Utilizing Twitter for Disaster Detection

Tofer Kim, Ritchie Kwan, and Will Stecher

INTRODUCTION

Traditional methods for alerting on disaster-related events like earthquakes and tsunamis rely on information derived from official sources (e.g. USGS). Social media platforms, such as Twitter, can also be a valuable resource for sharing information regarding disaster-related events.[1] For our project, we will attempt to identify *relevant* disaster-related tweets in order to investigate trends that can be used to detect natural disasters as they occur. Our methods can then be implemented to serve as an alert to first-responders and disaster relief efforts (e.g. FEMA) and potentially help save lives.

GATHERING DATA

In order to train a model that could classify disaster-related tweets as *relevant* or *not relevant*, we needed a very specific dataset: Luckily, in September 2015, Figure Eight[2] publicly published a dataset which included 10,877 disaster-related tweets, “culled with a variety of search terms like ‘ablaze’. ‘quarantine’, and ‘pandemonium’, then noted whether the tweet referred to a disaster event, as opposed to a joke with the word or a movie review or something non-disastrous.”

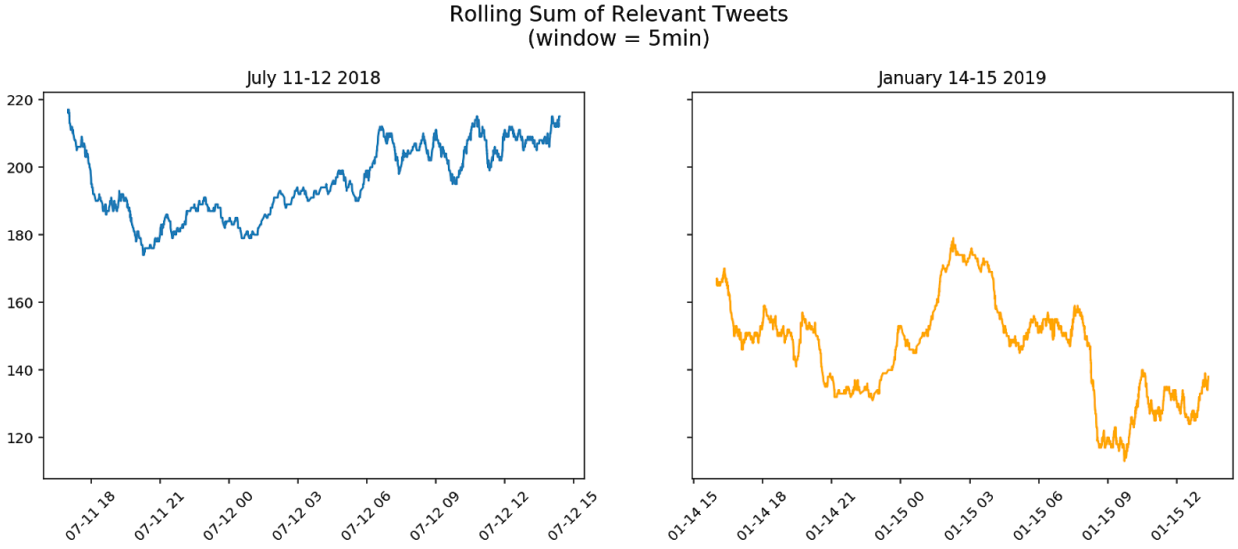
Using the available Twitter API to gather new tweets to test our model has many limitations; its date constraints restrict searches to only the past 7 days. Additionally, the API only allows the query to collect 100 tweets at a time. However, there are tools to get around these restrictions and provide the ability access older tweets more pertinent to our purposes. Jefferson Henrique[3], details exactly how to do so, and following his instructions, we were able to search for tweets using specific keywords and date-ranges. Here, we decide to narrow the scope of our project to detect a specific natural disaster--wildfires--using the keywords "wildfire" and "forest fire" to collect tweets from two distinct time periods: the beginning of the 2018 California wildfires and a more recent period after these fires had subsided.

MODELING

The most distinguishing features of a tweet are its words/text. NLP (Natural Language Processing) is a way to extract these features in a way that can be used to then train a classification or regression model. After removing links and tokenizing, then stemming each word in a tweet from our training dataset, we built a Doc2Vec algorithm to engineer a vector space for our observed vocabulary of words.[4] These vector features are then used to train a logistic regression model, which returns a probability used for classification: identifying whether a tweet is *relevant* or *not relevant* to a disaster-related event. We evaluate our trained model by its 83% accuracy score on unseen data, compared to a baseline prediction of 57%. (Our classification methods of using Doc2Vec for Logistic Regression were adapted from Susan Li's example[5]).

CONCLUSION

After running our trained logistic regression model to classify tweets from the specified time-ranges above, we analyzed the frequency of *relevant* tweets using a "rolling sum" function. Our findings show that at beginning of the 2018 California wildfire disaster, there were consistently over 180 "relevant" tweets within a window of 300 seconds (5 minutes); More recent tweets did not exceed this threshold of 180 "relevant" tweets. This distinction can be used to detect future wildfire disasters in as short of a delay as 5 minutes.



NEXT STEPS

We can implement our outlined methods with new keywords and date-ranges in order to observe and detect different types of natural disasters. Likewise, if we continue to train our Doc2Vec algorithm with more data (e.g. Wikipedia corpus), we can use its properties to classify detected disaster-related events and use cosine similarity in order to rank which tweets may be more urgent. With access to Twitter geolocation and population data, we can further improve our functionality and potentially estimate local areas and number of people affected by future natural disasters, as soon as they occur.

REFERENCES

Ford, Jordan. (2018, Jul 11). "Improving disaster response through Twitter data". [https://phys.org/news/2018-07-disaster-response-twitter.html]

Figure Eight. (2015, Sep 4). "Data For Everyone". [https://www.figure-eight.com/data-for-everyone/]

Henrique, Jefferson. (2018, Nov 21). "GetOldTweets-python". [https://github.com/Jefferson-Henrique/GetOldTweets-python]

Shperber, Gidi. (2017, Jul 25). "A gentle introduction to Doc2Vec". [https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e]

Li, Susan. (2018, Sep 17). "Multi-Class Text Classification with Doc2Vec & Logistic Regression". [https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4]