Determining the Type of Disaster With Natural Language Processing

Correlational Study

Jacob Voyles

CPTS 115 - Into to Data Analysis - Washington State University
Submitted 13 Dec 2020

Abstract

The use of tools like Twitter during disasters offers a valuable source of information for disaster response agencies, as it often provides critical up-to-date and on-location updates about an unfolding crisis. This precipitates an interest in robust processing and visualization tools. I explore the use of NLP models for the analysis of disaster-related Twitter data.

Introduction

Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programmatically monitoring Twitter (i.e. disaster relief organizations and news agencies).

But, it's not always clear whether a person's words are actually announcing a disaster. Take this example:

'On the plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE"

The author explicitly uses the word "ABLAZE" but means it metaphorically. This is clear to a human right away.. But it's less clear to a machine

Due to their ease of use and simplicity, social media platforms can provide efficient delivery of information that can give better situational awareness for emergency response. Unfortunately, this vast amount of information can be useless or even dangerous, since its reliability is often unclear and any uncertainties can result in chaos

The rapid growth of online information services, social media, and other digital format documents mean that large amounts of information are becoming immediately available and readily accessible to numerous end-users. However, the human ability to organize and understand a large number of social media texts is limited. Methods requiring extensive human attention and interpretation do not scale robustly to large social media data sets.

Moreover, most of these numerous posts on Twitter are not useful in providing information about disasters.

Methods

Tweet contents analysis can be applied to all kinds of text analysis but certain domains and modes of

communication tend to have more expressions of very short text messages.

Social media mining for disaster response and coordination has been receiving an increasing level of attention from the research community.

The lack of well-defined values in choosing machine learning algorithms suitable for a given problem remains a major challenge. To address these problems, I analyze statuses updated on Twitter about natural disasters and perform automatic classification.

The next section describes the three main functions of the proposed system: Data Collection, Tweet Preprocessing, Feature Extraction:

Data Collection

This data was provided and collected from Appen Datasets. The dataset includes 30,000 messages from events including an earthquake in Haiti in 2009, to floods in Pakistan in 2010 and spanning a large number of years and 100's of different disasters.

The dataset is split with 36 different categories related to disaster response but has been stripped of messages with sensitive information in their entirety.

Tweet Preprocessing:

This step preprocesses the tweet content before creating the numeric vector. Firstly, this task removes the tweets which already contains the same text as the previous preprocessed tweets to reduce the redundancy and noise.

Secondly, stop_words from collected tweets are used to reduce the dimensionality of the dataset, and thus terms left in the tweets can be identified more easily by the feature extraction process. Stop_words are common and high-frequency words such as "a", "the", "of", "and", "an" "in" etc.

Finally, the stemming process converts all the inflected words present in the text into a root form called a stem. For example, 'automatic,' 'automate,' and 'automation' are each converted into the stem 'automat'. For the purpose of stemming, this system uses a popular snowball stemmer.

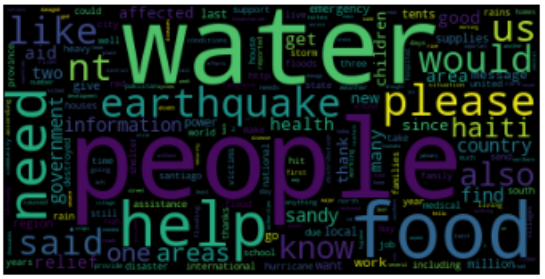

Figure 1: Word bubble, showing the most common words used in disasters, sorted by size

Feature Extraction

Feature extraction is the transformation of arbitrary data such as images or text into numerical features usable for classification. In this work, it is concerned with altering tweet contents into a simple numeric vector representation To do this, each tweet is tokenized into hashtags, user mention, URLs, special characters such as punctuation or emotion using LancasterStemmer.

These stems are then vectorized using TfidfVectorizer and mapped to a multinomial Naive Bayes classifier to make a prediction model.

Disaster	Accuracy Score
Child Alone	1.000000
Tools	0.998479

Shops	0.997337
Fires	0.995816
Aid Centers	0.994294
Security	0.994294
Missing People	0.993914
Hospitals	0.993153
Search and Rescue	0.989730
Clothing	0.985926
Other Infrastructure	0.984024
Electricity	0.984024
Cold	0.982883
Money	0.973754
Refugees	0.971852
Infrastructure	0.969951
Military	0.962343
Medical Products	0.959680
Other Weather	0.959680
Buildings	0.956257
Death	0.952834
Transport	0.950552
Earthquake	0.925827
Water	0.925447
Medical Help	0.916318
Other Aid	0.916318
Shelter	0.912514
Storm	0.912134
Floods	0.906428
Food	0.877900
Weather-Related	0.832256
Aid Related	0.773298

Table 1: Table showing the accuracy of different models at predicting a disaster.

Conclusion

Social media mining for disaster response and coordination has been receiving an increasing level of attention from the research community. It is still necessary to develop automated mechanisms to find critical and actionable information on Social Media in real-time. The proposed system combines effective feature extraction using NLP and machine learning approach to obtain the annotated datasets to improve disaster response efforts. In the future, I will investigate the specific variation of terms over different disasters to perform annotation on all disasters. I hope to create a disaster lexicon in more detail to improve accuracy.

References

1. “Multilingual Disaster Response Messages | Appen Datasets.” 2018. Appen. May 7, 2018. https://appen.com/datasets/combined-disaster-response-data/.