Extracting, exploring, summarizing information from Twitter social media, aggregated news during disasters

Mahesh Kumar Badam Venkata
IST700 – Deep Learning, NLP and
Computational Sciences – Project Report
School of Information Studies,
Syracuse University
mbadamve@syr.edu

Abstract

In this research report, the goal is to understand and extract information from tweets posted on Twitter social media during disasters and achieve providing knowledge extracted to the organizations, charities and in turn, help them take decisions about disaster recovery. Also, the information extraction process can be automated and will be discussed in the future scope of this research work. On a higher level, this is made possible by the use of powerful Natural Language Processing (NLP) algorithms like BERT and using word embeddings from Word2Vec technique.

1 Credits

This document has been developed with the references of prior researches in this field. The initial research work that shows the baseline performances of various models has been adapted from the paper titled *Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters* which contains much information about extraction of features used for model building, information retrieval via summarization techniques. Next, it is important to mention *word2vec* word embedding techniques developed by Google, that is applauded all over the world and BERT, language model again by Google has transformed the NLP tasks by achieving the state-of-the-art performance.

2 Introduction

Twitter has become a vital source of information, used by millions of users to share a variety of information and it plays a key role in obtaining situational information during *disaster events* and this is shown in many recent studies. Not only the news of the disaster, but also the sentiment of persons and status or updates of operations done are posted by users and it will be very helpful if organizations knew these updates via an automated system that extracts knowledge and presents with actions needed. Time is very critical in a disaster situation, if these tasks can be performed in a near real time, the authorities can develop actions quickly to recover the situation.

Recent studies attempted to overcome few challenged in summarizing information but there are few limitations, like unable to distinguish between situational and non-situational tweets and extending to less resourceful languages. Modifications to the methodologies need to be made to improve the framework of extracting this knowledge. Through this research work, I would like to analyze tweets various methodologies and provide potential improvements to this automated framework. I put forward performance of various models and compare them with baseline models' performance and later provide areas of improvement. In a nutshell, state-of-the-art models become key to identifying tweets and categorizing them according to the type of information in the tweet.

3 Significance

During disasters like floods, earthquakes and hurricanes or typhoons, massive spread of dangerous diseases in times of pandemics, and bomb blasts, sharing information is vital and through social media channels like Twitter, it has become easier to share news at almost any corner of the world. While it is not possible for government officials or social workers, charity organizations to go through millions of tweets and understand the vital information manually, the same work can be performed using powerful NLP algorithms and techniques with the knowledge from prior dataset of tweets already made ready for analysis. Through this research, aiming for automating the information retrieval process, the process of disaster recovery can be eventually made somewhat faster than existing methods of disaster management.

4 Prior Research

There were many attempts of extracting knowledge and summarizing information from social media in many fields of interest like during product launches or political campaigns etc. However, there were few drawbacks and limitations to applications to specific areas.

Through this research analysis, there would be a novel framework that goes beyond and attempting to mitigate the problems faced by other researchers. A few research attempts by few researchers include classification of tweets during disaster events, which use bag-of-words models to distinguish tweets which contain situation information, and which do not. One drawback for this was the classifier was trained on vocabulary of past disasters and couldn't generalize to a new disaster. This work would be focusing on ideas and solutions which could be used to solve such problems. Not just this one, I would be deep-diving into other problems and how they were or can be potentially taken care of.

Another important prior research was tweet summarization, focusing on summarizing a set of tweets, e.g., tweets coming during sports events. This was done using clustering techniques, graph-based abstract summarization, greedy summarization etc., however, as said before, few problems cause to fail these processes and the research analysis aims to converge the ideas and implementations by proper analysis and interpretation.

5 Research methods used and the process of analysis

The datasets have been collected from the *CrisisNLP* website which contains many datasets of user tweets during different disasters. Each dataset has tweet id, tweet text and the label showing the type of information in the tweet. The dataset was developed by *CrisisNLP*, annotated by paid workers with high level of annotation rules and restricting number of annotation tasks per user. Hence, it is assumed that the dataset will have higher degree of annotation efficiency. For this research purposes, I use 11 datasets which include tweet text and label for the tweets. 11 datasets have 11 different type of disasters and situational tweets during the time. So, the disasters that are in this research are Earthquake in Pakistan in 2013, California Earthquake in 2014, Chile Earthquake in 2014, Ebola virus pandemic in 2014, India floods in 2014, Hurricane in Mexico in 2014, Middle East Respiratory Syndrome in 2014, Pakistan floods in 2014, Cyclone in Vanuatu in 2015, Nepal Earthquake in 2015, Philippines Typhoon Hagupit in 2014. All the tweets are in English, also contains retweets as received from Twitter API.

Firstly, features are generated from the tweet text. A set of 11 low-level lexical and syntactic features to identify the more complex notions of subjectivity and formality of tweets, are used for predicting the label. There are different labels specific to each type of disaster. For example, for earth related, the labels used by annotators are

1- Injured or dead people---Reports of casualties and/or injured people due to the crisis

2- Missing, trapped, or found people---Reports and/or questions about missing or found people

3- Displaced people and evacuations---People who have relocated due to the crisis, even for a short time (includes evacuations)

4- Infrastructure and utilities damage---Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored

5- Donation needs or offers or volunteering services---Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood, and volunteering services

6- Caution and advice---Reports of warnings issued or lifted, guidance and tips

7- Sympathy and emotional support---Prayers, thoughts, and emotional support

8- Other useful information---Other useful information that helps understand the situation

9- Not related or irrelevant---Unrelated to the situation or irrelevant. There are similar labels to other disasters related tweets.

Support Vector Machines Classifier (SVM) and Random Forest Classifier (RF) are used to predict labels of tweets. The accuracy and F1-score of the two models are used as baseline performance and is compared with models using *Term Frequency – Inverse Document Frequency (TF-IDF)* statistic for classification. Also, as a final comparison it is compared with *BERT Classifier* and classification using *word2vec* word embeddings.

6 Research Findings

For each dataset type, set of 11 low-level lexical and syntactic features are used and support vector machines classifier and random forest classifier model results are set as *baseline* performance classification. It is compared with TF-IDF for a possible improvement and finally with BERT and word2vec word embeddings classification.

Disaster type (dataset)	11 Features - Baseline - SVM		11 Features - Baseline - RF	
	Accuracy	F1-score	Accuracy	F1-score
Earthquakes	38.68%	27.69 %	41.32%	32.32 %
Floods	44.73%	36.70 %	50.55%	42.63 %
Pandemic	30.56%	14.31 %	32.79%	23.70 %
Hurricanes	29.39%	17.53 %	32.84%	24.99 %

Table 1: Baseline performance of SVM and RF classifier using 11 low-level lexicon and syntactic features.

From Table 1, it can be understood that using decision trees over linear models for text classification has greater advantage which increased the accuracy and f1-score of the model. Next, it is compared to the TF-IDF features model classification. There features are actually vectors of each word based on their frequency in the entire tweets' data and also with other tweets in the same disaster type. The results of which are shown in Table 2.

Disaster type (dataset)	Using TF-IDF	
	Accuracy	F1-score
Earthquakes	63.10%	62.30%
Floods	70.00%	69.00%
Pandemic	65.00%	64.00%
Hurricanes	59.00%	59.00%

Table 2: Improved performance after using TF-TDF features for classification

Finally, it is further improved by using word embeddings from *word2vec* model, and also using BERT algorithm. The results of which are shown in Table 3.

Disaster type (dataset)	BERT		Word2Vec	
	Accuracy	F1-score	Accuracy	F1-score
Earthquakes	69.2%	62.00 %	67.00%	66.00 %
Floods	70.9%	75.42 %	65.00%	63.00 %
Pandemic	67.3%	69.93 %	58.00%	57.00 %
Hurricanes	70.0%	71.60 %	59.00%	58.00 %

Table 3: Performance of BERT and word2vec text classification models

The comparison is further illustrated via two charts. Chart 1 showing just comparing baseline performances and Chart 2 showing performance comparisons of TF-IDF, BERT, and word2vec classifications.

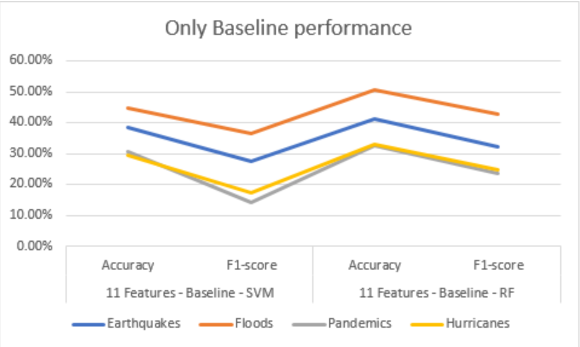
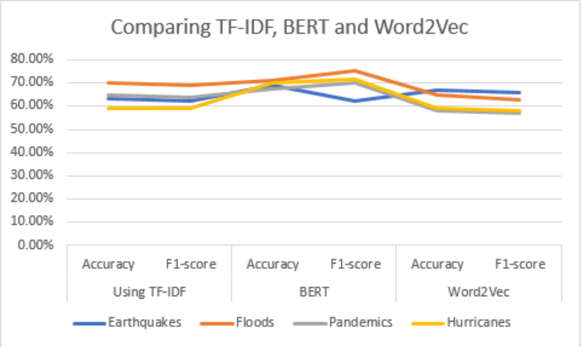


Chart 1: Line Chart showing accuracy and f1-score of baseline models.

From this, we can say that using state-of-the-art NLP models on text classification of tweet data during disasters would be more beneficial than using low-level features that has been done in prior researches in this field.



From both the charts, what becomes apparent is using term frequencies and inverse document frequencies during disasters for tweet classification would provide information that is worthy of the 9 labels as classifications. Through this model classifications, we can classify future tweets into any one of the labels and in real time, this information can be published on the public websites and telecasted resulting in the management disasters.

This process may not be completely reliable in times of natural calamities but by automating this process of information extraction and knowledge will help government officials and authorities to take necessary actions that aid the recovery.

7 Conclusion

In a nutshell, this research report discusses the scope of using state-of-the-art in times of disasters around the world irrespective of the type of the calamity, a broad level labels used previously for categorizing tweets would help important levels of departments in disaster-recovery team to understand the vital information present in tweets. The scope of this research is not limited to twitter data, the same text classification techniques can be applied to other research areas like topic identification of news articles, knowledge extraction from research papers etc.

8 Future Scope

As a future scope of this research analysis, text classification will always be better when using statistic like TF-IDF, word embeddings, BERT. It is recommended to progress by understanding how these models work and make improvements from there. It will sometime lead to void if we traditional lexicon and syntactic features as the word usage and internet slangs on social media is increasing more and more.

Acknowledgments

I would like to express my special thanks of gratitude to my professor *Bei Yu* at *Syracuse University* who gave me the golden opportunity to do this wonderful research project, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them.

References

1. Koustav Rudra, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2018. Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters. ACM Trans. Web 12, 3, Article 17 (July 2018), 35 pages. DOI:https://doi.org/10.1145/3178541

2. Arno Scharl, Alexander Hubmann-Haidvogel, Albert Weichselbraun, Gerhard Wohlgenannt, Heinz-Peter Lang, and Marta Sabou. 2012. Extraction and interactive exploration of knowledge from aggregated news and social media content. In Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems (EICS '12). Association for Computing Machinery, New York, NY, USA, 163–168. DOI:https://doi.org/10.1145/2305484.2305511

3. Carlo Lipizzi, Luca Iandoli, José Emmanuel Ramirez Marquez, Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers’ reactions to the launch of new products using Twitter streams, International Journal of Information Management, Volume 35, Issue 4, 2015, Pages 490-503, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2015.04.001

4. Wikipedia contributors. (2020, November 24). BERT (language model). Wikipedia.

https://en.wikipedia.org/wiki/BERT_(language_model)

5. Google Code Archive - Long-term storage for Google Code Project Hosting. (2013). Google Code Archive. https://code.google.com/archive/p/word2vec/
6. CrisisNLP.(2019).CrisisNLP. https://crisisnlp.qcri.org/lrec2016/lrec2016.html
7. Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242.