Detection and Analysis of Disaster-Related Tweets

Daniel Solomon
solomond@mail.tau.ac.il

Gal Ron
galr1@mail.tau.ac.il

Omri Ben-Horin
EMAIL@mail.tau.ac.il

Abstract

TODO:

1 Introduction

The popular microblogging service Twitter is a fruitful source of user-created content. With hundreds of millions of new tweets every day, Twitter has become a probe to human behavior and opinions from around the globe. The Twitter 'corpus' reflects political and social trends, popular culture, global and local happenings, and more. In addition, tweets are easy to access and aggregate in real-time. Therefore, we experience an increased interest in natural language processing research of Twitter data.

As one of the world's most widely used social networks, Twitter is an effective channel of communication and plays an important role during a crisis or emergency. The live stream of tweets can be used to identify reports and calls for help in emergency situations, such as accidents, violent crimes, natural disasters and terror attacks (which we all refer to as 'disasters' in this paper).

In this work we utilize techniques from the natural language processing pipeline (tokenization, part-of-speech tagging and named-entity recognition) to work on Twitter data, as opposed to traditional corpora, in order to detect and analyze disaster-related tweets.

The Dataset We present our experiments on a dataset of 10,877 tweets[1], labeled to *'disaster-related'* and *'not disaster-related'* with confidence in the range $[0, 1]$. For example, the following tweet is *'disaster-related'* with confidence 1,

> Thunderstorms with little rain expected in Central California. High fire danger. #weather #cawx http://t.co/A5GNzbuSqq

while the following tweet is *'not disaster-related'* with confidence 0.59,

> It's been raining since you left me // Now I'm drowning in the flood // You see I've always been a fighter // But without you I give up

Even for one who is not familiar with the Bon Jovi lyrics in the latter tweet, it is clear that the tweet does not refer to a real natural disaster. However, this observation is hard to make examining only the vocabulary used; the latter tweet contains a variety of 'disastrous' words (e.g. raining, drowning, flood, fighter). This example hints that in order to reach meaningful results we may want to examine additional linguistic features of colloquial writing, as well as Twitter-specific features such as hashtags (#), user at-mentions (@), internet links and emojis.

Our Contribution In this paper we present our work tackling three missions involving disaster-related tweets.

The first mission is *identification* of disaster-related tweets among a variety of tweets. We implemented several classifiers, the best of which achieved TODO: % accuracy on the dataset. We note that this method could have easily been adjusted to identify tweets related to themes other than disasters (e.g. politics-related, sports-related, etc.), given the appropriate dataset.

The second missions is binary classification of disaster-related tweets to one of two categories, *subjective tweets* (i.e. tweets that express an emotion) vs. *objective tweets* (such as news reports on disasters). To achieve this we manually tagged 2,410 disaster-related tweets. The rationale behind this task is that objective tweets like informative news reports are likely to be published after the event had already become clear to emergency services, while subjective tweets may contain invaluable first-person testimonies.

Finally, we extracted named entities to enrich our knowledge on the disaster (mostly location)... TODO: Omri - short description of method and achievements.

To demonstrate our framework we aggregated recent tweets from various locations in the US, extracted

disaster-related tweets using our classifier, and used named-entity recognition to discover entities related to ongoing disasters. For example, "Hurricane Harvey" appeared as a top named-entity among recent tweets sent from Houston, TX, which we identified as *disaster-related*.

The code of our project is available at `https://github.com/glrn/nlp-disaster-analysis`.

1.1 Twitter vs. Traditional Corpora

Tweet datasets have some unique features that differ from traditional corpora (such as WSJ corpus). These features should be taken into consideration when implementing natural language processing techniques.

> Heard about #earthquake is different cities, stay safe everyone.

TODO: (Gal) Complete this section

2 Analysis Wokrflow

keywords TODO

- A
- B
- C

TODO: (Gal) Complete this section

3 Classification of Disaster-Related Tweets

TODO:

4 Sentiment Analysis of Tweets

TODO:

5 Named-Entity Recognition in Tweets

TODO:

6 Experimenting with Recent Tweets

keywords Twitter's Search API

7 Conclusions

Future work TODO

References

[1] KEVIN GIMPEL, NATHAN SCHNEIDER, B. O. D. D. D. M. J. E. M. H. D. Y. J. F., AND SMITH, N. A. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proc. ACL* (2011).

[2] RITTER, A., CLARK, S., MAUSAM, AND ETZIONI, O. Named entity recognition in tweets: An experimental study. In *Proc. EMNLP* (2011).

Notes

[1] **"Disasters on social media" Dataset by CrowdFlower**: Contributors looked at over 10,000 tweets culled with a variety of searches like "ablaze", "quarantine", and "pandemonium", then noted whether the tweet referred to a disaster. https://www.crowdflower.com/wp-content/uploads/2016/03/socialmedia-disaster-tweets-DFE.csv