G02 MODERN DISASTER ANALYSIS APPLICATIONS OF SOCIAL MEDIA

Abstract

The aim of this project is to bring together diverse research communities such as information retrieval, data mining and natural language processing, to explore potential applications of Twitter data in disaster analysis and provide an assessment of potential implementation of social media data analytics in potentially contributing towards building an AI-based next-generation information processing systems for an effective utilization of social media data for disaster related business applications.

Introduction

Social media platforms (such as Twitter) are relied upon increasingly by the general public, businesses, and governments as a platform to retrieve real-time information in addition to disseminating information. In situations like natural disasters, social media information is often faster than official news sources. The massive amount of available data in social media also presents an opportunity to explore potential applications in disaster analysis.

The premise of this project was inspired by LexisNexis Risk Solutions, with the idea that if insurance information on natural disasters such as total loss, claim numbers, and risk assessment factors could be integrated with natural disaster statistics and potential implementations of social media data analytics, then a new method could be offered to the age-old methods of predicting and managing disaster risk.

Our aim is to perform an initial analysis to evaluate the usefulness of Twitter data in natural disaster assessment for potential business applications in insurance risk assessment.
.

Research Question(s)

- What useful metrics and trends can we extract from Twitter data about natural disasters?
- What is the feasibility of Twitter text analysis and publicly available severe weather data to generate disaster analysis?

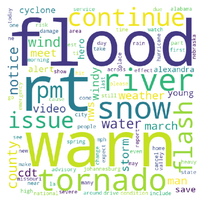
Materials and Methods

- Deployed web crawler using Tweepy to randomly collect tweets related to natural disasters
- Selected natural disaster keywords: hurricane, tornado, storm, flood, snow, rain, wind, earthquake, cyclone
- Tweets filtered for language: English, location: United States
- Utilized LexisNexis HPCC to load, merge and store 17.836MB of raw data
- Performed text analysis using Python NLTK on 3,392 relevant tweets on natural disasters
- Examined tweets against the collected severe weather data during same 2-month web crawling period

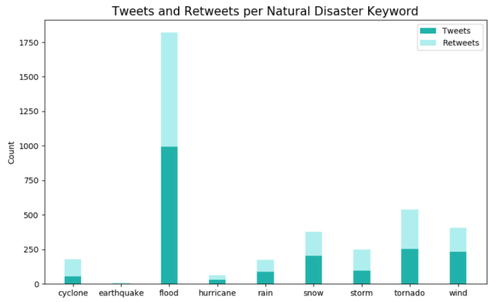
Results

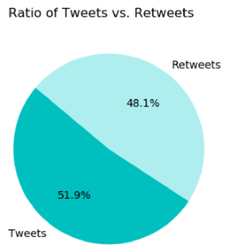
Total days tweets randomly crawled/collected: 26 days
Total tweets (uncleansed) collected: 68,571
Total tweets (cleansed) relevant to natural disasters: 3,392

Word Frequency Analysis result of most commonly occurring keywords:

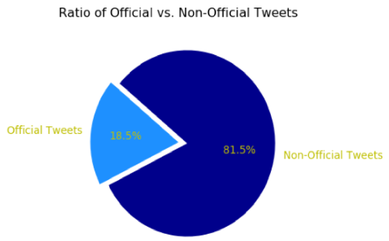


Analysis of tweets vs. retweets per natural disaster keyword showed that roughly half were retweets:

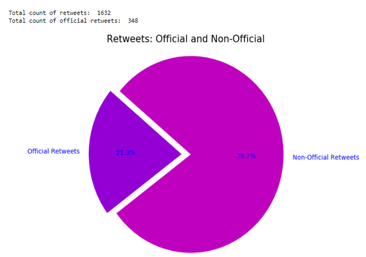




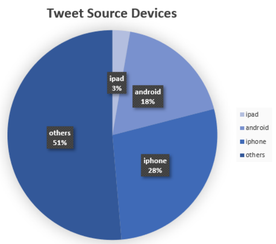
Analysis of official vs. non-official tweets showed that 81.5% were non-official:



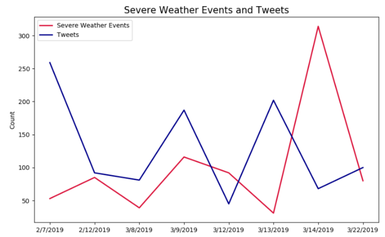
Are official tweets retweeted more? No



Analysis of tweet sources showed that more than half of tweets originated from non-mobile devices:



Is there any correlation between tweets volumes and natural disaster incidents? Inconclusive



Conclusions

This project examined a number of features in Tweets to extract useful metrics for potential natural disaster analysis applications. Word frequency analysis results show that in addition to the selected natural disaster keywords, the most commonly occurring tokens in BOW include lemmatized words such as warn, notice, emergency, damage, and alert, which may be of particular significance for insurance risk assessment applications.

Tweet vs. retweet breakdowns and official vs. non-official occurrences within retweets were some of the metrics evaluated to explore whether they could merit focus when collecting Tweets for future disaster analysis or risk predictions. In a similar pursuit, tweet sources were examined to determine the existence of any interesting insight for future collection of Twitter data based on whether natural disaster tweets are more likely to originate from "mobile" or stationary devices.

Finally, the severe weather events (i.e. natural disaster incidents) and tweet volumes were examined to uncover any positive or negative correlations that would support an assumption such as whether tweets volumes are likely to increase with severity of the natural disaster event, but the result was inconclusive.

As an initial analysis to evaluate the usefulness of social media such as Twitter in potential applications of disaster analysis, the most valuable findings are in the identification of the challenges and the limitations in the evaluation of the feasibility of disaster analysis applications of Twitter, and the suggestions for the direction of future works in finding potential business applications of Twitter analysis in insurance risk assessment.

Challenges for Future Works

- User-Disabled geo/location/coordinates feature
- Region-based filtering (eg- users with US accounts while in the US tweet about natural disaster in Africa)
- Content filtering (eg- non-emergency Tweets about selected keywords are within topic but not relevant for purpose at hand)
- Content analysis (eg- qualifying the intensity/severity of disaster event based on user-reported info)
- Polysemes
- Privacy policies and restrictions on collection of Twitter/social media data

Acknowledgments

- LexisNexis Risk Solutions
- Dr. Meng Han
- Dr. Lei Li
- Jennifer Watson, Mike Dunbar- The Weather Channel

Contact Information

Project Website: https://disastercapstone.com
Team Lead: Karis Kim (skim144@students.kennesaw.edu)
Lead Technical Specialist: Anthony Darden (adarde11@students.kennesaw.edu)
Advisor: Dr. Meng Han (mhan9@kennesaw.edu)
Dr. Lei Li (lli13@Kennesaw.edu)

KENNESAW STATE UNIVERSITY

Author(s) Karis Kim, Anthony Darden, Darlene Echols, Bricesen Ross
Advisors(s) Dr. Meng Han, Dr. Lei Li