Paper- 3330-2015

Analyzing and visualizing the sentiment of the Ebola outbreak via tweets

Sharat Dwibhasi, Dheeraj Jami, Shivkanth Lanka, Goutam Chakraborty

Oklahoma State University

ABSTRACT

The Ebola virus outbreak has produced some of the most significant and fastest trending news throughout the globe today. There has been a lot of buzz surrounding the deadly disease and the drastic consequences that it potentially poses to mankind.

Social media provides the basic platforms for millions of people to discuss the issue, and allows them to openly voice their opinions. There has been a significant increase in the magnitude of responses all over the world since the death of an Ebola patient in a Dallas, Texas hospital. In this paper, we aim to analyze the overall sentiment that is prevailing in the world of media. For this, we extracted the live streaming data from Twitter over time and studied its pattern based on the Ebola Timeline. We used SAS® Enterprise Miner and SAS® Sentiment Analysis Studio to evaluate key questions pertaining to the outbreak such as, how seriously are people taking the outbreak, the geographical areas where people are most concerned, percentage of tweets which emphasize awareness and how the of people regarding Ebola outbreak changed over a period of time.

INTRODUCTION

Microblogging today evolved as a basic medium of communication among the Internet users worldwide. Millions of messages are being sent out daily in the social media websites such as Facebook, YouTube and Twitter. These massive data corpus have created a platform for Social Media Analytics which can be efficiently used for marketing and sentiment analysis. Spurred by these opportunities, companies such as Twitalyzer, Brandwatch and several others are offering services of Twitter sentiment analysis.

In this paper, we have used tweets to study the sentiment of the people regarding the #Ebola outbreak and model public mood and emotion. Ebola is a disease which has a high risk of death, killing between 25 percent and 90 percent of those infected with an average of about 50 percent [1]. The disease was first registered in 1976 in two simultaneous outbreaks, one in Nzara, and the other in Yambuku. The latter one occurred in a village near the Ebola River where the disease takes its name.[2] Through 2013, the World Health Organization reported a total of 1,716 cases in 24 outbreaks.[1] The largest outbreak to-date is the ongoing epidemic in West Africa, which is centered in Guinea, Sierra Leone and Liberia. As of Mar. 25th 2015, this outbreak has 24,927 reported cases resulting in 10,338 deaths [3]. This has catalyzed an enormous response worldwide on Twitter.

It has been observed that the tweets posted can be categorized into two bins: tweets containing opinions of the user and tweets to share information (retweeted from various news broadcasting centers). Both these types of tweets reflect back the mood state of the author. In the first case, expressions are evident from explicit sharing of subjectivity [4]. For example, “An Ebola vaccine is on the horizon”. In the second case, even though the user is not explicitly tweeting about his/her personal emotion but the tweets reflect their mood over the issue. For example, “Very Good News, Sierra is now Ebola free.” Thus, tweets can be used to represent sentiment and when collected and analyzed over a given time period can reveal the changes in the state of public sentiments on a larger scale.

In this paper, we analyze and visualize the sentiment patterns as evidenced from the tweets published during Oct. 8th 2014 to Feb 15th 2015 and relate them to the recent happenings about Ebola over the same time period. On the basis of a large database of tweets downloaded, we specifically look at the interplay between a) death of patient, such as the death of the Ebola patient Thomas Eric Duncan and the treatment of Amber Vinson and b) the steps taken by U.S government to tackle the outbreak.

PROCEDURE

We first extracted data from Twitter by accessing the live stream API of Twitter, using the tweepy package in Python. But in this process of collecting tweets, we had to make sure that the crawler runs for the entire day. In this case, as we had to download data for around six months it impractical. Also, there were some instances when the crawler abruptly stopped and the data gathering process halted for that period of time which resulted in the irrecoverable loss of tweets as we were dealing with the live streaming API. As it resulted in a non-continuous data, it necessitated the need of a better data extraction process which could run interminably. It was a project within itself and we used Google Spreadsheets for this task. Using the script editor in Google Spreadsheet we wrote a crawler that runs continuously and stores data on the cloud which ensures that there is no loss of a single tweet for a given hashtag.

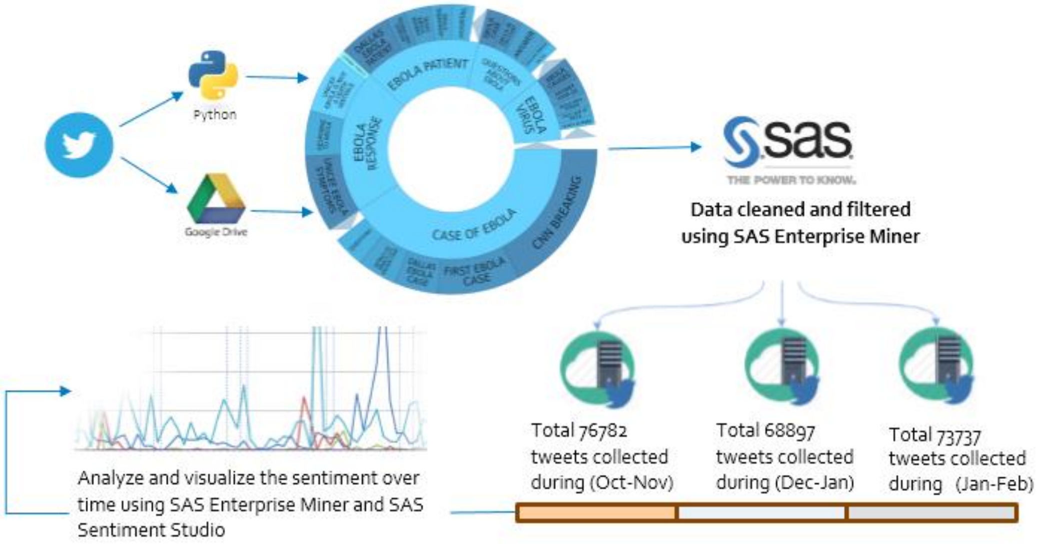

Python
Google Drive
EBOLA PATIENT
EBOLA VIRUS
EBOLA RESPONSE
CASE OF EBOLA
CNN BREAKING
FIRST EBOLA CASE
DALLAS EBOLA CASE
sas
THE POWER TO KNOW.
Data cleaned and filtered using SAS Enterprise Miner
Total 76782 tweets collected during (Oct-Nov)
Total 68897 tweets collected during (Dec-Jan)
Total 73737 tweets collected during (Jan-Feb)
Analyze and visualize the sentiment over time using SAS Enterprise Miner and SAS Sentiment Studio

Figure 1 Execution Plan of this project

The metadata of the datasets used is as shown below:

Variable Name	Type	Format	Length	Details
Date	Date	DATE9	8	Date on which the tweet was posted
Time	Time	Time5	8	Time at which the tweet was posted (24Hour Format)
Twitter User	String	Char20	20	Unique username of the tweeter
Followers	Number	BEST12	8	No of followers for that tweeter
Follows	Number	BEST12	8	To number of fellow tweeters he/she follows
Retweets	Number	BEST12	8	Number of times his tweet has been re-tweeted
Favorites	Number	BEST12	8	Total number of likes for that tweet
Tweet Text	String	Char182	182	The actual text of the tweet (max.140 characters)
Lang	String	Char3	3	The original language code in which the tweet was posted
Source	String	Char135	135	The source of the tweet
Geo	String	Char54	54	The geographical location of the tweet
Location	String	Char150	150	The location of the tweet
User Dis	String	Char160	160	User ids of the tweeter
Created At	Date	Date9	8	Date on which the twitter account has been created
Status Count	Number	BEST12	8	The total number of tweets posted till date

Table 1 Metadata of the final dataset downloaded by the crawler

APPROACH

A timeline of important clinical, political and social activities occurred and widely covered by the media relating to Ebola between Oct. 8th 2014 and Mar. 15th 2015. For this project we have the Ebola timeline as shown in *Figure 2.*

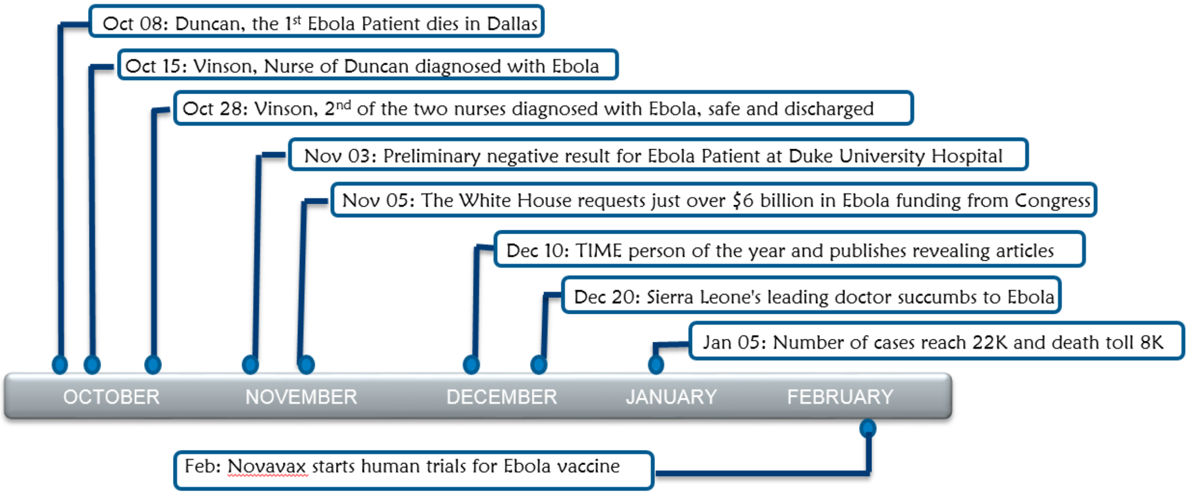

Oct 08: Duncan, the 1st Ebola Patient dies in Dallas
Oct 15: Vinson, Nurse of Duncan diagnosed with Ebola
Oct 28: Vinson, 2nd of the two nurses diagnosed with Ebola, safe and discharged
Nov 03: Preliminary negative result for Ebola Patient at Duke University Hospital
Nov 05: The White House requests just over $6 billion in Ebola funding from Congress
Dec 10: TIME person of the year and publishes revealing articles
Dec 20: Sierra Leone's leading doctor succumbs to Ebola
Jan 05: Number of cases reach 22K and death toll 8K
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
Feb: Novavax starts human trials for Ebola vaccine

Figure 2 Ebola timeline considered for this project

A corpus of around 270,000 tweets in English language published by Twitter users in the same time period and distributed as shown in *Figure 3* were considered for this project.

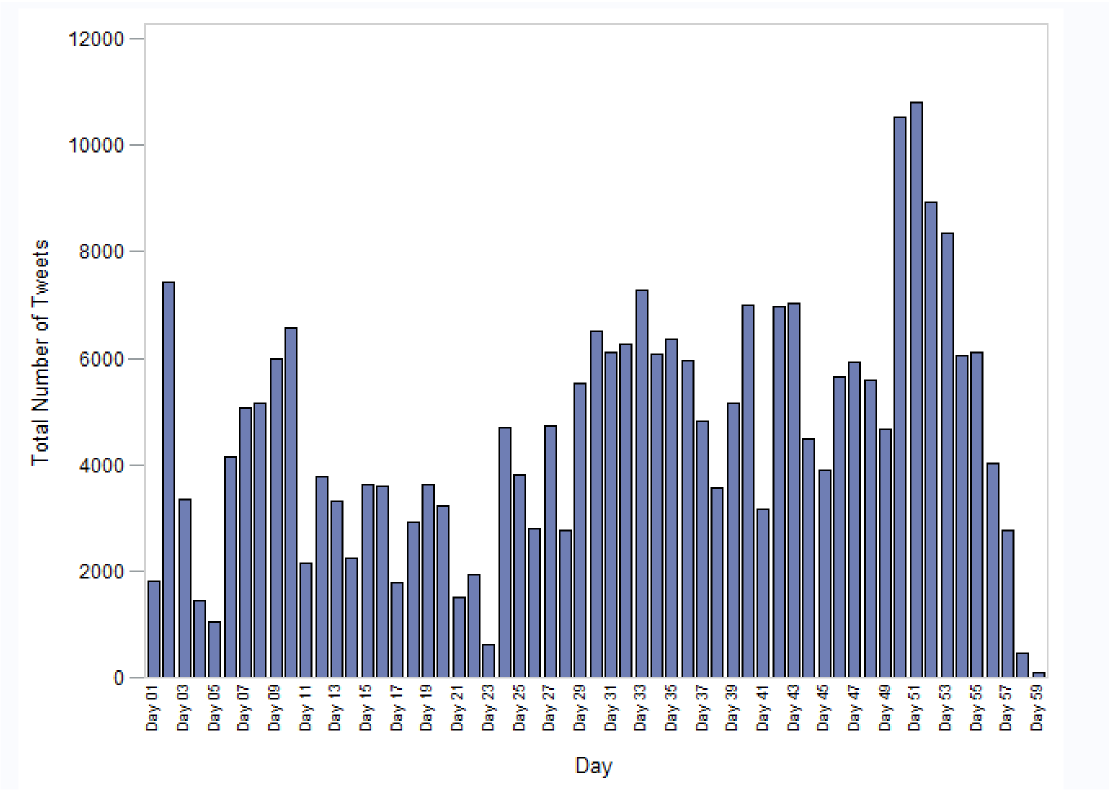


Figure 3 Volume of tweets: 8th Oct 2014 to Feb 3rd 2015

The entire tweet corpus has been divided into three datasets for easy categorization of the tweets and comparing the change in moods. The first dataset contains tweets starting from Oct. 8th 2014 to Dec. 26th 2014. The second dataset contains tweets starting from Jan. 2nd 2015 to Jan. 20th 2015 and the third dataset contains tweets from Jan. 21st 2015 to Feb. 3rd 2015. For simplicity and software limitations, we have considered only tweets in English language and made sure that each dataset contains approximately equal numbers of tweets and are evenly distributed.

It can be assumed that the retweet count signifies the interest of the users in that particular topic and a distribution of the same can help in an easy visualization of variation. The below needle plots depict the distribution of retweets over time in the three datasets (shown by the black partition line):

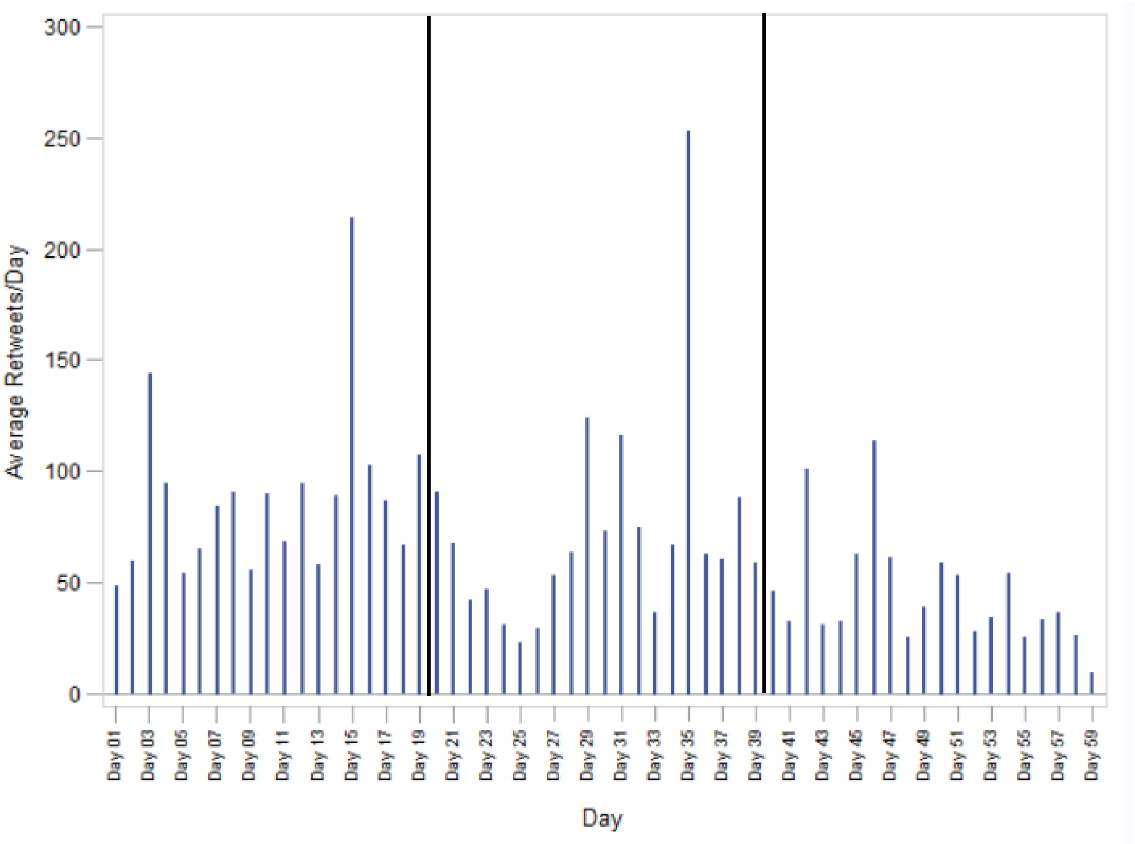


Figure 4 Average retweets per day from 8th Oct 2014 to Feb 3rd 2015 along with the partition of datasets

DATA PROCESSING USING SAS® ENTERPRISE MINER:

For our analysis, we considered all the tweets in the above mentioned timeline and followed the general text analytic approach suggested by Chakraborty, Pagolu and Garla (2013). This involved using appropriate NLP techniques, lemmatization, concept linking, use of synonyms, etc.

Concept links help in understanding the relationship between words (terms) based on the co-occurrence of words (terms) in the documents. It is a hyperbolic tree graph with Ebola (in this case) in the center of the tree structure. It shows the terms that are strongly associated with the term Ebola [5].Thick links indicate strong association between the terms. The *Figure 5* below shows the concept links of the first dataset

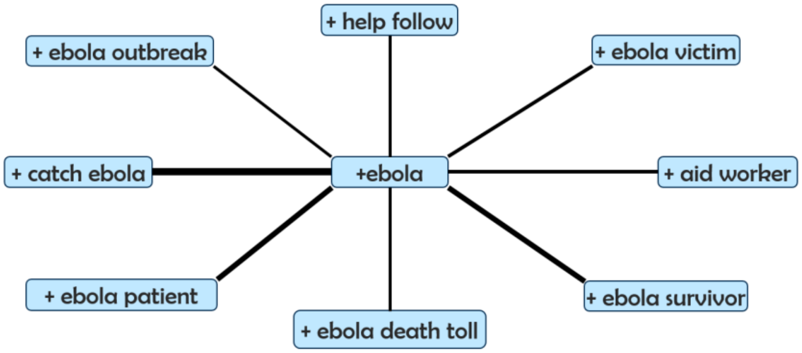


Figure 5 Concept links for dataset 1

In order to create the concept links we have treated some of the keywords as synonyms such as "*sufferer*", "*suffering*", "*Ebola sufferer*", "*Ebola patient*" and "*Ebola victim*" were all treated as "*Ebola sufferer*". Similarly in the concept link for the second dataset "*Sierra*", "*Leone*" and "*Africa*" are treated as "*Africa*" itself.

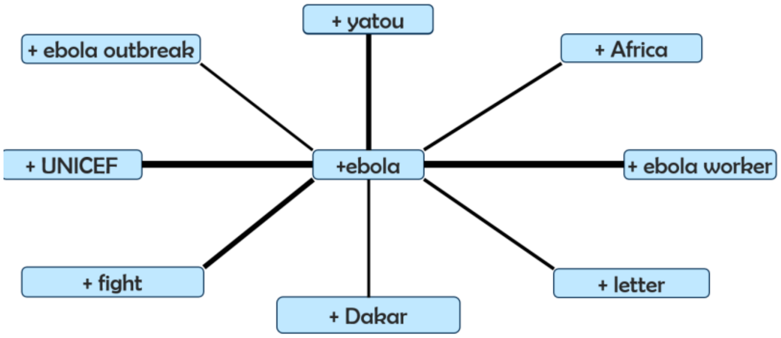


Figure 6 Concept links for dataset 2

There is a clear difference in the concept links of these datasets. In the first case, people were more concerned or rather taken aback by the number of deaths caused by Ebola, whereas in the second case people appreciated the Ebola workers, who are risking their life and saving the world from a drastic epidemic. Yatou and Dakar are the patients who were treated by the Ebola Workers and they wrote thanking letters to the Ebola Workers which was shared by the users all over the world as a mark of appreciation and extension of support to the Ebola workers. Kudos!. In the third dataset, concept links tells us that people tweeted more about the health facilities provided at Ebola prone areas and specifically they were satisfied by the Ebola Labs built specifically for the women which is highlighted by the terms +all woman, +ebola lab and +scale recovery.

The concept link for dataset 3 is as shown below:

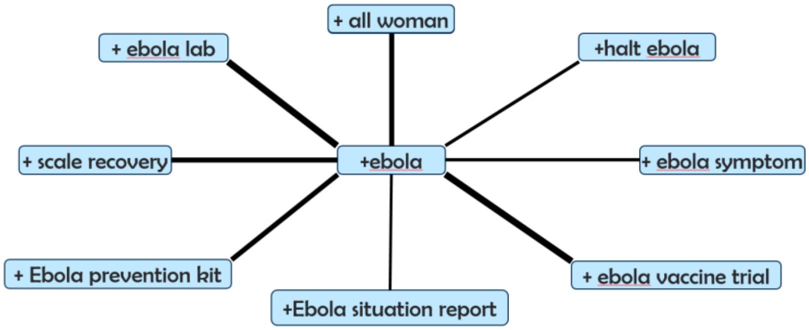


Figure 7 Concept links for dataset 3

Going further, to better understand the pattern in the tweets, we have used customized topics. We created 6 customized topics using following terms in the topics as shown in the table below:

Topic	Terms
Africa	Leone, Africa, West, Sierra
Prevention	Lockdown, Hospital, Symptom, Health, Prevention, Cure, Control, Vaccine, Treatment, Quarantine, Airport, Screening
Ebola Patients	First Person, Patient, Family, Friend, Dead
Concern Over Ebola	Tweet, Minute, Check, Christmas, Christmas Day, Ebola Spread, Outbreak, Fight, Obama , Congress, Fund
Ebola Workers	Health Worker, Fight, Help, Donate, Zuckerberg, Organization, Good, UNICEF
Ebola Death	Population, Population Control, Form, Shit, Death Toll, Case People, Die, American, Kill virus, Spread

Table 2 Table depicting the categorization of Terms into six topics

When we examine the distribution of these tweets over the six different topics for each dataset, we can clearly see how the sentiments/opinions change over a period of time.

In the first dataset, people were worried more about themselves/their family members getting infected from Ebola. But over a period of time people started to care about Africa/Sierra/Leone where the number of Ebola cases were predominant. Later everyone including Queen Elizabeth started appreciating the selflessness of Ebola workers who risked their life in protecting the victims and also showed interest in a permanent cure (vaccine) of this deadly disease.

The below screenshots depicts the distribution of the tweets over these topics for all the datasets as shown (datasets are color coded):

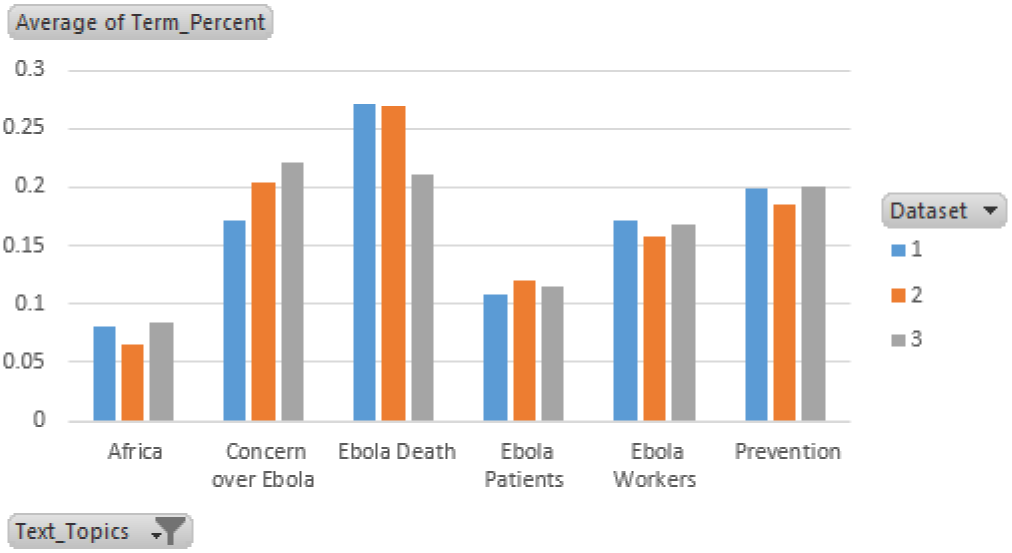


Figure 8 Number of documents by topic ID using clustered approach for all the datasets

In the second case, the topics tell us a different story. Over the period of December end to January first week, the entire focus of the world was changed from Dallas to Africa. Although the first US Ebola victim succumbed, it followed several preventive measures and people were treated successfully. This brought in positive hope and energy among the people worldwide. Media glorified the backend Ebola workers. TIME magazine mentioned them as the Man of the year and Twitter was flooded with positive tweets. Where as in the third dataset, people tweeted more about providing health facilities to the Ebola patients all over the world. Also we can see that there was a major increase in the percentage of topic ids for the term "Concern over Ebola".

BUILDING A RULE BASED MODEL USING SAS® ENTERPRISE MINER:

So far we have gone through the descriptive analysis of the tweets and studied the concept links and the clustering of tweets. Now to get an idea of sentiment variation regarding Ebola, we built a rule-based model to predict the sentiment of the tweets into two bins: positive and negative.

For this, we took a random sample of tweets from the entire dataset containing around 1,400 tweets and three different reviewers read through each of the tweets diligently, and coded them as either positive/negative. We then compared each of these tweets and considered only those tweets for modeling which were coded undisputedly as positive/negative by each of the reviewer.

The below bar chart shows the percent Positive/Negative tweets collectively coded by the reviewers.

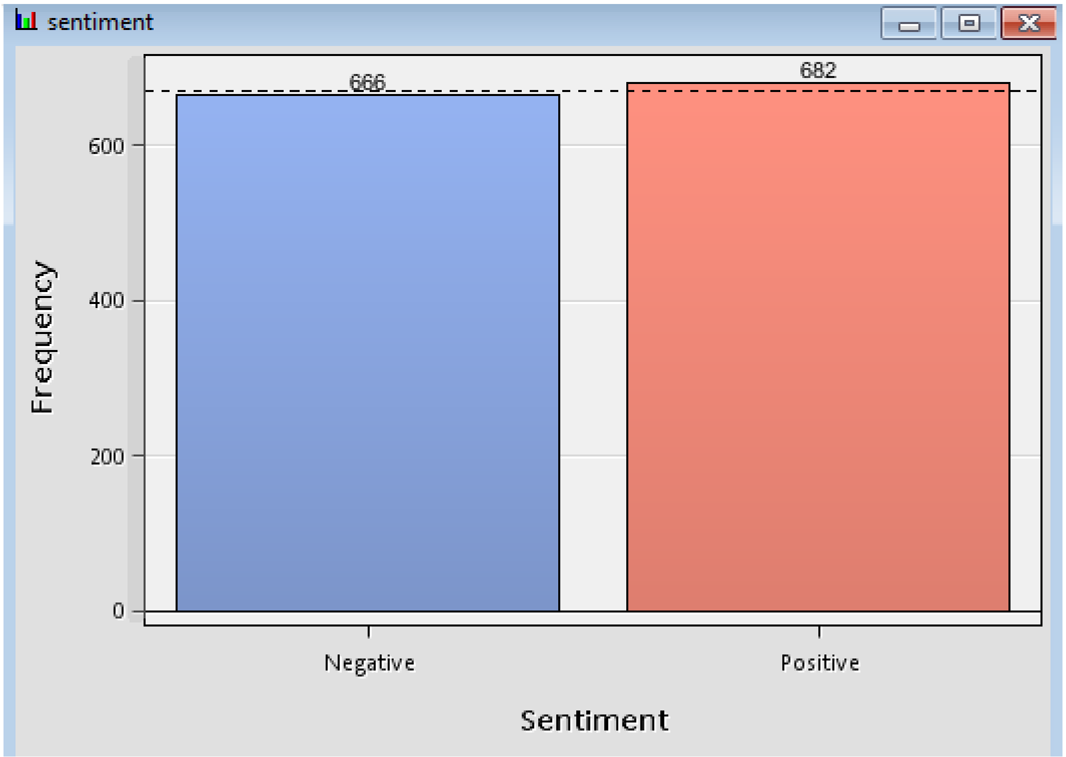


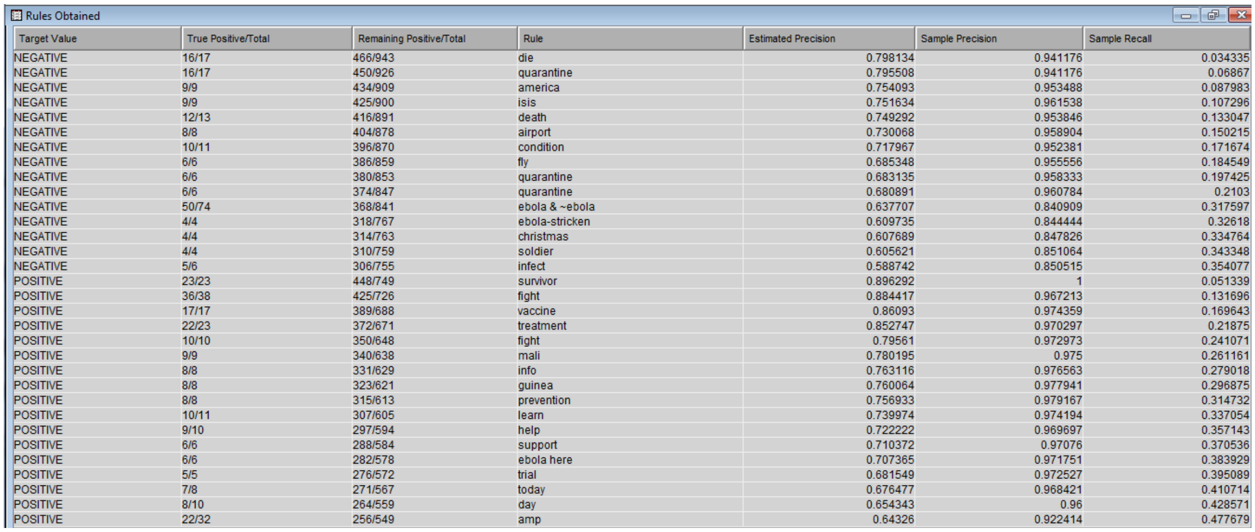
Figure 9 Distribution of positive and negative tweets in the training dataset

Overall, there are 666 negative tweets and 682 positive tweets in the training dataset used for modeling.

We then used a data partition node which split 70% of the data as the training dataset and 30% of the data as validation. It was followed by a text-parsing node creates the term-by-document matrix.[6] Generally terms are single words considered along with their synonyms/stems, multi word phrases,, parts of speech etc. It was followed by the text filter node which helped us to eliminate extraneous variation caused by the presence of noise terms and other terms that were not pertinent to our analysis. For running this node we also provided a complete English dictionary containing terms and synonyms. The end result of this node was a compact yet rich information contained in the tweets.

This was followed by the “Text Rule Builder” node which generates an ordered set of rules from small subsets of terms that together are useful in describing and predicting our target variable – sentiment of tweets (positive/negative). Rules are defined for each target category containing a conjunction that indicates the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)). A particular document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3. [6]

This set of derived rules creates a model that is both descriptive and predictive. When categorizing a new document, the model will proceed through the ordered set and choose the target that is associated with the first rule that matches that document. The rules given by this node are as shown below:



Rules Obtained

Target Value	True Positive/Total	Remaining Positive/Total	Rule	Estimated Precision	Sample Precision	Sample Recall
NEGATIVE	16/17	466/943	die	0.798134	0.941176	0.034335
NEGATIVE	16/17	450/926	quarantine	0.795508	0.941176	0.06867
NEGATIVE	9/9	434/909	america	0.754093	0.953488	0.087983
NEGATIVE	9/9	425/900	isis	0.751634	0.961538	0.107296
NEGATIVE	12/13	416/891	death	0.749292	0.953846	0.133047
NEGATIVE	8/8	404/878	airport	0.730068	0.958904	0.150215
NEGATIVE	10/11	396/870	condition	0.717967	0.952381	0.171674
NEGATIVE	6/6	386/859	fly	0.685348	0.955556	0.184549
NEGATIVE	6/6	380/853	quarantine	0.683135	0.958333	0.197425
NEGATIVE	6/6	374/847	quarantine	0.680891	0.960784	0.2103
NEGATIVE	50/74	368/841	ebola & ~ebola	0.637707	0.840909	0.317597
NEGATIVE	4/4	318/767	ebola-stricken	0.609735	0.844444	0.32618
NEGATIVE	4/4	314/763	christmas	0.607689	0.847826	0.334764
NEGATIVE	4/4	310/759	soldier	0.605621	0.851064	0.343348
NEGATIVE	5/6	306/755	infect	0.588742	0.850515	0.354077
POSITIVE	23/23	448/749	survivor	0.896292	1	0.051339
POSITIVE	36/38	425/726	fight	0.884417	0.967213	0.131696
POSITIVE	17/17	389/688	vaccine	0.86093	0.974359	0.169643
POSITIVE	22/23	372/671	treatment	0.852747	0.970297	0.21875
POSITIVE	10/10	350/648	fight	0.79561	0.972973	0.241071
POSITIVE	9/9	340/638	mali	0.780195	0.975	0.261161
POSITIVE	8/8	331/629	info	0.763116	0.976563	0.279018
POSITIVE	8/8	323/621	guinea	0.760064	0.977941	0.296875
POSITIVE	8/8	315/613	prevention	0.756933	0.979167	0.314732
POSITIVE	10/11	307/605	learn	0.739974	0.974194	0.337054
POSITIVE	9/10	297/594	help	0.722222	0.969697	0.357143
POSITIVE	6/6	288/584	support	0.710372	0.97076	0.370536
POSITIVE	6/6	282/578	ebola here	0.707365	0.971751	0.383929
POSITIVE	5/5	276/572	trial	0.681549	0.972527	0.395089
POSITIVE	7/8	271/567	today	0.676477	0.968421	0.410714
POSITIVE	8/10	264/559	day	0.654343	0.96	0.428571
POSITIVE	22/32	256/549	amp	0.64326	0.922414	0.477679

Figure 10 Rules obtained from the text rule builder model

The graph below shows the cumulative lift for graph for the training and validation datasets suggesting that they are pretty much close to each other.

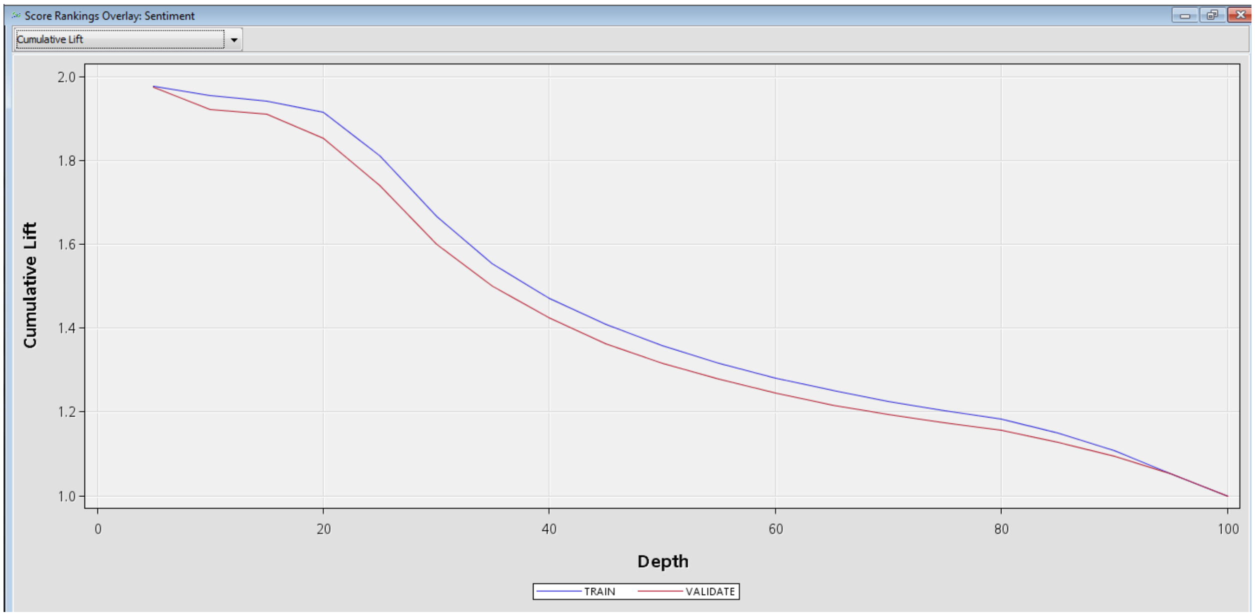


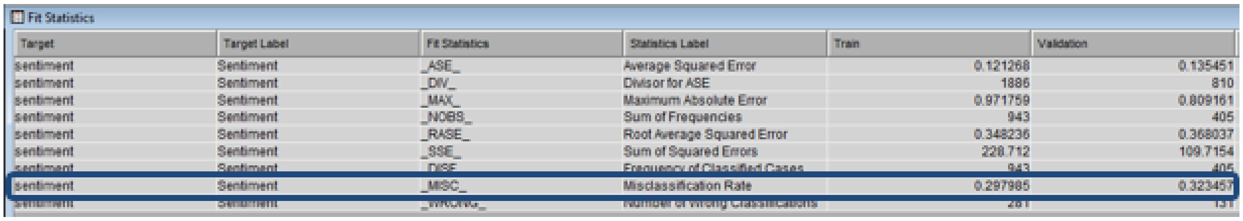
Figure 11 Cumulative lift of the training and the validation datasets.

There are serious challenges in categorizing tweets for sentiments about an emerging issue such as Ebola because of the issues mentioned below:

a. In most of the tweets there is a link which conveys the actual sense of the user behind his tweet. Since we have ignored them for text mining, understanding and predicting the actual sentiment becomes tough.

b. There are a lot of tweets which are very short in length, making it problematic to code them as positive/negative

c. In most of the tweets, they post about some other topics but include a # for Ebola as it is a viral environment. In such cases, even though our model codes them as positive/negative it doesn't make any business sense.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
sentiment	Sentiment	_ASE_	Average Squared Error	0.121268	0.135451
sentiment	Sentiment	_DIV_	Divisor for ASE	1886	810
sentiment	Sentiment	_MAX_	Maximum Absolute Error	0.971759	0.809161
sentiment	Sentiment	_NOBS_	Sum of Frequencies	943	405
sentiment	Sentiment	_RASE_	Root Average Squared Error	0.348236	0.368037
sentiment	Sentiment	_SSE_	Sum of Squared Errors	228.712	109.7154
sentiment	Sentiment	_DISF_	Frequency of Classified Cases	943	405
sentiment	Sentiment	_MISC_	Misclassification Rate	0.297985	0.323457
sentiment	Sentiment	_WRONG_	Number of Wrong Classifications	281	131

Figure 12 Fit statistics of the text rule builder model

The validation misclassification rate his slightly higher than desired, but considering the above challenges, we can use this model to sore the entire dataset. For scoring we used a Score Node and then connected it to the dataset containing all the tweets from October to February. We then used a SAS Code node to export the scored dataset.

The graph below shows the sentiment pattern over time. It is clear from the display below that initially the response over Ebola was negative, as depicted by the huge gap between the two graphs. But with time the number of positive tweets increased and the gap between them narrowed as shown:

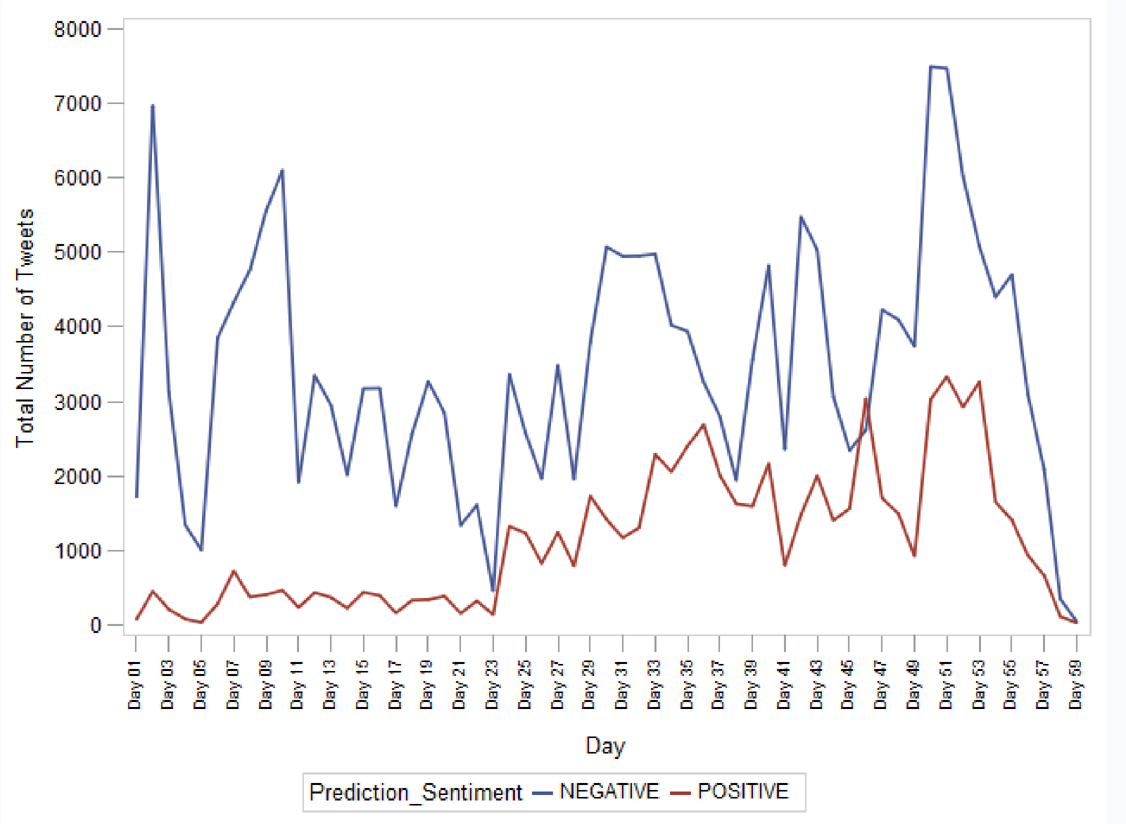


Figure 13 Timeline of number of tweets color coded by the sentiment (positive/negative)

There is a twist in the tale. In the first look it seems that overall there is a dominance of negative tweets in the dataset. But analyzing further we found that tweets which were positive were retweeted more than the negative tweets. Retweets gives us the number of times a particular tweet has been retweeted again by other users. The graph below shows that although a lot of people were tweeting in a negative sense, these received less attention for retweeting when compared to the tweets that were in a positive sense.

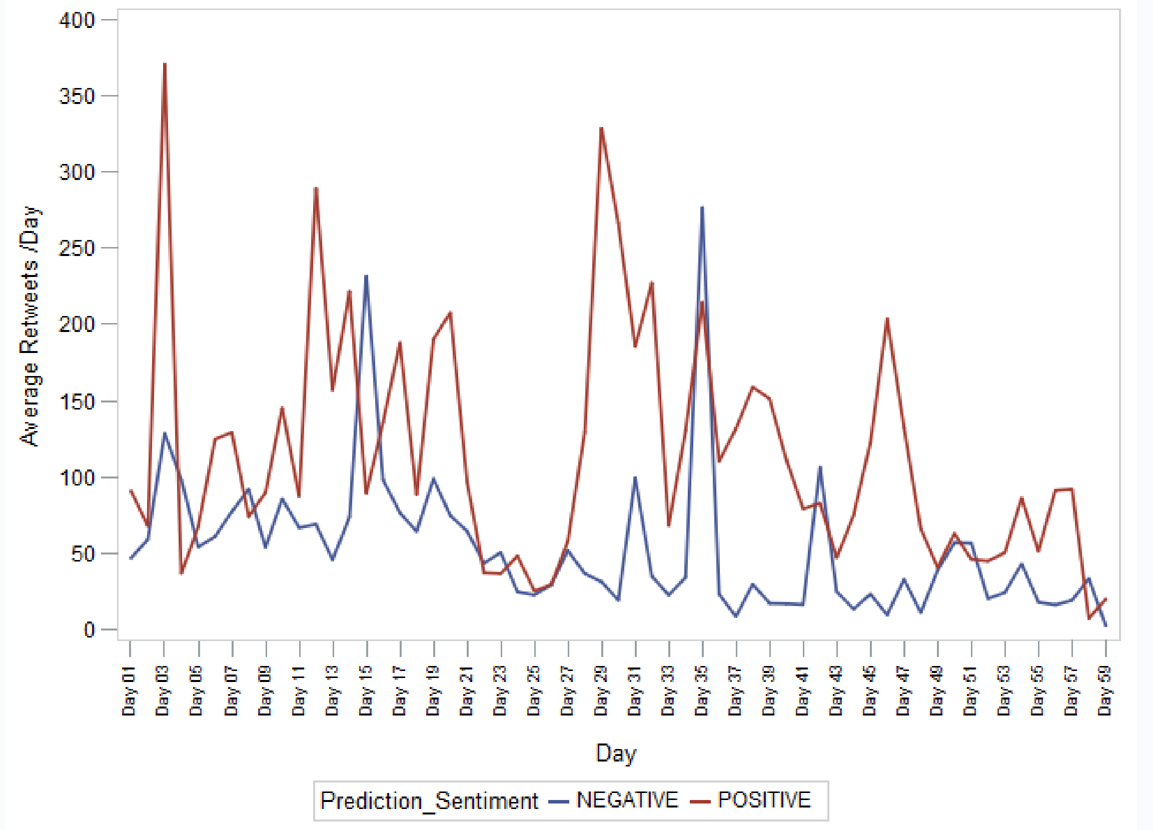


Figure 14 Timeline of average retweets color coded by sentiment

CONCLUSION

In this paper, we performed text mining and sentiment analysis of tweets regarding Ebola in the time frame of October'14 – February'15 using Google Drive/Python to download the real time tweets and SAS to clean/analyze them. We used concept links to understand relationship between terms used in the tweets. We built six customized topics (and tracked how they changed over a period of time. We then categorized sentiment of each tweet in a sample and use this sample to build a Text Rule Based Model and compared our results to a timeline of notable events that took place in that time period. We find that social, political, cultural and economic events are correlated with significant, even if delayed fluctuations of public mood levels along a range of different mood dimensions. .

REFERENCES

World Health Organization, Fact Sheet 103 Ebola Virus Disease
http://www.who.int/mediacentre/factsheets/fs103/en/

"Ebola virus disease, Fact sheet N°103, Updated September 2014". World Health Organization. September 2014. Retrieved 2014-12-15. http://www.who.int/mediacentre/factsheets/fs103/en/

"Liberia Ebola SitRep no. 234". 8 January 2015. Retrieved 9 January 2015. http://www.humanitarianresponse.info/operations/liberia/document/liberia-ebola-daily-sitrep-no-234-4-th-january-2015

Johan Bollen. 2011 "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena" *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*

Chakraborty, Goutam, Murali Pagolu and Satish Garla. November 2013. Book *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®* SAS Institute.

Getting Started with SAS(R) Text Miner 12.1- Documentation on Using Text-Filters and Text Rule Builder Nodes. http://support.sas.com/documentation/cdl/en/tmgs/65668/HTML/default/viewer.htm#bookInfo.htm

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sharat Dwibhasi, Oklahoma State University, Stillwater OK
E-mail: Sharat.dwibhasi@okstate.edu

Sharat Dwibhasi is graduate student majoring in Management Science and Information Systems at Oklahoma State University. He has been using SAS® tools over a year for Data Mining, Texting Mining, and Sentiment Analysis projects. He has published a paper in SCSUG, 2014 Proceedings.

Dr. Goutam Chakraborty, Oklahoma State University, Stillwater OK
Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.