Improving Emergency Service Resource Allocation by Analyzing Sentiment in Twitter Data

Gregory Burton, Indiana University Bloomington
Micheal Smoot, Indiana University Bloomington
Zhongkun Xiang, Indiana University Bloomington

Natural disasters happen all around the world, despite substantial efforts by governments in relieving negative disaster related aftereffects, significant work remains in improving those services. Monitoring individuals level of satisfaction can be potentially important in informing FEMA in their performance. English language Tweets were collected that contained the keyword FEMA from Florida and California; two states that were recently affected by natural disasters, Hurricane Irma and the Sonoma region wildfires. The Tweets were analyzed for sentiment scores to compare the level of FEMA relief effort satisfaction between Florida and California. Interestingly enough, Tweets from both states contained large amounts of text relating to President Donald Trump and politics, therefore each dataset was further divided into two subsets: Tweets that contain Donald Trump content and Tweets not containing Donald Trump content. After running sentiment analysis on all datasets, it was discovered that California and Florida Tweet datasets expressed similar sentiment toward FEMA, and in both states, sentiment scores of Tweets containing Donald Trump content were measurably more negative than tweets not containing Donald Trump related text.

Additional Key Words and Phrases: Social media, Twitter, tweets, data mining

1. INTRODUCTION

With the recent catastrophic natural disasters throughout the United States, emergency service resources have been stretched thin. Efficient resource allocation of emergency capabilities is critical during large scale disaster incidents. Improved positioning of capabilities will lead to cost reductions in regard to relief efforts and expedited recovery operations. With the recent interest in analyzing social media data for sentiment [Mishne Rijke, 2006, Yang, Hsin-Yih Lin, Chen, 2007], this study will further the field by applying sentiment analysis to natural disaster events with the aim to advance relief efforts.

In this initial study, Twitter.com datasets were chosen for use with a focus on two locations, Florida and California, for a period near the Hurricane Irma event and California Wildfires. It is expected that a difference in sentiment in regard to the Federal Emergency Management Agency (FEMA) led efforts between the locations will be discovered through the analysis. Additionally, the study may prove useful in providing additional relief capabilities during emergency events outside of natural disasters such as large-scale terrorist attacks.

2. LITERATURE REVIEW

While a single tweet may contain multiples features such as text emoticons and hashtags, not all of these features are useful to sentiment analysis. Choosing the most relevant feature will be critical to our study. Kouloumpis and his team [2011] performed sentiment analysis on Twitter data with various features and concluded that sentiment lexicon, emoticons, and abbreviations were proven more useful. In contrast, Poddar, Halder, and Jia performed a study using POS tags, and achieved a lower accuracy than other similar studies with better feature selection [2016].

Similarly, Go and his team [2009] conducted research regarding tweet features in sentiment analysis, using multiple machine learning techniques (e.g. Nave Bayes, Support Vector Machine) to classify different sentiments, the combination of unigrams and bigrams achieved the highest accuracy. While some have attempted more manual methods of assigning sentiment to various parts of a message, greatest accuracy lies within these machine learning models [Thakkar, H., Patel, D., 2015].

In the context of disaster and tragedy, Wang, Varghese, Donnelly suggest sentiment towards a particular political view, such as being pro-gun vs. anti-gun (2016). The idea of leveraging sentiment to a specific opinion may help identify those suffering from less-than-ideal relief efforts relative to other areas in a similar situation. Many researchers have attempted to infer disaster effects using sentiment such as severity of storm damage, but fell short. Though its important to understand the effects of disaster, existing studies do not leverage appropriate features for this use-case and did not achieve statistically significant results [Kryvasheyeu, et al. 2015].

3. DATASET DESCRIPTION

Twitter.com is an ideal data source for monitoring sentiment [Saif, Fernandez, He, Alani, 2013]. Twitter is a web based social media platform that enables users to write and read brief 280-character messages colloquially referred to as Tweets. Twitter.com Tweets provide a real-time tapestry of on the ground sentiments.

The Twitter.com datasets were harvested using the Twitter provided Application Programing Interface (API) and Python 3.6. The Florida dataset was harvested from 19 October 2017 to 21 October 2017 using the geocode function with a grid coordinate of 27.157829, -81.490444 and a 700-kilometer radius. The California dataset was harvested from 18 October 2017 to 27 October 2017 using the geocode function with a grid coordinate of 36.778261, -119.4179324 and a 500-kilometer radius. Multiple radius ranges from 10-kilometer to 1000-kilometer were tested with the above ranges providing adequate data for analysis without geographic overlap. All Tweets harvested had geolocation services enabled. Non-geolocation tagged Tweets were not harvested. The data gathering was restricted to the two regions for this initial proof of concept and gathering only data required to perform analysis. In an effort to limit the data harvesting to those Tweets directly related to FEMA and relief efforts the Twitter search API was called using keywords fema, FEMA, and Fema, with language constraints set to English. Additional Python scripted data cleaning was performed removing aberrant non-English language data, setting all Tweets to lowercase, common stop words were removed such as a, and the. The datasets were extracted into a CSV format for further manual inspection. After the automated cleaning the Tweet data requiring additional refinement, such as duplicates with unique universal resource locators were manually removed. The cleaned datasets were then loaded into Pandas data frames with the following headers: Date, Text, Hashtag, and BinaryID. The columns of tweets without hashtag will be empty string. The California dataset was labeled with a BindaryID of 0 and the Florida dataset was labeled with a BinaryID of 1.

4. RESEARCH DESIGN AND METHODS

The goal of the study is to identify overall satisfaction with the Federal Emergency Management Agency responses surrounding Hurricane Irma and the California wildfires in the summer of 2017. To identify the satisfaction between both populations, sentiment analysis models were leveraged to obtain a general quantification of the citys attitude toward the agency. Because the Twitter output was pre-processed to contain clean text data, two distinct approaches were used to generate sentiment estimates of the population:

1.Lexicon-based Sentiment Scoring
2.Machine Learning Classification

First, the datasets were analyzed by observation to identify universal anomalies or obvious patterns in the extracted tweets. While the sample sizes are relatively small per city, it is crucial to achieve a holistic understanding of the data prior to automated analysis. It was found that 21.9 percent of the tweets gathered mentioned President Donald Trump (i.e. identifying @realDonaldTrump within the tweet body), which has great potential to influence the sentiment analysis unrelated to FEMA.

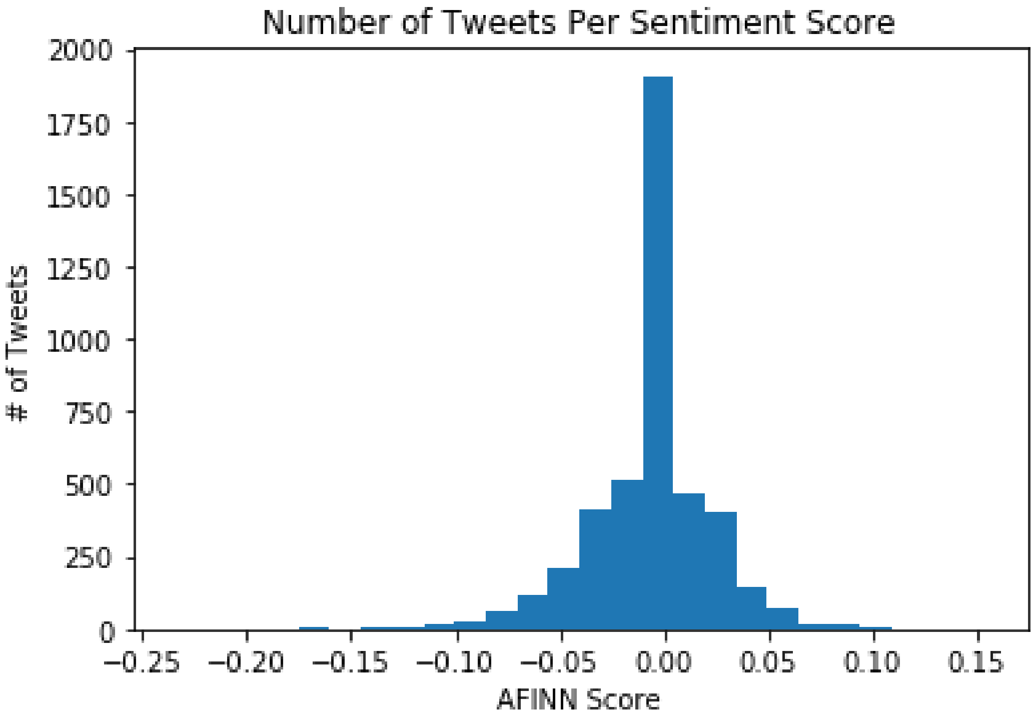
6Lexicon-based Sentiment ScoringUsing a vocabulary with pre-computed sentiment values will be used to achieve the first method. This approach allows for quick analysis on the tweet dataset without manual classification. The chosen lexicon is the AFINN vocabulary. This list of English words contains 2477 labelled words and phrases with a sentiment score of -5 to +5 (negative to positive, respectively).

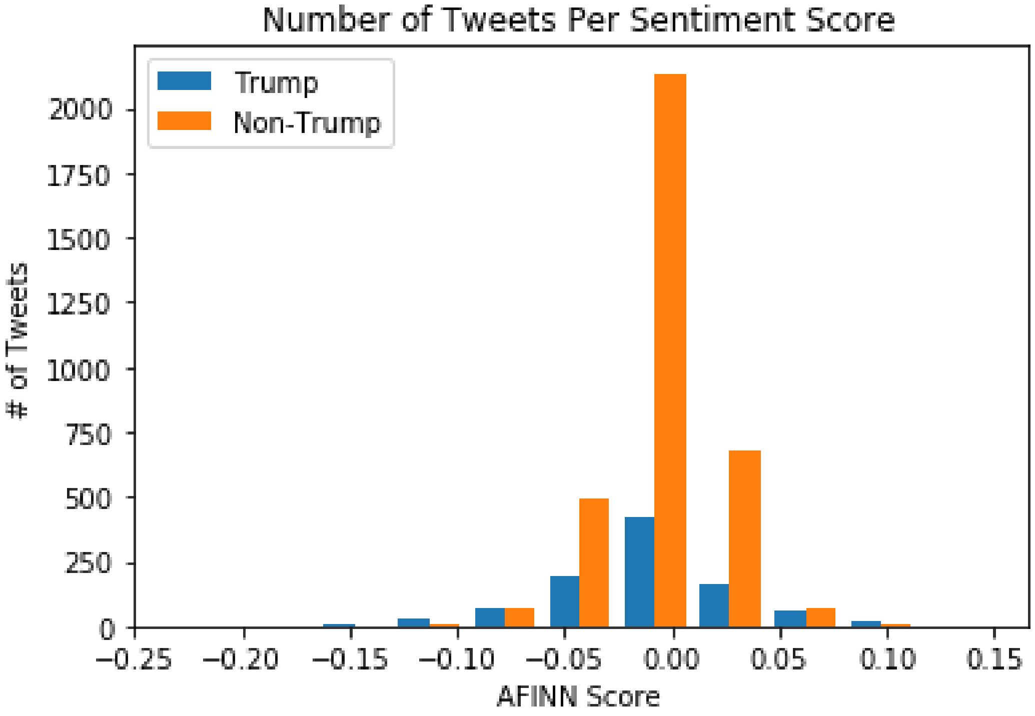
The AFINN sentiment scoring algorithm is applied to each tweet in the dataset for both cities. Finally, once each tweet for both cities have an attributed sentiment score, the datasets are normalized separately to produce scaled results for easier visualization and analysis.

	Trump-Related	No Trump
Florida	2.4637	7.4225
California	2.3696	6.0348

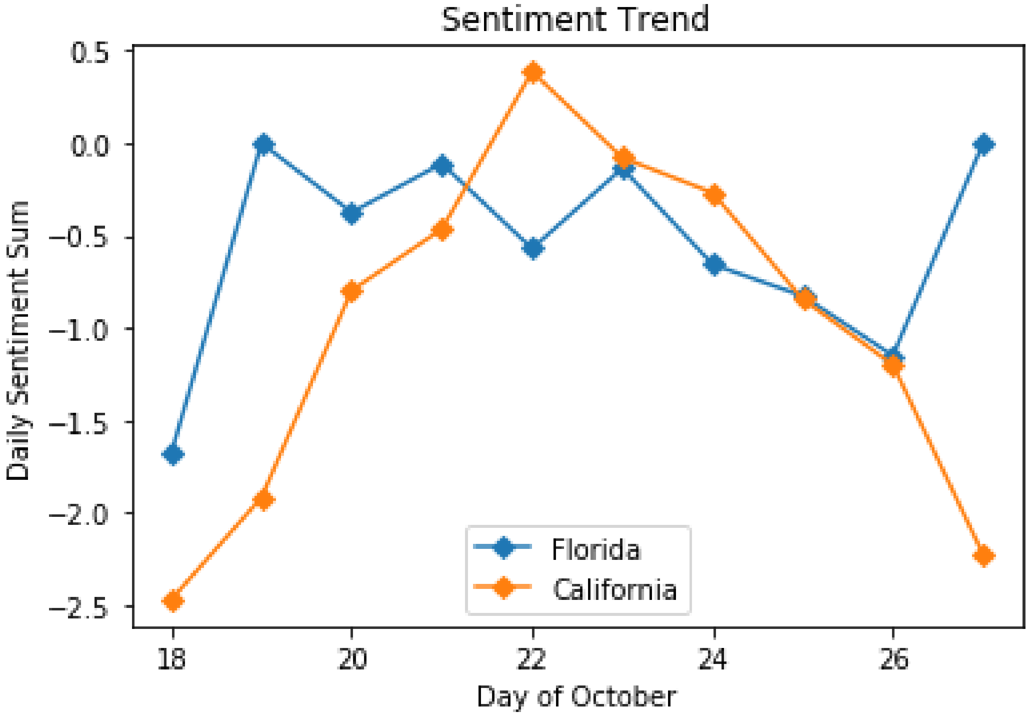
As indicated by the normalized AFINN scores, California tweets expressed a less negative sentiment during time of disaster, albeit only slightly. While this indicates a significant difference between cities during time of disaster, theres several considerations that must be accepted as possible factors that exemplify this difference.First, there is no control sentiment for this timeframe to which they can be compared. If the population often tweets with a less negative sentiment than Florida, these results may not indicate an effect of FEMAs performance. Additionally, the nature of both disasters are vastly different a hurricane may be much more detrimental to a populations attitude than a wildfire, and understanding the drop in sentiment during these times proportionally will help identify if a geographic location is received disproportionate attention than necessary.

Finally, other events outside of the disaster within the timeframe could also affect the emotions of tweeters and would server as indirect bias for the sentiment negatively. For example, recent political headlines could reduce morale overall as society, which has potential to affect the sentiment of tweets regardless of topic.


Number of Tweets Per Sentiment Score
# of Tweets
2000
1750
1500
1250
1000
750
500
250
0
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
AFINN Score


Number of Tweets Per Sentiment Score
Trump
Non-Trump
# of Tweets
2000
1750
1500
1250
1000
750
500
250
0
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
AFINN Score

Lastly, using this time frame we can attempt to understand when sentiment towards FEMA experienced noticeable changes. When attempting to leverage these techniques to identify gaps in resource allotment, finding when sentiment is at its lowest in comparison to disaster times can aid in understanding the impact of provisioning delays and lag times. When analyzing the trends, its clear that the California population reached a maximum sentiment during time of disaster around the 22ndof October. However, this increase in sentiment is counterintuitive to disaster, and could have been influenced by outside factors. Florida, in contrast, appears less stable and shows a maximum towards the end of disaster.



While attributing FEMA sentiment on Twitter to resource provisioning presents significant challenges, it is also uncertain that a lexicon-based approach offers a confident measurement of sentiment when analyzing small and sometimes cryptic bodies of text. Attempting to classify sentiment based on previous data can help improve our indication of whether the lexicon-based approach is reliable. However, to train models for application on the rest of the data set, several tweets must already be classified. To solve for this, 300 tweets of the California tweet set were hand-classified with a sentiment polarity of -1 for a negative sentiment, +1 for positive, and 0 for neutral. While one judge is sufficient, two team members independently classified their set and a correlation was generated using the Cohen-Kappa score, 8achieving roughly 0.400 (indicating a majority of classifications were agreed upon between both raters).Once the set of tweets was hand-classified, the featureswere generated by vectorizing the tweet body with TF-IDF (term frequency-inverse document frequency) to identify important words without considering commonly occurring words. All words that are tokenized for including in the feature set are lowercased prior.Three models in total were chosen to train for classifying the remaining data set: Support Vector Machine, Logistic Regression, and a generalized Linear Regression model. Upon random test runs, the Logistic Regression model appeared to have the highest accuracy. To verify, using 5-fold cross validation generated a 76.5 percent accuracy using an 80 percent training / 20 percent testing split on the 300 hand-labelled.

Support Vector Machine	Linear Model	Logistic Regression
0.683	0.800	0.817

While the accuracies for the trained models are promising, a significant issue is presented. One problem is the class distribution in the training set. Not only is the training size is relatively small (only about 8 percent of the total data set), but further investigation exposes a very small population of positive tweets. This leaves the model classifying positive tweets very poorly, and helps reduce accuracy

5. DISCUSSION

While Twitter data has proven useful for specific sentiment analysis problem sets, using Tweet data as a sole data source for disaster relief satisfaction measurement is not sufficient with current machine learning techniques [Saif, Fernandez, He, Alani, 2013]. The harvested Twitter data was not delivered in sufficient volume when constrained to specific geographic locations surrounding an event, such as a grid coordinate of 27.157829, -81.490444 and a 25-kilometer radius. With the narrow geographic radius setting and very few Tweets being harvested the data is not representative of the 20 million plus residents of an area such as Florida. Additionally, individuals tend to stray significantly from the intended target conversation such as FEMA and Hurricane Irma. When the data was manually reviewed a significant portion of the Tweets were directed at President Trump and politics or the current contracting disputes currently circulating in the media regarding whitefish and Puerto Rico, and not the targeted subject of disaster relief efforts. The non-disaster related Tweets tended to be severely negative. A recommendation to include multiple data sources such as local news comments, Facebook, Twitter, and other Social Media platforms for a more holistic approach to the problem would be recommended. Discussions in news feeds tend to be more complete than short Tweets and may provide the additional data to enable higher machine learning performance.

Having a robust historical Twitter dataset would also improve analysis. The original study intent was to harvest data surrounding the events temporally with a dataset harvested a week prior, during, and after the disasters. Unfortunately, difficulty in getting the Python Get Old Tweets 3 (GOT3) package to perform well prohibited the desired dataset collection. The Twitter API does not allow access to Tweets older than seven days without a fee. A full Twitter dataset for each region studied would be useful in allowing a baseline sentiment to be calculated. As an example, California may often have a negative sentiment and when compared to the extracted datasets the team had no way of observing the data from that specific viewpoint.

6. CONCLUSIONS

Detecting peoples general satisfaction using social media is one of the fastest and most effective method to provide the government useful feedback and help them improve their services. From the findings of our study, people on both California and Florida are generally dissatisfied with Federal Emergency Management Agency.

However, there are a few limitations in this study. The biggest one is the quality of data. Tweets collected during and right after a catastrophic event can potentially be the best representative data, and may reflects peoples attitude and emotions toward FEMA accurately without too much noisy information. In our study, we tried using GOT3 package to collect old tweets during hurricane Irma and California wildfire but failed.

Let alone the number of data we have collected is relatively small, after manually inspecting the data, there is a large portion of tweets that are mainly about Donald Trump rather than FEMA directly, which brings up another limitation, the ability to clean the data properly. We only collected tweets that contain the keyword fema (both upper and lower case), which is not enough for detecting peoples attitudes toward FEMA directly. In future work, we expect to develop a better algorithm to identify the content of each tweet, and classify them accordingly.

Another feature we can develop is the future is keyword detection, after getting accurate data, we can use TFIDF or deep learning to find what are the most dissatisfactory factors people are talking about, and the government may use those feedback to adjust their service accurately.

7. TEAM MEMBER CONTRIBUTIONS

Greg worked the initial project idea, harvested data for both Florida and California, provided the initial cleaning, consulted on the analysis, and drafted a template for the paper. Micheal focused his efforts on the analysis portion and scripting in Python to process and analyze the datasets. Micheal was also the primary driver of the machine learning techniques chosen to use on the Twitter data analysis, and drafted the methods portion of the final document. Xiang harvested data for Florida and California, wrote a script to complete the cleaning of the harvested datasets, drafted the abstract section, formatted the document in ACM, consulted on the machine learning and Python script for Python, and closed out the work with a final summary section. All team members supported each other throughout the project and attended Zoom meetings as required.

8. REFERENCE

Go, A., Bhayani, R., Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Academia.org. Retrieved September 22, 2017.

Kouloumpis, E., Wilson, T., Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Icwsm. Retrieved September 21, 2017.

Kryvasheyeu, Y., Chen, H., Obradovich, N., et al. (2015). Nowcasting Disaster Damage. Arxiv.org. Retrieved September 23, 2017.

Mishne, G., Rijke, M. d. (2006). Capturing Global Mood Levels Using Blog Post. (pp. 145-152). Stanford: AAAI Spring Symposium - Technical Report.

Poddar, L., Halder, K., Jia, X. (2016). Sentiment Analysis for Twitter : Going Beyond Tweet Text. Arxiv.org. Retrieved September 23, 2017.

Saif, H., Fernandez, M., He, Y., Alani, H. (2013). Evaluation of Datasets for Twitter Analysis, A survey and a New Dataset, the STS-Gold. Emotion and Sentiment in Social and Expressive Media. 1096. Turin: CEUR Workshop Proceedings.

Thakkar, H., Patel, D. (2015). Approaches for Sentiment Analysis on Twitter: A State-of-Art study. Arxiv.org. Retrieved September 23, 2017.

Wang, N., Varghese, B., Donnelly, P. (2016). A Machine Learning Analysis of Twitter Sentiment to the Sandy Hook Shootings. Arxiv.org. Retrieved September 23, 2017.

Yang, C., Hsin-Yih Lin, K., Chen, H.-H. (2007). Emotion Classification Using Web Blog Corpora. 00, pp. 275-278. Web Intelligence, IEEE / WIC / ACM Inernational Conference on.