Understanding Flood Situation in India Using Social Media

Abdul Rehman[a] , Aishwarya Ahuja[b], Alok Bhardwaj[c]

[a]Department of Computer Science Engineering, Sant Longowal Institute of Engineering and Technology, Punjab
[b]Department of Biosciences and Bioengineering, IIT Roorkee
[c]Department of Civil Engineering, IIT Roorkee

Introduction and Motivation

- Social media platforms such as Twitter can be used for information distribution, community engagement, and real-time monitoring of natural hazards.
- Among natural hazards, floods cause huge losses to life and property across the globe[1]. In India, floods occur every year during the monsoon season affecting different parts of the country[2].
- Our motivation in this work is to use Twitter to accurately determine flooded v/s non-flooded tweets and subsequently extract information on flooded locations across India.

1) https://link.springer.com/content/pdf/10.1007/s11069-004-8891-3.pdf?pdf=inline%20link
2) Ray, Kamaljit, et al. “On the Recent Floods in India.” *Current Science*, vol. 117, no. 2, 2019, pp. 204–18. *JSTOR*, https://www.jstor.org/stable/27138236. Accessed 5 July 2023.

Study Area and Dataset

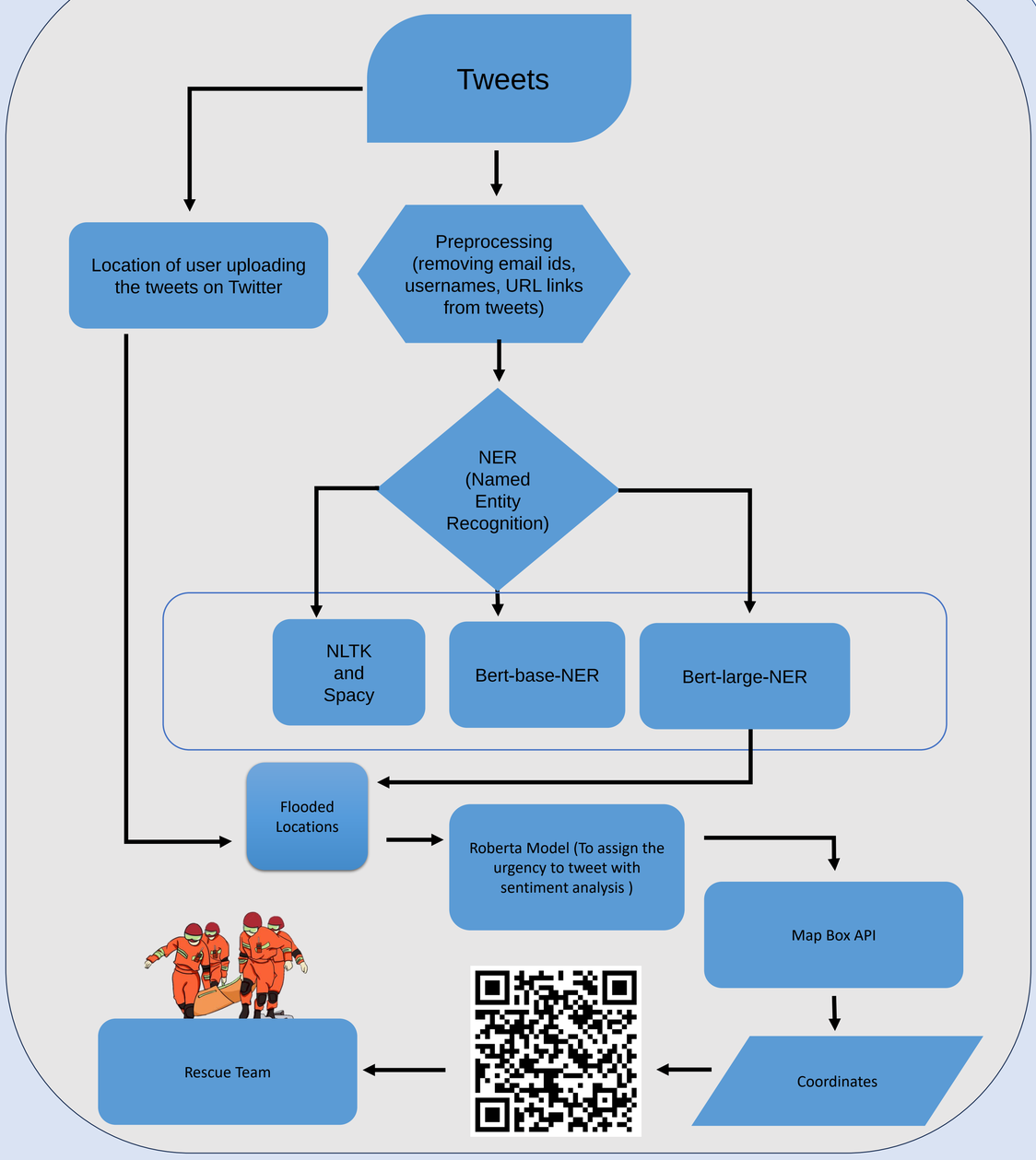
- We worked with a twitter dataset extracted using snscraper (snscraper is an open-source scraper for social networking services (SNS). The extracted dataset contains user location , hashtags (e.g. #flood #floodinAssam), searches (live tweets, top tweets, and users), tweets (single or surrounding thread), list posts and trends.
- For example, we have used the following search in snscraper: 'flood *"flood" (flood OR Monsoon) -Pakistan -Bangladesh -California(#Flood) min_replies:2 min_faves:50 min_retweets:5 lang:en since:2022-06-01 to 2023-06-01’*
- Using the above search, we obtained a dataset of 2334 tweets extracted
- We then applied certain techniques to the extracted tweets to obtain details of flood. These techniques are described next.

Techniques

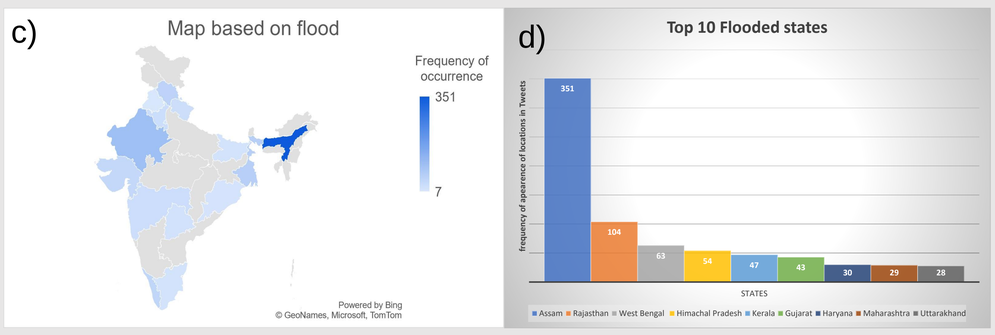
- We used Natural Language Toolkit (NLTK) and spacy, and Transformers (BERT-base-NER and BERT-large-NER) for Named Entity Recognition.
- Named Entity Recognition (NER) extracts information from text. NER involves detecting and categorizing important information in text
- We used Bert-large-NER[3] for extracting location from extracted tweets using snscraper since Bert-large-NER is a fine-tuned BERT model for NER and achieves state-of-the-art performance for any given NER task.
- The Bert-large-NER model is trained to recognize four types of entities: location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC).
- RoBERTa [4]: A Robustly Optimized BERT Pretraining Approach which is built on BERT by modifying key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

3) https://arxiv.org/pdf/1810.04805.pdf
4) https://arxiv.org/abs/1907.11692

Methodology



Results



c) We have successfully extracted locations from given tweets using Bert-large-NER model.
d) We have successfully identified Top10 flood affected states in India during the period spanning from June 1, 2022 to June 1,2023 based on frequency of flood occurrences in each state.

Conclusion and Future Work

- We have observed that Bert-Large-NER model outperforms the Bert-base-NER and NLTK and Spacy library based model.
- In future we will improve the Roberta Model based classifier to classify tweets into “Flooded” and “Non Flooded”.

Acknowledgement

The author would like to thank Twitter social networking site for providing tweets datasets. Also, the author is very much grateful to the SPARK 2023 committee of the Indian Institute of Technology Roorkee for providing the necessary infrastructure to carry out this research work. The author would also like to thank Prof. Alok Bhardwaj for guiding and supervising during the research internship