Using Social Media Data to Improve Situational Awareness during Emergent Events

Emergent Research Forum (ERF)

Naga Vemprala
University of Texas at San Antonio
naga.vemprala@gmail.com

Charles Zhechao Liu
University of Texas at San Antonio
Charles.Liu@utsa.edu

Rachael Ruizhu Xiong
University of Texas at San Antonio
ruizhu.xiong@utsa.edu

Kim-Kwang Raymond Choo
University of Texas at San Antonio
raymond.choo@fulbrightmail.org

Abstract

Social media is a potential source of information that can be used to improve situational awareness during emergent events, such as hurricanes and riots. In this paper, we propose to take advantage of the rich information on social media and develop a set of machine learning and data analytics tools to dynamically learn from the ongoing supply of fragmented social media messages, continuously filter the unrelated information to reduce the noise contained in these messages, and efficiently convert scattered messages to convergent topics. This approach will allow us to quickly identify the emergence of the high-level discussion topics/themes/hot spots and monitor the ongoing development of these topics/themes/hot spots in a timely manner. Moreover, our proposed approach will focus on deriving patterns of events under development and probabilities of the occurrence of future events by examining the association between topics over continuous time intervals, and can be used to mitigate scam and rumor propagation. These applications allow more efficient utilization of resources in a timely manner, particularly during adverse situations such as natural disasters.

To test the outcome of this proposed approach, we used Twitter REST API to collect data (tweets) for the Charlottesville rally between August 10, 2017, two days prior to the violence, and August 25, 2017. In order to circumvent Twitter’s restriction of seven days of non-segmented analytics data, we developed our own scraping tool to collect archival information (e.g. user mentions, hashtags, URLs, comments, retweets, favorites and replies). We then performed data cleansing (removing stop words and special characters), segregated the tweets into documents (collection of tweets in hourly intervals) and bags of words based on the contextual meaning derived from the synonyms of the words (i.e. lemmatization), and applied the widely used machine learning classification algorithm, Latent Dirichlet Allocation (LDA), to classify the large volume of unstructured data (i.e. 795,075 tweets over the 15 days) into new topic categories. Preliminary results from the analysis of the Twitter data collected for the Charlottesville rally demonstrated the utility of our proposed approach.