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 much 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).

The Dataset In this work we present our experiments on a dataset of 10,877 tweets[1], labeled to *'disaster-related'* and *'not disaster-related'* with confidence in the scale $[0, 1]$. 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

and the following tweet (containing a Bon Jovi lyrics) 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

TODO (Gal): explain here that the 2nd tweet contains more 'disasterous' words (drowning, fighter, give up, raining, flood), which explains why the task is challenging; we'd have to look at more than unigrams to succeed.

Our Contribution TODO (Gal): organize and expand this section a bit

In this work we present a model trained to identify disaster-related tweets from other messages, using a natural language processing pipeline adjusted to the special features of Twitter tweets. In addition, we present two experiments conducted on disaster-related tweets.

First, we learn to separate *subjective* tweets (example: TWEET) from *objective* reports on disasters (example: TWEET). This involved manual tagging of 2100 tweets. We also recognize named-entities in disaster-related tweets to enrich our knowledge on the disaster (mostly location).

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.

Here's a tweet:

RT This is an #awsome tweet lmao :O

TODO (Gal): Complete this section

2 Analysis Wokrflow

keywords TODO

- A
- B
- C

TODO (Gal): Complete this section

3 Tweet Classification

keywords TODO

4 Named-Entity Recognition in Tweets

keywords TODO

5 Experimenting with Recent Tweets

keywords Twitter’s Search API

6 Conclusions

Future work TODO

References

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