Data Analysis for

#ebola occurrences on Twitter

## **Introduction**

A lot of buzz is happening around Ebola outbreak, its transmission, symptoms, risks and preventions. Although a majority of the outbreak is in West Africa, a handful of cases were reported and confirmed in the United States Lots of discussions are going around in news debates, government’s response in handling the situation and its action to fight the epidemic. While Ebola has been contained in its entirety in United States, it has created a lot of stir among the people with discussions happening all over. Social media is no exception to the topic. All major social media outlets has shown enormous response to the topic in different ways and the increasing number of cases created more interest among non-participants of social media.

The fact that the virus could spread at a faster rate has created lot of confusion in general public and the Center for Disease control and prevention has put forward detailed information on facts and myths about Ebola and its prevention. Considering all the discussions around, a predictive analysis study is conducted on twitter data by Social Media Analytics Command Center (SMACC) at Illinois State University using data retrieved over a period of 11 days, to understand the trends and discussion patterns on Ebola along with other discussion topics happening in parallel.

## **The Data**

Data is retrieved from NUVI for a period of 11 days, from Sep 14th 2014 to Sep 25th 2014. Data contains the actual 140 character tweet along with the Author name, Time, profile name, location information, if available, sentiment pattern of the post as determined by NUVI, number of users the tweet might reach (which depends on followers for a user), influence, type of social network, etc.

Data is retrieved as a comma separated values file and all the values are segregated based on the type of data, with the twitter feed being the largest chunk of data. URL to profiles and tweets are also captured.

## **Mining**

Various Data mining operators are used to identify patterns using Rapidminer, a software platform that provides an integrated environment for data mining, text mining and predictive analysis.

While majority of the data retrieved is structured, tweets from various users are unstructured and has a variety of information to be considered. Unstructured tweet text data is separated from structured data and is processed in the form of a text file to feed as an input for text mining.

Text mining in RapidMiner uses a variety of operators to retrieve words through which a pattern can be observed. Data is tokenized and is filtered to only keep tokens that have 3 or more words. This will essentially eliminate all the prepositions. Filters are then applied to retrieve tokens that start with ‘#’, essentially, hashtags. Running the process on all the available text data gives us the total number of hashtags and their count. By end of text mining, hash tags and their occurrences are retrieved and top 10 hashtags are considered for analysis.

Through data mining, counts for various hashtags are retrieved and top 10 are chosen based on the number of occurrences.

## **The Breakdown**

#ebola is at the top of the list with 29796 occurrences among a total of 30754 tweets retrieved, which is a staggering 97% of tweets, followed by #ebolaoutbreak with 1599 occurrences, closely followed by #liberia with 1292 occurrences and #africa 1064 occurrences. The vast difference in the count of hashtag’s between first and the rest shows us that a majority of discussion has happened around Ebola.

More than half of the posting activity to #ebola is retweets.

Out of 30,754 total tweets, 15,932 (about 53%) of them were retweets.

The remaining 13,933 were original tweets – 7,390 weren’t directed to anyone, 1,789 replied directly to another user, and 4,754 publicly mentioned one or more users.

In total, 13,615 user account data is retrieved, of which 13,312 (about 98%) posted to #ebola in the given dates. Excluding retweets, that number drops down to 5,569 users.

A large amount of tweets have posted various links to pictures, webpages and videos etc. 21,590 tweets for #ebola among the 29,865 tweets have links to various sources apart from text.

## **Evaluating Sentiment**

Sentiment Analysis of the tweets are done at NUVI level and data retrieved has sentiment values computed, Positive, Negative or Neutral. Sentiment depicts the emotional aspect of the words used in the posts. It is typically determined by identifying number of positive, negative and neutral words that denote a meaning and analyzing what words dominate the text. As with any other analysis, this may not completely represent the intent of the user and depends on the assumption that there are more positive words than negative and neutral in the post.

With the help of available hash tags and sentiment column, the total number of Positive, negative and neutral posts for a hash tag are calculated to identify the overall sentiment around a hash tag.

It turns out that the overall sentiment around #ebola is neutral, with 13,921 tweets, followed closely by 9,481 negative tweets. We can say that the sentiment around #ebola is more neutral than negative, owing to the not so much difference in counts.

Similarly, #ebolaoutbreak and the rest of the hashtags except for #isis are also more neutral than negative to positive. The reaction of users on social media about all the events going on at the time of retrieving data is more Neutral. Interestingly, after Neutral, the highest percentage of tweets are more negative.

This gives an interesting pattern, out of the top ten hashtag counts retrieved, 7 hashtags has a relation to #ebola and #africa. We could say that most of the discussion is surrounded by #ebola and #africa related issues.

## **Location**

Users have tweeted from various geographic locations within United States. Location information is captured for majority of tweets wherever available. A quick run on the location data gives us the details as shown below

Majority of tweets were from New York State, followed by California and Kansas. It is interesting to know that California and Kansas tweeted more even though there are no confirmed ebola cases in those states.