Listening to Data: Interpreting Twitter Sentiment Analysis using Tone Analyzer and Personality Insights of PAG-ASA and Phivolcs

Christianne Lynnette Cabanban, Stephan Kupsch
Randy Joy M. Ventayen, Thelma D. Palaoag

University of the Cordilleras

ABSTRACT

Twitter, a microblogging site plays a vital role in spreading information during natural disasters. The volume of tweets posted during crisis and disaster tend to be extremely high, making it hard for disaster-affected communities and disaster management team of a local government unit to process the information in a timely manner. In this research, we describe different data mining techniques that can be used for extracting information from microblog posts that will be a basis of creating a machine learning called Disastweet: An Open-Source Tweet Mining Tool for Disaster Management. Specifically, we focus on extracting valuable information from tweets that is brief, self-contained relevant to disaster response.

Keywords: NA
Classification: K.8.1, K.4.2
Language: English

London Journals Press

LJP Copyright ID: 975711
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science and Technology

1 9 2 U K

Volume 17 | Issue 2 | Compilation 1.0

Listening to Data: Interpreting Twitter Sentiment Analysis using Tone Analyzer and Personality Insights of PAG-ASA and Phivolcs

Stephan Kupsch[α], Randy Joy M. Ventayen[σ], Christianne Lynnette G. Cabanban[ρ] & Thelma D. Palaoag[¥]

I. Abstract

Twitter, a microblogging site plays a vital role in spreading information during natural disasters. The volume of tweets posted during crisis and disaster tend to be extremely high, making it hard for disaster-affected communities and disaster management team of a local government unit to process the information in a timely manner. In this research, we describe different data mining techniques that can be used for extracting information from microblog posts that will be a basis of creating a machine learning called Disastweet: An Open-Source Tweet Mining Tool for Disaster Management. Specifically, we focus on extracting valuable information from tweets that is brief, self-contained relevant to disaster response.

Author α σ ρ ¥: University of the Cordilleras, Baguio City, Philippines.

II. INTRODUCTION

Social media networks and microblogging services such as Twitter has redefined the floodgates of information dissemination through pervasiveness of Information and Communication Technology (ICT).

Members of the society turn to social networking sites, microblogging services and similar technologies to better understand and communicate during emergency situations by focusing on Twitter communications also called as tweets generated during mass emergency, and eventually showcasing.

Natural Language Processing (NLP) techniques to contribute to the task of analyzing through massive datasets infeasible for the emergency-management community, including the medical and public health professionals, to effectively find it, make sense of, and act on it. By merely sharing images, status updates and tweeting, the members of the public is already becoming part of a larger response network rather than a mere bystanders or casualties.

Typhoon Haiyan hit the central Philippines on November 8, 2013, killing 6,190 people and leaving 14.1 million people in need of immediate assistance. Over four million people were forced from their homes with more than a million houses destroyed or damaged. Many of the people who were displaced were already amongst the poorest in the Philippines and following the typhoon found themselves living in tents or evacuation centres.[1]

When disaster events capture global attention users of Twitter form transient interest communities that disseminate information and other messages online. Twitter data were collected the day before super typhoon Haiyan (locally known as Yolanda) and for 18 days afterwards. Haiyan was a trending topic on Twitter for over two weeks, with activity coming from many countries.

London Journal of Research in Computer Science and Technology

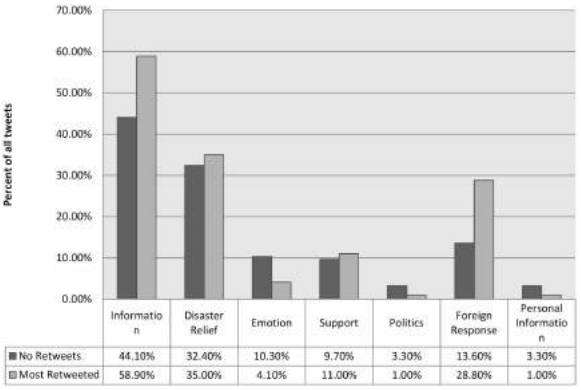

Figure 1: Content distribution of tweets during the Typhoon Haiyan

Figure 1 shows general content category distributions for both the most retweeted messages and messages without retweets. While 59% of the most retweeted messages contain information about the typhoon, a smaller portion (43%) of the general Twitter messages contains information. Results show that these tend to contain more messages with emotion, personal information, and commentary about politics, although these do not occur frequently. Among the original tweets, 10% of messages have content about emotions, compared to a smaller 4% among the most retweeted messages. Personal information in general does not get tweeted often, however given the scale of Twitter, even small proportions could be consequential. There are more instances of personal information tweets from original posts compared to the retweeted messages. Of note is a dearth of messages related to politics. [2]

In this research, we describe different algorithms that can be used for extracting information from microblog posts using Apache Spark on IBM Bluemix. We use the twitter handle of Department of Science and Technology - PAG-ASA and Phivolcs as a tool in assessing sentiment analysis of its' followers.

III. METHODOLOGY

The focus on creating a machine learning app that uses Spark Streaming is to create a feed that captures live tweets from Twitter. The user can optionally filter the tweets that contain the hashtag(s) of their choice. The tweet data is enriched in real time with various sentiment scores provided by the Watson Tone Analyzer service.

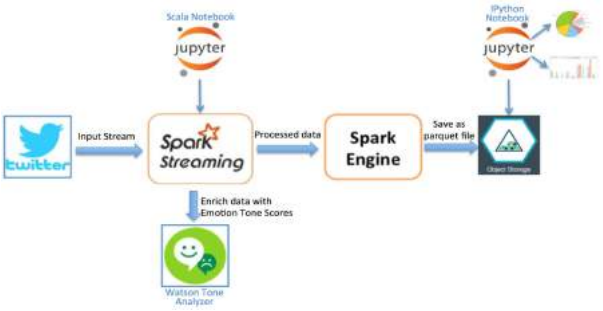

Figure 2: Basic Architecture of Spark Streaming

This service provides insight into sentiment and then use Spark SQL to load the data into a DataFrame for further analysis as seen in Figure 2.

The three algorithms utilized are Naïve Bayes, Logistic Regression and Decision Trees in this research.

3.1 Naive Bayes

Naive Bayes is a simple multi-class classification algorithm based on the application of Bayes' theorem. Each instance of the problem is represented as a feature vector, and it is assumed that the value of each feature is independent of the value of any other feature. One of the advantages of this algorithm is that it can be trained very efficiently as it needs only a single pass to the training data.

3.2 Logistic Regression

Logistic regression is a regression model where the dependent variable can take one out of a fixed number of values. It utilizes a logistic function to measure the relationship between the instance class, and the features extracted from the input.

3.3 Decision Trees

The decision tree is a classification algorithm that is based on a tree structure whose leaves

represent class labels while branches represent combinations of features that result in the aforementioned classes. Essentially, it executes a recursive binary partitioning of the feature space.

IV. RESULTS AND DISCUSSION

The tweets are analyzed using Watson Personality Insights as deemed in Figure 3 that requires at least 100 words from its lexicon to be available, which may not exist for each user. This is why the getPersonlityInsight helper function guards against exceptions from calling Watson PI. If an exception occurs, then an empty array is returned. Each record with empty array is filtered out of the resulting RDD.

userid	Emotional range	Agreeableness	Extraversion	Conscientiousness	Openness
phivolcs_dost	0.4298609	0.7124939	0.53390217	0.59583825	0.7434547
dost_pagasa	0.37862363	0.71802986	0.5235555	0.59549797	0.7920747

Figure 3: Watson Personality Insight Results

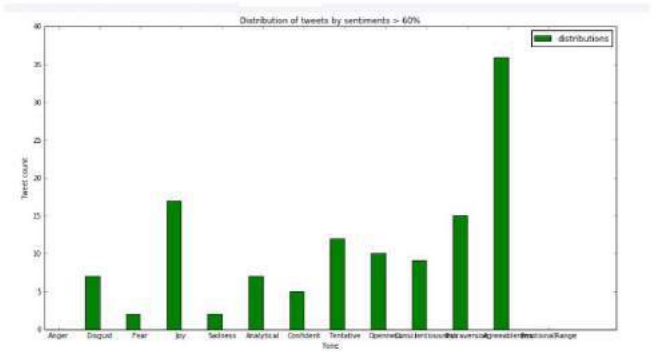

Figure 4: Real time visualization

V. CONCLUSION

Social Media Networks specifically Twitter have emerged as an important source of information. These sources may not be primarily used in prediction of the disasters, hence it contributed significantly to early detection and adaption for appropriate disaster management response. The researcher propose to build a machine learning based on these concepts for Phase 1. Our aim for the phase II is to extract meaningful information from tweets during natural disasters. We can use machine learning tool that the researcher will be developing to extract valuable information from noisy social media data.

REFERENCES

1. (n.d.). Typhoon Haiyan: Community research into the relocation... - ReliefWeb. Retrieved March 30, 2017, from http://reliefweb.int/sites/reliefweb.int/files/resources/cs-typhoon-haiyan-internally-displaced-people-relocation-150515-en.pdf
2. (2016, March 28). Tweeting Supertyphoon Haiyan: Evolving Functions of Twitter... - PLOS
3. Retrieved March 30, 2017, from http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0150190&type=printable
4. Extracting Information Nuggets from Disaster- Related Messages in...." http://chato.cl/papers/imran_elbassuoni_castillo_diaz_meier_2013_extracting_information_nuggets_disasters.pdf. Accessed 2 Apr. 2017.

This page is intentionally left blank