Classification of Disaster Tweets using NLP

Team: Manasa Hari, Amit Kamboj, Feiyu Cai, Tushar Sharma

Abstract— Twitter has been the most important part of digital communication these days. It has been the channel of communication across the world for daily news starting from television gossips to case-of-emergency news. But due to large amounts of data flooding into the twitter base every minute, sometimes, a highly important piece of news such as a disaster occurrence can be submerged amidst not at all important tweets. Due to this, it may take days or even weeks before the disaster news reaches to people, meanwhile, a huge irrevocable loss can happen already. This project aims to highlight the disaster emergency news so that users can be constantly notified on the emergency issues out of millions of other tweets. We take the challenge to build a machine learning model that classifies between tweets about real disasters and the rest. The key challenge is to distinguish metaphorical usage of tragedy vocabulary and the real intended usage of disaster terms. For example, a user tweets 'Thoughts are a storm, unexpected'. This is clearly a metaphorical statement. Even though it is obvious for humans to interpret that this tweet is not about a real disaster, but it is less clear to a machine. This problem is also an actively ongoing Kaggle competition. We want to explore possible predictors and conclude on the right predictors to solve this classification problem. The dataset has 10,000 tweets that were hand classified[1]. We will explore Naïve Bayes, LSTM and CNN classifiers to solve this problem.

REFERENCES

[1] Data and Problem source: https://www.kaggle.com/c/nlp-getting-started/data.

DATASET AND DESCRIPTION

Link: https://www.kaggle.com/c/nlp-getting-started/data

1. Data format: Each sample in the train and test set has the following information:
   a. The text of a tweet
   b. A keyword from that tweet (although this may be blank!)
   c. The location the tweet was sent from (may also be blank)

2. Files we have:
   a. train.csv - the training set
   b. test.csv - the test set
   c. sample_submission.csv - a sample submission file in the correct format

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

4. Output: We will predict whether a given tweet is about a real disaster or not. If so, predict a 1. If not, predict a 0.

TECHNOLOGIES WE PLAN TO USE: Python, Keras, Tensorflow.