Project Proposal
Natural Language Processing with Disaster Tweets

Sina Abbasi

Motivation

Nowadays, thanks to smartphones and easy access to internet across the world the free social media applications are one of the best places to get the breaking news instantly. Especially after COVID-19 hit, news consumption from social medias rapidly increased where based on Pew Research Center survey conducted 31 August - 7 September 2020, about half of US adults (53%) get their news "often" or "sometimes" from social media [1]. Moreover, for example, about half of Twitter's users get their news from there as you can see in Fig. 1 [1]. Moreover, people use Twitter to report emergency news that are seeing them in real-time. Therefore, many companies and news agencies are interested to monitor the Twitter to get real-time and breaking news. More importantly, organizations such as disaster relief organization want to know the news about disasters as soon as possible. This monitoring can not be done by humans because we are dealing with on average 6000 tweets per second.

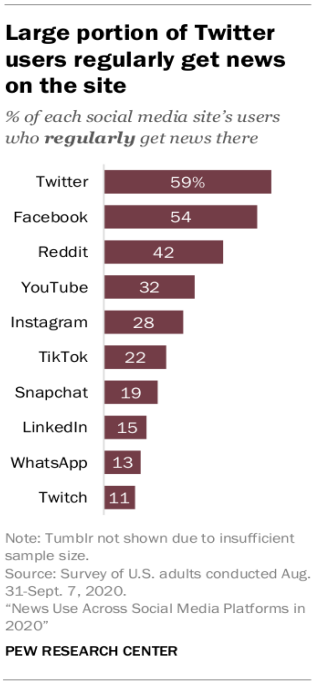

Figure 1: Percentage of users that get their news from that social media.

Pragmatically monitoring Twitter by just analyzing the words in each sentence separately is not a promising approach. For example, the author of a tweet said: "On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE" [2]; the word "ABLAZE" is used not in actual meaning but metaphorically.

Project details

In this project we are going to build a deep learning model that can classify which tweets are pointing a real disaster or not. Obviously, since we will work on text data, this task will be a Natural Language Processing (NLP) task. This challenge is a classification task and the sequential model (not trained model) can be used in other classification tasks in different domains, i.e. stock market predication based on tweets or news, sentiment classification for reviews of an online market like amazon, spam news classification and etc.

The input of the task will be text and the binary output should determine that weather the text describes an real disaster or not. The evaluation will be based on F1 score between predicted and expected answer and F1 is calculated as follow:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \tag{1}$$

where

$$precision = \frac{TP}{TP + FP}, \tag{2}$$

$$recall = \frac{TP}{TP + FN}, \tag{3}$$

TP, FP and FN stands for true positive, false positive and false negative, respectively. $F1 = 1$ is the best value that means perfect precision and recall and $F1 = 0$ is the worst.

Data set

In Fig. 2, you can see data set includes five columns where
`keyword`: a particular keyword from the tweet (may be blank)

	id	keyword	location	text	target
2333	3357	demolition	Lisbon, Portugal	Draw Day Demolition Daily football selection s...	0
1491	2149	catastrophe	Portugal	Alaska's #Wolves face catastrophe Denali Wolve...	0
5318	7593	outbreak	New York, NY	An outbreak of Legionnaires' disease in New Yo...	1
2252	3225	deluged	Clearwater, FL	@LisaToddSutton The reason I bring this up bc...	0
6653	9533	terrorist	????? ???? ????	#UdhampurAgain 2 terrorist shot dead.. #Udhampur	1

Figure 2: Data samples.

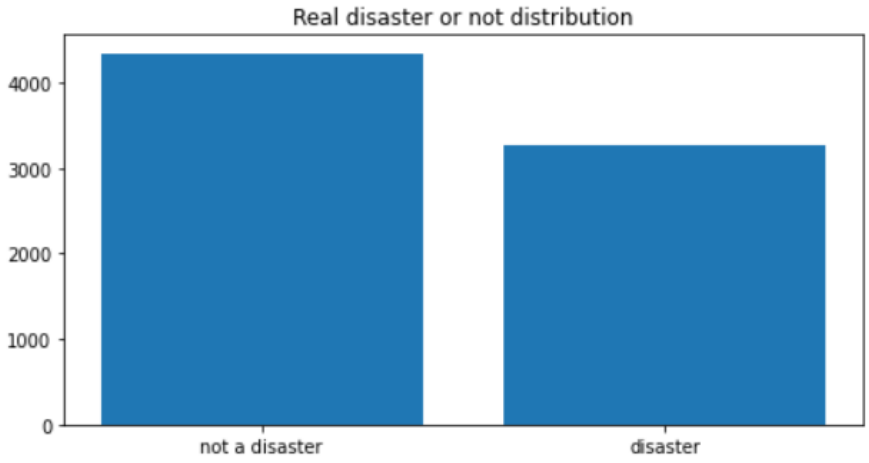

Figure 3: Distribution of target values.

`location`: the location the tweet was sent from (may be blank)
`text`: the text of the tweet
`target`: this denotes whether a tweet is about a real disaster (1) or not (0).
Train data has about 7.5k sample rows, and target values has fair distribution as you can see in Fig. 3.

Competition details

The task is an active kaggle competition with currently 830 competitors that you can find it *here*. Best score is 1.0 (perfect score) and top 50 scores vary between 0.84

to 1.0.

Related works

Authors in [3] works on similar data sets but the tweets related to hurricane. They get the data from Twitter and aiming to first classify and then analyse disaster-related tweets. In their binary classification phase, classification methods: LSTM, CNN, SVM, Logistic Regression and Ridge are performed and based on their evaluation Long Short-Term Memory (LSTM) get better fit to sequential order of textual data.

Looking through the top solutions in Kaggle website where they used LSTM structure to solve this task, we found and run a notebook which gave us 0.81 F1 score. This is a decent score for a NLP task. However, based on new viral method published by google researchers at 2018 called "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" which has already 34000 citation, the BERT can learn better contextual text [4]. Therefore, we are going to use this transformer for our task and try to get better F1 performance than LSTM.

References

[1] E. Shearer and A. Mitchell, "News use across social media platforms in 2020," 2021.

[2] "tweet: On plus side look at the sky last night it was ablaze," https://twitter.com/AnyOtherAnnaK/status/629195955506708480.

[3] M. A. Sit, C. Koylu, and I. Demir, "Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma," *International Journal of Digital Earth*, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.