Data Mining and Sentiment Analysis of Real-Time Twitter Messages for Monitoring and Predicting Events

Maya D. Albayrak and William Gray-Roncal
Johns Hopkins University Applied Physics Laboratory, {maya.albayrak, william.gray.roncal}@jhuapl.edu

Abstract

Real-time geolocation data can be used to map natural and social hazards, facilitating predictions and preventive actions. Currently, organizations such as the U.S Geological Survey, the U.S. Center for Disease Control, Weather.com and Google Trends attempt to provide information, but at a high cost and with limitations. This research represents an effort to overcome some of these challenges by offering an inexpensive alternative based on analyzed real-time data with precise geolocation, and including a dashboard informed by human factors analysis.

Twitter is a social networking service that allows access to messages and geolocations created by members of the public. For this research, data was collected from Twitter and filtered through data mining techniques based on keywords and phrases. To eliminate tweets with a context other than the focus of this research, sentiment analysis was performed using machine learning algorithms. Visual representations of the results were created including maps of precipitation and earthquakes, and a dashboard showing flu spread. Human factors analysis techniques, mouse and eye trackers, and a small survey-based study were used to verify that users could make accurate interpretations and conclusions.

Introduction

Technological advances that allow fast processing of large amounts of data have motivated researchers to look for real-time sources of information to predict events. Challenges for data collection are cost, public access, and accuracy. An innovative way to address these challenges is through social media data. Twitter is an online social networking service that lets its users communicate through short messages consisting of text, location (latitude and longitude), and time. The service created an Application Programming Interface (API) [1], that allows public access to tweets in real time. Research in social media analytics provides rapid identification of areas affected by environmental, health, and other social threats, and exciting progress has been made in developing research algorithms; however real-world adoption has been limited [2].

Process

In this work, we extracted tweets for three use cases: rain, earthquakes, and flu spread. Tweets were collected using the API, and then filtered using keywords. Data mining techniques were then used to filter messages and identify trends based on key phrases in addition to key words. Sentiment analysis was performed using machine learning software to eliminate tweets containing keywords used in a different context. We created a tool to visually represent the results and to aid in quick decision-making. We leveraged Plotly, a Python graphing library, to provide maps of rain and earthquakes, and a dashboard consisting of flu symptoms, flu spread, and graphs of related hashtags. To determine the efficiency of the dashboard, we used human factor techniques. For example, mouse and eye trackers were used to create heat maps, identifying areas where the users focused. In addition, test participants were surveyed to verify that the graphs contained information necessary to support accurate interpretation.

978-1-7281-1502-3/19/$31.00 ©2019 IEEE

Results

The results we obtained using these methods were compared to ground truth. For example, we compare to the results shared by Weather.com and the U.S Geological Survey, which use expensive satellites and seismometer data. The U.S. Center for Disease Control provides flu maps, but with a week delay (3). Google Trends (4) collects data in real-time, but results are based on raw data and mapped by state, creating imprecise reporting. Our work showed that we were able to effectively produce predictions that were consistent with these data sources, while also providing location specific information. The solution proposed by this project is an inexpensive method to obtain real-time data with precise geolocation and includes human factors methodology to help ensure that the researcher can interpret their results effectively. Our real-time weather data has been used to validate NASA satellite data through a citizen science project.

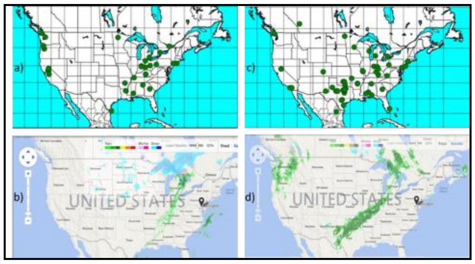

Figure 1. Comparison of precipitation maps created from tweets and weather.com maps. a), b) Event1: November 4th c), d) Event 2: November 11th, 2014.

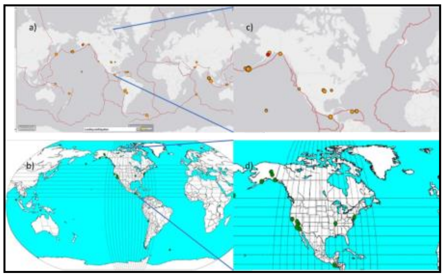

Figure 2. a) United States Geological Survey (USGS) map of earthquakes, b) Generated map, using Twitter data Nov. 8, 2015 c) and d) are the focused versions.

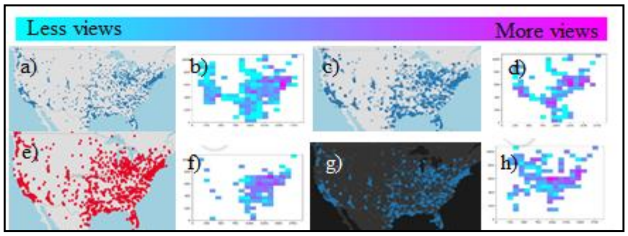

Figure 3. a), c), e), and g) are maps of tweets related to flu. b), d), f), and h) are heat maps generated from an eye tracker, purple represent focus areas. The analysis is paired as follows: (a,b) (c,d) (e,f) and (g,h).

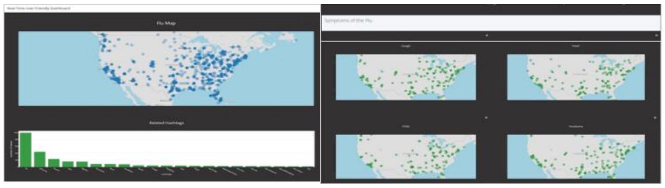

Figure 4: Dashboard developed through the Plotly library in Python, displaying a map of the flu, its symptoms, and related hashtags.

Future Work

We plan to extend this research beyond the initial dashboard reports to allow users to input their desired keywords and make calls directly to the Twitter API. This will provide the public with an easy interface to explore trends and predictions of diverse real-time information, facilitating the decision-making process in a variety of areas.

References

[1] L. E. Charles-Smith *et al.*, “Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review,” *PLoS One*, vol. 10, no. 10, pp. 1–20, 2015.
[2] Twitter API,” [Online], Available: https://developer.twitter.com/ [Accessed 16-Jan-2019]
[3] “CDC Flu Data.” [Online]. Available: https://www.cdc.gov/flu/weekly/usmap.htm. [Accessed: 16-Jan-2019].
[4]“Google Trends Data Collections,”[Online], Available: https://trends.google.com/trends/explore?q=flu&geo=AU [Accessed: 16-Jan-2019].

Acknowledgements

The authors would like to acknowledge Scott Simpkins, the Citizen Science for Earth Systems Program, NASA ROSES NNH16ZDA001N-CSESP, and AFCEA Central Maryland Chapter’s STEM Kickstarter grant program for their scientific support and contributions to this work.