	FALSE	2019-10-0	#Ã¤¬Ã¤¿Ã	0	3	0		not useful					
23	PrinceSingh	1509		16	59	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¿Ã¤°	0	2	0		not useful					
24	Suvasit			1018	1825	FALSE	FALSE	2019-10-0	Ã¤¯Ã¤—Ã¤	0	3	0		not useful					
25	khabarser	Bihar, India		5286	248	FALSE	FALSE	2019-10-0	Ã¤•Ã¤¿Ã¤	1	4	3		not useful					
26	LogicalNe	Lucknow, India		1007	83	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¿Ã¤	0	0	0		not useful					
27	manojsing	Sasaram, Rohtas, Bih		1018	2117	FALSE	FALSE	2019-10-0	Ã¤ªÃ¤Ÿà¤	0	2	1		not useful					
28	sbrsins	Kurnool, India		23	1210	FALSE	FALSE	2019-10-0	Bihar is pa	0	1	0		not useful					
29	Ankitdu63	Murliganj, India		1393	1623	FALSE	FALSE	2019-10-0	I am	0	3	0		not useful					
30	archu_pea	Vishakhapatnam, Inc		18	168	FALSE	FALSE	2019-10-0	#BIHARflo	2	3	0		not useful					
31	nawaneet	Patna, India		20	124	FALSE	FALSE	2019-10-0	â€™Ã¤¸Ã¥	0	0	0		not useful					
32	Anandkun	New Delhi/Gorakhp		5989	2882	FALSE	FALSE	2019-10-0	Ã¤¬Ã¤¿Ã¤	0	3	1		not useful					
33	raviran22	Ranchi, India		433	940	FALSE	FALSE	2019-10-0	No big	0	0	0		not useful					
34	Akdubey0	Noida, India		159	1081	FALSE	FALSE	2019-10-0	#Ã¤¬Ã¤¿Ã	0	0	0		not useful					
35	Vishnut75	276167		71	133	FALSE	FALSE	2019-10-0	#bihar flo	0	0	0		caution_and_advice					
36	sujit2427	India		3412	2662	FALSE	FALSE	2019-10-0	Ã¤°Ã¤¾Ã	0	9	10		Not useful					
37	SkymetW	India		94004	488	FALSE	TRUE	2019-10-0	(1/n) #We	1	5	1		caution_and_advice					

In order to train and validate our model, sufficient tweets related to an event are needed, which should reflect the realistic scenario of that event. We used Twitter API to capture live tweets related to floods in southern and eastern states of India. The data collection was done using snscrape's twitter python library. A total of 32,400 tweets were collected with keywords "flood", "water". The collected tweets were in English, Hindi, and some other regional languages. For this study, we concentrated only on tweets in English and Hindi languages. One of the major problems with data collected from Twitter is that it may contain a lot of irrelevant tweets such as advertisements. There are many spammers, also known as 'spambots', sending huge number of tweets. Finding spammers is a very difficult task and a number of researchers (Benevenuto et al. 2010; Gayo-Avello 2013; Li and Du 2014; Yardi et al. 2010) are focusing on fixing this issue. In our case, spamming does not pose a significant problem, as we were collecting tweets originating from mobile phones only. The rationale behind this is that hand-held devices are used as personal devices, and they are hardly used for mass tweet dissemination. To filter out the tweets coming from hand-held devices, the source field of the tweets is used. To further reduce the effect of spambots, only tweets from users having a ratio of the number of followers to the number of those following less than one was stored.

Dataset Labelling

Event Name	Affected Ir	Caution &ad	donation & volunteerin	infrastructure &u	not Hummanita	sympathy & Support
Assam Flood	61	52	65	48	35	39
Bihar Flood	16	14	15	7	5	23
Gujarat Flood	62	37	41	34	12	77
Maharahtra Flood	32	10	34	1	4	19
Srinagar Flood	76	29	34	30	9	45
Odisha Flood	59	72	15	31	19	70
Chennai Flood	1219	1475	1703	1005	913	1385

Models

	A	B	C	D	E	F
1	Event Name	Event type	Total Tweets	Date Range	Duplicate Tweets	Other Language
2	Assam Flood	Floods	3300	30-09-2019-2017-04-01	2900	100
3	Bihar Flood	Floods	725	2019-10-29 -2019-09-28	637	200
4	Gujarat Flood	Floods	2713	14-08-2017-2017-06-03	2300	150
5	Maharahtra Flood	Floods	832	2021-07-23-2021-07-22	354	300
6	Srinagar Flood	Floods	1172	29-09-2014-2014-09-03	740	200
7	Odisha Flood	Floods	1795	2020-11-29-2019-09-04	1250	279
8	Chennai Flood	Floods	30002	2016-03-24-2015-12-05	17500	4800

Train	Test	Accuracy(%)	P(%)	R(%)	F1(%)
CrisisLex(6C)	CrisisLex(6C)	77	83	83	78
CrisisNlp(10C)	CrisisNlp(10C)	67	72	72	66
Consolidated(11C)	Consolidated(11C)	77	75	77	76

Results of CNN model on Humanitarian datasets

Train	Test	Accuracy(%)	P(%)	R(%)	F1(%)
CrisisLex(2C)	1- CrisisLex	94	94	95	94
	2- CrisisNlp	69	69	68	69
	3- Consolidated	81	80	80	81
CrisisNlp(2C)	1- CrisisNlp	82	82	82	82
	2- CrisisLex	71	80	71	70
	3- Consolidated	72	76	73	72
Consolidated(2C)	1- Consolidated	87	88	90	89
	2- CrisisLex	82	83	84	82
	3- CrisisNlp	83	84	84	83

Results of CNN model on Informativeness datasets

Train	Test	Accuracy	P(%)	R(%)	F1(%)
CrisisLex(6C)	CrisisLex(6C)	93.156	93.251	93.156	92.798
CrisisNLP(10C)	CrisisNLP(10C)	88.071	87.071	88.071	85.977
Consolidated(11C)	Consolidated(11C)	94.116	94.315	94.116	93.487

Results of BERT model on humanitarian datasets

Train	Test	Accuracy	P(%)	R(%)	F1(%)
CrisisLex(2C)	1-CrisisLex	98.5375	98.559	98.537	98.534
	2-CrisisNLP	71.329	73.905	71.329	69.189
	3-Consolidated	82.177	82.160	82.177	81.926
CrisisNLP(2C)	1-CrisisLex	78.575	83.703	78.575	78.502
	2-CrisisNLP	95.218	95.221	95.218	95.219
	3-Consolidated	77.546	79.289	77.546	77.767
Consolidated(2C)	1-CrisisLex	85.325	85.125	85.325	84.930
	2-CrisisNLP	86.624	86.612	86.624	86.513
	3-Consolidated	96.389	96.398	96.389	96.380

Results of BERT model on Informativeness datasets

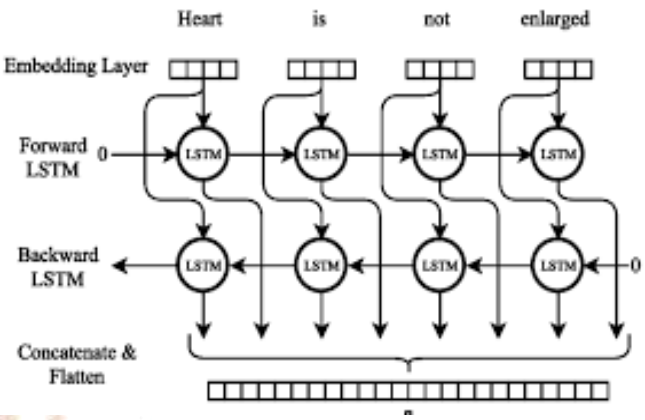
Comparison between CNN and BERT

As BERT is bidirectional encoders representation from transformer, the accuracy which we got is higher than CNN model. We ran both model on same training and testing dataset and acquired greater accuracy on BERT. As shown in the above tables, the precision, recall and F1 scores of BERT model is higher than the CNN model. So, it can be said that BERT model is more effective than CNN model on tweets.

Named Entity Recognition Results

For named entity recognition(ner) we are using BiLstm (bidirectional LSTM) A Bidirectional LSTM, or biLSTM, is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. We have trained our model on a tweeter dataset having tagged entities. The accuracy which we are getting is around 94 % for all the entities.

Bi LSTM



Result of Bi LSTM model on train Datase

```
---------- Train set quality: ----------
processed 105778 tokens with 4489 phrases; found: 4544 phrases; correct: 4396.

precision:  96.74%; recall:  97.93%; F1:  97.33

             company: precision:   97.24%; recall:   98.60%; F1:   97.92; predicted:   652
            facility: precision:   91.89%; recall:   97.45%; F1:   94.59; predicted:   333
             geo-loc: precision:   97.81%; recall:   98.80%; F1:   98.30; predicted:  1006
               movie: precision:   94.12%; recall:   94.12%; F1:   94.12; predicted:    68
        musicartist: precision:   96.55%; recall:   96.55%; F1:   96.55; predicted:   232
               other: precision:   96.33%; recall:   97.09%; F1:   96.71; predicted:   763
              person: precision:   98.54%; recall:   98.87%; F1:   98.70; predicted:   889
             product: precision:   97.52%; recall:   98.74%; F1:   98.12; predicted:   322
          sportsteam: precision:   96.35%; recall:   97.24%; F1:   96.79; predicted:   219
              tvshow: precision:   80.00%; recall:   82.76%; F1:   81.36; predicted:    60
```

Result of Bi LSTM model on test dataset

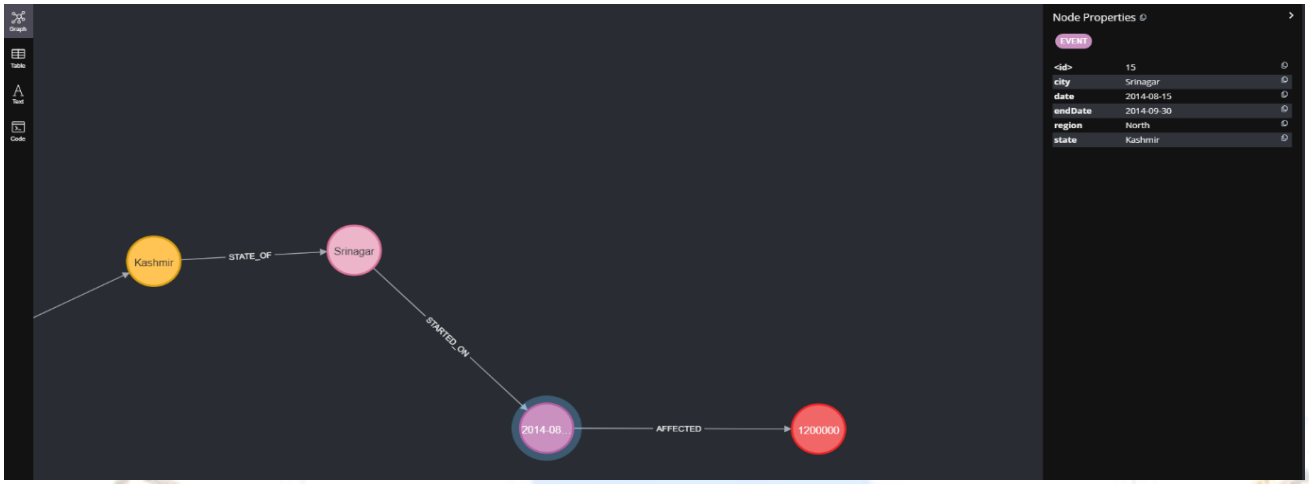
```
---------- test set quality: ----------
processed 118614 tokens with 5026 phrases; found: 4979 phrases; correct: 4591.

precision:  92.21%; recall:  91.35%; F1:  91.77

             company: precision:   91.38%; recall:   92.24%; F1:   91.81; predicted:   754
            facility: precision:   88.61%; recall:   91.67%; F1:   90.11; predicted:   360
             geo-loc: precision:   95.07%; recall:   93.87%; F1:   94.46; predicted:  1095
               movie: precision:   83.12%; recall:   85.33%; F1:   84.21; predicted:    77
        musicartist: precision:   93.85%; recall:   88.08%; F1:   90.87; predicted:   244
               other: precision:   90.58%; recall:   90.69%; F1:   90.64; predicted:   839
              person: precision:   95.57%; recall:   90.88%; F1:   93.17; predicted:   949
             product: precision:   90.86%; recall:   90.34%; F1:   90.60; predicted:   350
          sportsteam: precision:   91.14%; recall:   91.14%; F1:   91.14; predicted:   237
              tvshow: precision:   64.86%; recall:   77.42%; F1:   70.59; predicted:    74
```

Graph Database:

Information Graph are a way of structuring information in graph form by representing entities as nodes and relationship between entities as edges.

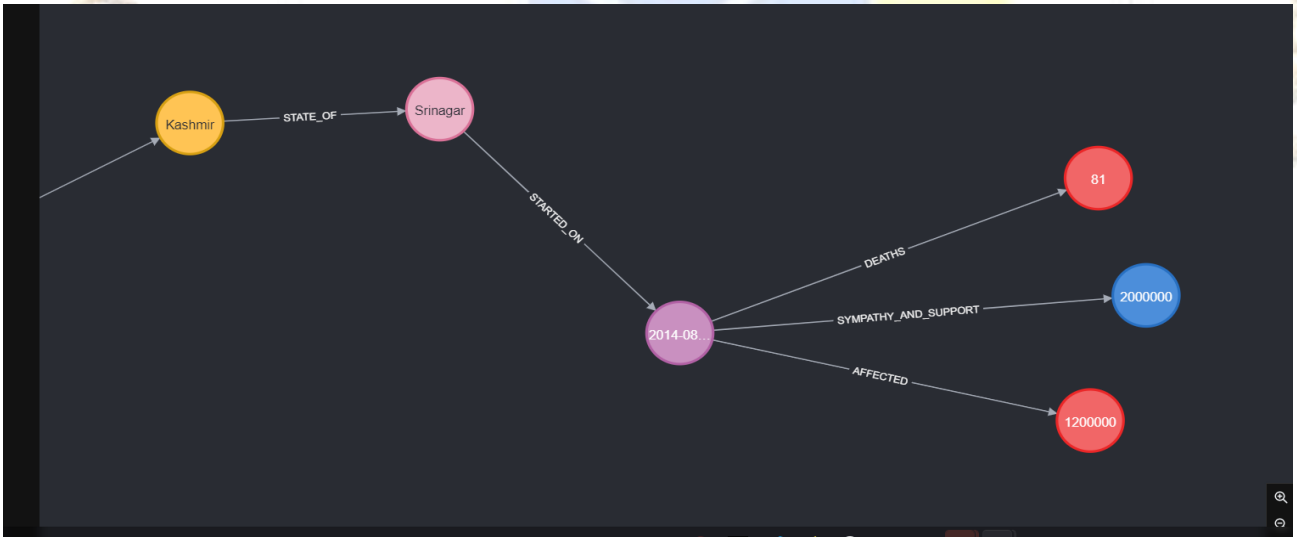
Following is an example of how the database increases in size as number of tweets increases-

Graph | csv file after 15 Tweets



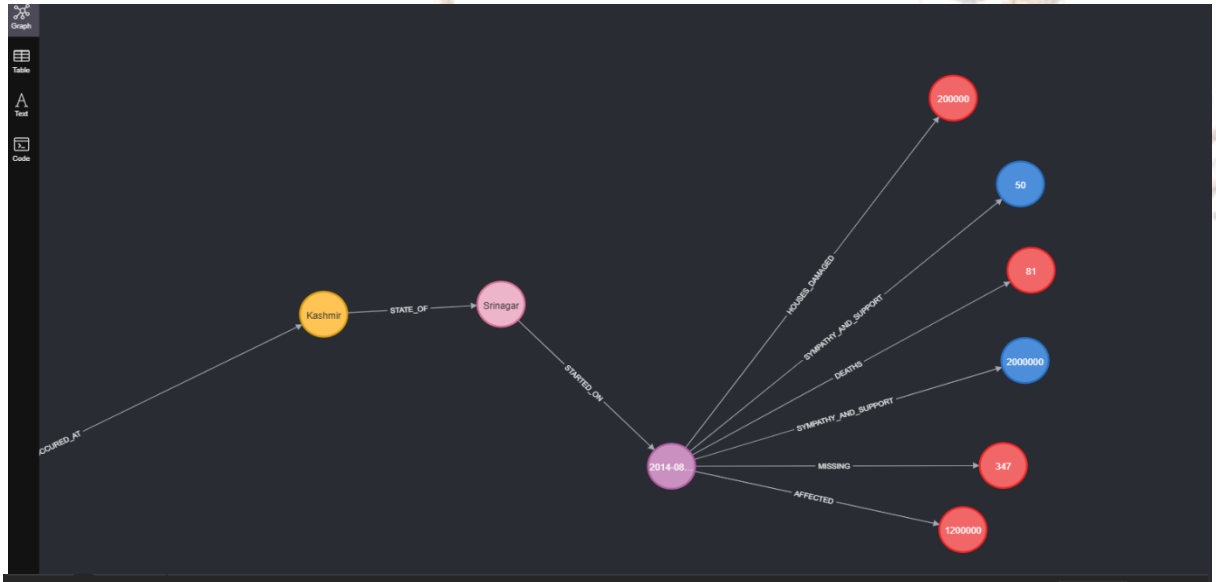
	A	B	C	D	E	F
1	Disaster	State	City	Started on	Affected	
2	Flood	Kashmir	Srinagar	15-08-2014	1200000	
3						
4						
5						
6						
7						

Graph | csv file after 50 tweets



Graph | csv file after 100 tweets

Disaster	State	City	Started on	Affected	Deaths	Donation
Flood	Kashmir	Srinagar	15-08-2014	1200000	81	AAP MPs MLAs will contribute Rs 20 lakh each for the Kashmir flood relief from their development fund



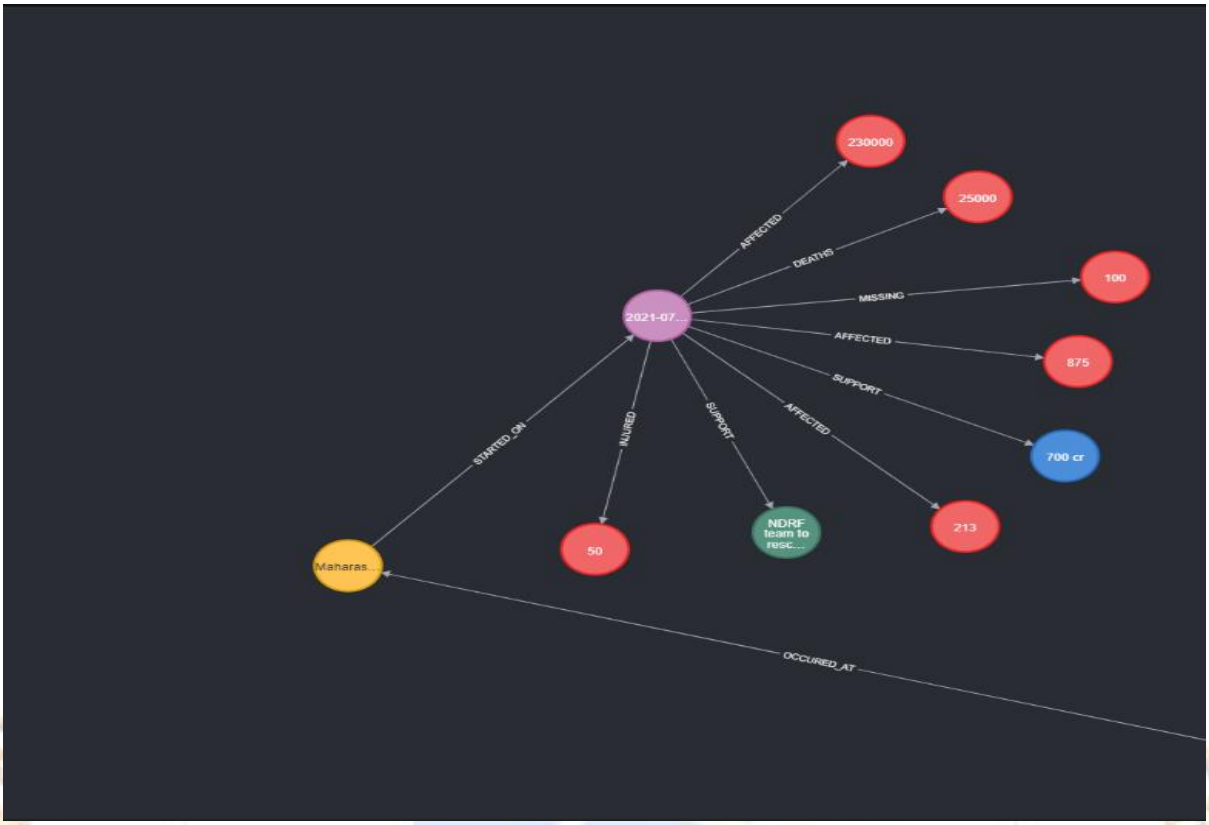
Disaster	State	Started on	Affected	Missing	Deaths	Houses Da	Donation	Support	Villages aff	Injured	Animal dea	Donation
Flood	Maharasht	2021-07-15	230000	100	213	200000	Centre ap	34 NDRF t	81	50	25000	
Flood	Gujarat	01-06-2017	33000		224		101 cr don	12 Relief camps				2500 food packets donated by various organization

The final graph that is formed can be queried in many ways. Some of the queries are given below-

Graph which shows the floods that occurred in western region of India between 2016 and 2022.

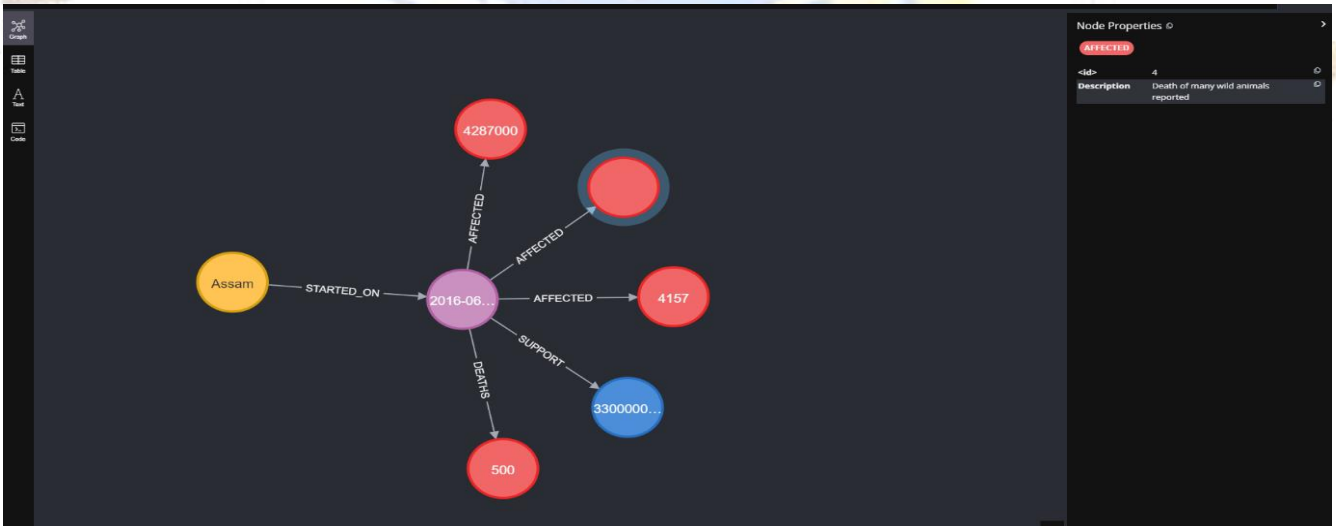
Disaster	State	Started on	Affected	Missing	Deaths	Houses Da	Donation	Support	Villages aff	Injured	Animal dea	Donation
Flood	Maharasht	2021-07-15	230000	100	213	200000	Centre ap	34 NDRF t	81	50	25000	
Flood	Gujarat	01-06-2017	33000		224		101 cr don	12 Relief camps				2500 food packets donated by various organization

CSV of Graph which shows the floods that occurred in western region of India/Local Storyline between 2016 and 2022.

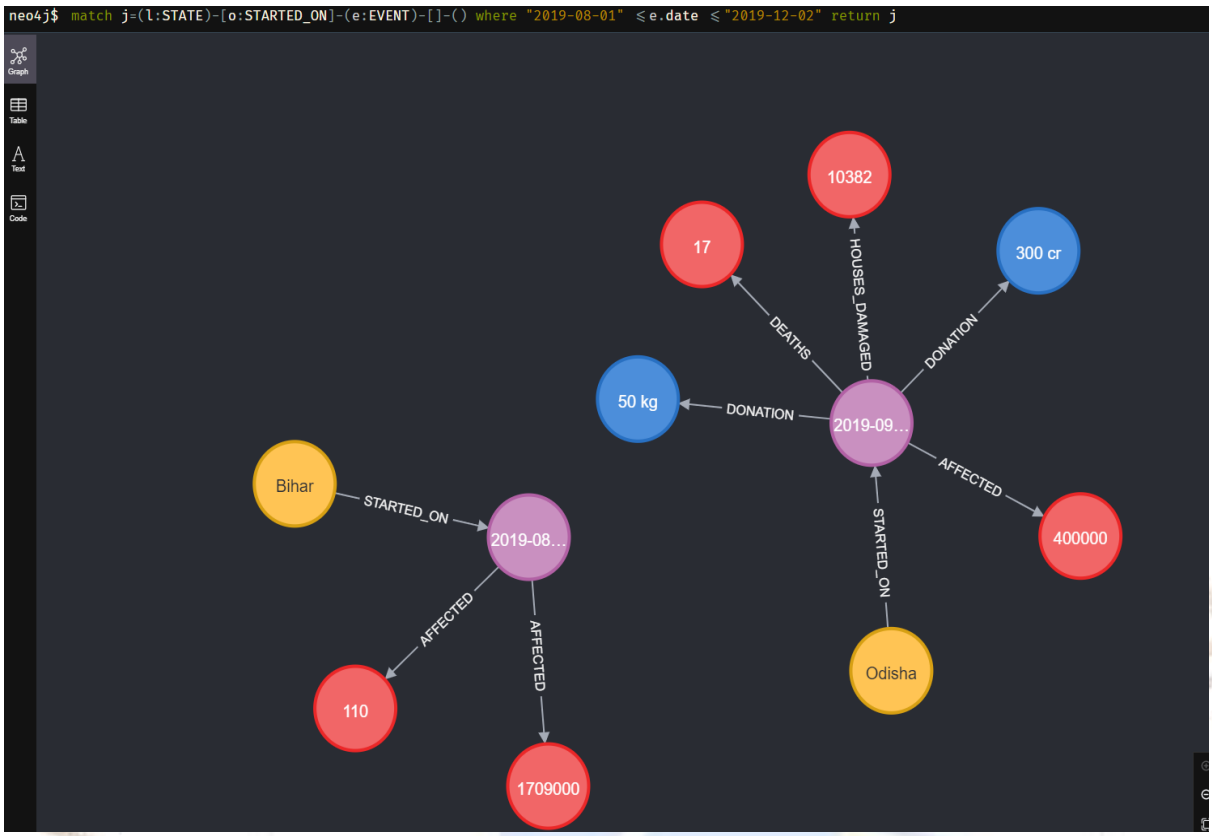


Graph that shows the flood that happened on a particular date (2017/03/16).

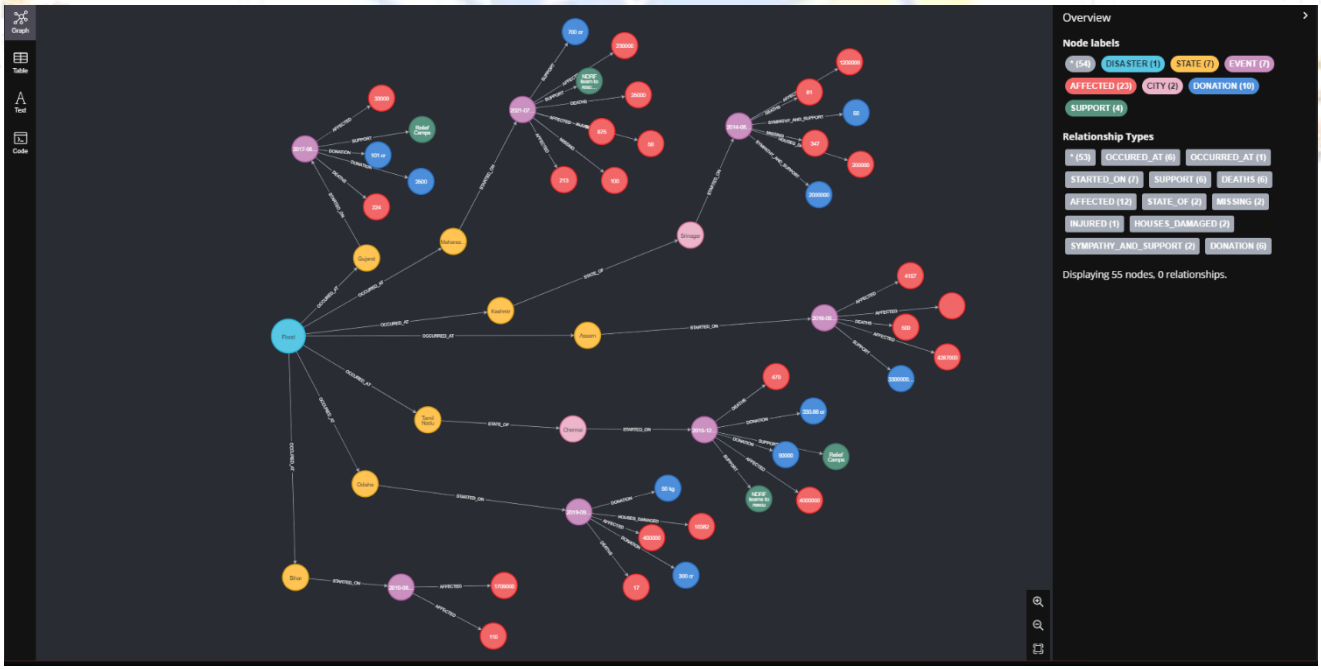
Local Storyline/Maharashtra



Graph which shows the floods that occurred in between August and December of 2019



Global Storyline/India



Conclusion

Disaster related casualties have an extensive impact in our society. Disaster management focuses on preventing and minimizing the risks these scenarios face humanity with. Social media, although raw, presents a way of using humans as sensors to detect such hazards with the utmost brevity. Our model builds upon existing tools, contributing with a dynamic way of extracting and representing spatial and temporal relationships, as well as providing this knowledge to decision-makers, in real-time. As we've already completed the model design for informative and non- informativeness, humanitarian classes using BERT and CNN along with that we are done with NER Bi LSTM model. So, this project can be further enhanced by including other disasters as well so that more people will benefit from this program and can be made accessible to more people by creating a web application where we can display this information present in Neo4j about all the disasters to several organizations like NGOs and government which will help them in providing support to potentially lakhs of people.

References

1. Dhanya, V.G., Jacob, M.S., Dhanalakshmi, R. (2021). Twitter-Based Disaster Management System Using Data Mining. In: Pandian, A., Fernando, X., Islam, S.M.S. (eds) Computer Networks, Big Data and IoT. Lecture Notes on Data Engineering and Communications Technologies, vol 66. Springer, Singapore.
2. Chanda, Ashis Kumar. "Efficacy of BERT embeddings on predicting disaster from Twitter data." arXiv preprint arXiv:2108.10698 (2021).
3. Singh, Jyoti Prakash, et al. "Event classification and location prediction from tweets during disasters." Annals of Operations Research 283.1 (2019): 737-757.
4. Boné, João, et al. "DisKnow: a social-driven disaster support knowledge extraction system." Applied Sciences 10.17 (2020): 6083.
5. Hu, Yingjie, and Jimin Wang. "How do people describe locations during a natural disaster: an analysis of tweets from Hurricane Harvey." arXiv preprint arXiv:2009.12914 (2020).
6. Lee K, Ganti R, Srivatsa M, & Mohapatra P (2013) Spatio-temporal provenance: identifying location information from unstructured text. In 2013 I.E. International Conference on Pervasive Computing and Communications Workshops (PERCOM workshops) (pp. 499– 504). IEEE. https://doi.org/10.1109/PerComW.2013.6529548.
7. Kumar, Abhinav, and Jyoti Prakash Singh. "Location reference identification from tweets during emergencies: A deep learning approach." International journal of disaster risk reduction 33 (2019): 365-375.
8. Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 237–246). ACM.
9. Kumar, A., Singh, J. P., & Rana, N. P. (2017). Authenticity of geo-location and place name in tweets. In 23rd Americas Conference on Information Systems (AMCIS).
10. Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate?: a gis approach for user location inference. In Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (pp. 1–8). ACM.
11. Nakaji, Y., & Yanai, K. (2012). Visualization of real-world events with geotagged tweet photos. In Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on (pp. 272–277). IEEE.
12. Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M. (2013). Who, where, when and what: discover spatio-temporal topics for twitter users. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 605–613). ACM.
13. Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 759–768). ACM.
14. Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In ICWSM.

15. Itoh, M., Yoshinaga, N., & Toyoda, M. (2016). Spatio-temporal event visualization from a geo-parsed microblog stream. In Companion Publication of the 21st International Conference on Intelligent User Interfaces (pp. 58–61). ACM.
16. Malmasi, S., & Dras, M. (2015). Location mention detection in tweets and microblogs. In International Conference of the Pacific Association for Computational Linguistics (pp. 123–134). Springer.
17. Middleton, S. E.,Middleton, L., &Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. IEEE Intelligent Systems, 29, 9–17.
18. Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. Journal of Spatial Information Science, 2014, 37–70.
19. Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. Transactions in GIS, 15, 753–773.
20. Giridhar, P., Abdelzaher, T., George, J., & Kaplan, L. (2015). On quality of event localization from social network feeds. In Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on (pp. 75–80). IEEE.
21. Unankard, S., Li, X., & Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. World Wide Web, 18, 1393–1417.
22. Xu., & Li, S. Knowledge graph visualization of intelligent emergency research. International Journal of Innovative Computing, Information and Control ICIC International ISSN 1349-4198 Volume 17, Number 4 (2021).
23. Son, J., Lim, C. S., Shim, H. S., & Kang, J. S. Development of Knowledge Graph for Data Management Related to Flooding Disasters Using Open Data. Future Internet, 13(5), 124 (2021).