Machine Learning Engineer Nanodegree Capstone Proposal

Real or Not? NLP with Disaster Tweets Challenge (Kaggle Competition)

Molise Molefi

January 16th, 2020

Proposal

Domain Background

Twitter is a 'microblogging' system that allows people to send and receive short posts called tweets. It is a platform to connect with people and catch on on everything that is happening around us, either be the news, emergencies, blogs and many more. This provides and encourages communication between multiple parties and also informs the public about events that are around them and that may affect them.

Twitter has approximately over 126 million daily users worldwide. 6,000 tweets are tweeted every second, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. Due to the high usage of Twitter and the ubiquitousness of smartphones, Twitter has become an important communication channel in times of emergency. People prefer to announce an emergency they're observing in real-time on Twitter because their posts can reach many people instantly on Twitter. Because of this, more agencies are interested in programatically monitoring Twitter for disasters(i.e. disaster relief organizations and news agencies).But, it's not always clear whether a person's words are actually announcing a disaster.

Problem Statement

The goal of this project is to predict which tweets are about real disasters and which ones aren’t. To solve the problem, I am going to use Natural Language Processing and Machine Learning techniques to develop a model that can classify real disasters from unreal ones.

Datasets and Inputs

This dataset was created by the company figure-eight and originally shared on their ‘Data For Everyone’ website. It contains three .csv files which are:

- train.csv - the training set
- test.csv - the test set
- sample_submission.csv - a sample submission file in the correct format

1. Train.csv and Test.csv Column names

- id - a unique identifier for each tweet
- text - the text of the tweet
- location - the location the tweet was sent from (may be blank)
- keyword - a particular keyword from the tweet (may be blank)
- target - in train.csv only, this denotes whether a tweet is about a real disaster (1) or not (0)

2. Train.csv

- Number of columns: 5
- Number of rows: 7614

3. Test.csv

- Number of columns: 5
- Number of rows: 3264

4. Sample_submission.csv

- The file is not important for the current project because it is used for model evaluation.

The datasets can be found at Kaggle by following this link

Solution Statement

The proposed solution to this problem is to apply Machine learning and Natural Language Processing techniques to process and classify the data(tweets) as Real Disaster or Unreal Disaster.

Firstly, the tweets will be converted to lowercase and tokenized into words. Then form a wordcloud based on the most occuring words. All tweets will be converted to numbers by using the index of the words in the WordCloud. Because we are dealing with a classification problem, either RNN or XGBoost model will be used for classification of tweets as real and unreal disasters.

After training the model we are going to use the evaluation metrics described in later sections to analyse the performance of the model.

Benchmark Model

For this problem, the F1 Score of 1 is the best benchmark model. Also, comparing our model prediction results to the submission file offered by kaggle will also serve as a benchmark model. The model predictions will have to match all the results in the Kaggle Submission file.

Evaluation Metrics

Prediction results are going to be evaluated using F1 score between the predicted and expected answers. F1 which is a function of Precision and Recall that is used to test the accuracy of the model. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results returned by the classifier, and *r* is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). If The model has F1 score that is close to 1 then it is accurate.

Project Design

Data Preprocessing

- Remove noise such as '@user' from all tweets
- Stem the cleaned tweets (i.e Using PorterStemmer)
- Tokenize the tweets into words
- Convert every word to a number using a vocabulary

Data Splitting *Provided data is already split into test and training datasets

Model Training and Evaluation *Different classifications models will be tested and the best performing model (i.e. RNNs, XGBoost) will be picked.

- The evaluation will be done based on the Evaluation Metrics section above.

Reference

[1] Accuracy, Precision, Recall or F1? TowardsDataScience
[2] F1 score Wikipedia
[3] Real or Not? NLP with Disaster Tweets Kaggle
[4]Twitter keeps losing monthly users, so it’s going to stop sharing how many TheVerge
[5] The Number of tweets per day in 2019 Dsayce