https://doi.org/10.48009/4_iis_2017_9-15

SENTIMENT ANALYSIS OF TWITTER DATA FOR DEMONETIZATION IN INDIA – A TEXT MINING APPROACH

Kaustav Roy, Missouri University of Science & Technology, krgd9@mst.edu
Disha Kohli, Missouri University of Science & Technology, dpk3y7@mst.edu
Rakeshkumar Kathirvel Senthil Kumar, Missouri University of Science & Technology, rkp58@mst.edu
Rupaksh Sahgal, Missouri University of Science & Technology, rsc7d@mst.edu
Wen-Bin Yu, Missouri University of Science & Technology, yuwen@mst.edu

ABSTRACT

*In recent years, a merely boom was witnessed on analysis of opinions from social media because usually it is very difficult to obtain such a high volume of opinions through any other normal means of collecting opinion similar to surveys, polls etc. One such social media where opinions are expressed is Twitter and it plays a significant influence on any phenomenon. Hence, an accurate method for predicting sentiments could enable us to understand public view, and the impact of the event on social and economic setup can be analyzed. These have the potential to create positive or negative impact on general mindset of the society. The main focus of this paper is to analyze the sentiments expressed on Demonetization on Twitter so that public opinions and views are extracted, analyzed and used to understand the negative and positive impact of this act on the people of India.***

Keywords: Twitter, Opinion Mining, Sentimental Analysis, Demonetization, SAS

INTRODUCTION

The idea of demonetization is nothing but a well known phenomenon all over the world which enables refreshments in the on-boarded economy of a particular country and pushes the financial growth up for a better future. Whilst for a country like India which is the fastest growing economy with an average rate of 7% per year over the last two decades, is also encountered with massive corruption as well as severely victimized in terms of both internal and cross-border terrorism, meanwhile to counter that the Central Govt. of India declared suppression on two highest currency notes of 500 INR and 1000 INR to be used as legally approved entities for both online and offline cash transactions effective from Nov. 8 2016 onwards. But the twist lies in the fact the aforementioned declaration came from the Govt. just on the eve of Nov. 8 2016 without any prior notice, took the entire nation by surprise. In addition to that the bureaucrats also mentioned the recollection of the said currency notes and generation of new age alternative notes of 500INR and 2000INR currencies to be available from respective nationalized banks a day after.

This demonetization procedure took a whole new turn when the official body had to circulate the new age currency notes in such large scales that could reach out to all standards of people in the society. Moreover, the circular restrains the timeline to pull the old cash out from mass market and citizens in exchange of new age currency notes to the end of the year, i.e., 31st Dec, 2016. Initially, the aggressive move received support from renowned economists, scholars, several bankers as well as from some international commentators. On the other hand, it was heavily criticized by members of the opposition political parties, leading to debates in parliament and triggering organized protests against the proactive approach of the government in several places across India. Demonetization was supposedly implemented to combat corruption, terrorism financing and inflation. But it was often criticized as poorly designed and implemented, with scant attention paid to the laws of the market. Thus, the act had as much negative consequences as the positives.

The aim of this study is to perform text mining on Social media to examine the impact of this demonetization from public opinions which in turn, could be used by concerned authorities to smoothen the governmental process.

LITERATURE REVIEW

Text mining is the process of deriving useful information from natural text. Also known as text analytics, this process works by identifying various patterns in the unstructured data that can be used to bring out the underlining information. Though manual text mining was introduced in mid 1980s, the modern approach of this technique, evolved from a research by Prof. Marti A. Hearst in the late 1990s. Over the time text mining techniques have improved significantly with the help of many research and cutting-edge complex techniques. (Kiritchenko et al. 2014; Lipizzi et al. 2016).

There have been prior studies in the field of analyzing social media with text mining approaches. Hari Hara Sudhan and colleagues (2012) demonstrated the collection and summarization of tweets using the SAS® macro (%GetTweet) and then conducted sentiment analysis on the fetched tweets using SAS Text Miner. They used directed search and summarization of specific text items in SAS Text Miner. By analyzing the result, 6 clusters were formed in each group of data and were used to compare the results. Implementation of stop lists in each group removed the unwanted terms in forming clusters, and descriptive terms to understand each cluster.

Younggue Bae and Hongchul Lee (2012) demonstrated how the sentiment analysis technique can be used as a valid popularity indicator or measure. The authors developed a positive-negative measure based on their two findings. First, they distinguished between the positive and negative audiences of popular users. Second, they found that the sentiments expressed in the tweets by popular users influenced the sentiment of their audience. Lastly, the positive-negative measure was used on this influence on Granger causality analysis and to find that the time-series based positive-negative sentiment change of the audience was related to the real-world sentiment landscape of popular users.

Also, Hridoy and colleagues (2015) discussed a methodology which allows utilization and interpretation of twitter data to determine public opinions. The main focus of this research was to analyze the tweets about iphone6 to note the opinions of people based on feature specific popularity analysis and male-female specific analysis. The tweets were analyzed to be mixed but there were general consistency with outside reviews and comments.

Meire and colleagues (2016) conducted a study to measure the value of information available before and after the focal post's creation time in sentiment analysis of Facebook posts by building a sentiment prediction model which included the ‘before and after’ information along with the traditional post variables. The results indicated, ‘before and after’ information increase the model’s predictive performance. Their study also showed that the most important predictors include the number of uppercase letters, the number of likes and the number of negative comments. Another interesting outcome of their research is- the more the number of comments, the more likely it is to be a negative post.

Paltoglou (2016) analyzed if sentiment analysis can be successfully used for detecting significant events that occur in the world. The author explored whether sudden changes in the positivity or negativity that keywords are typically associated with can be exploited for this purpose. He showed that the number of tweets that are used for event detection is more important factor than the number of days used to extract token frequency or sentiment averages. Focusing on detecting local events he presented results that concluded that all approaches are dependent on the level of coverage that such events receive in social media.

RESEARCH METHODOLOGY

In this paper, initial sentiment of the people of India on demonetization is analyzed with twitter data captured between 8th Nov 2016 and 13th Nov 2016.

Aim of the Study

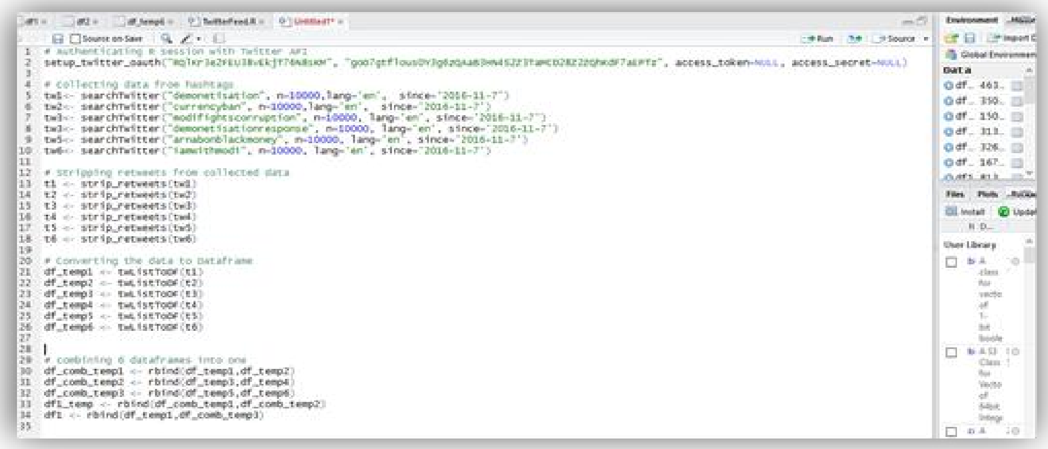
This study was performed using SAS Enterprise Miner, and R with the aim of achieving the following: (1) To extract Twitter data related to the demonetization of India using various hashtags. (2) To create refined clusters of descriptive terms corresponding to the various sentiments involved. (3) To analyze the sentiments on created clusters to find if the demonetization act has public support to the date till data has been collected.

Software and Programming Used

The following software and programming languages were used for our study: (1) SAS Enterprise Miner Workstation 12.1 (2) SAS 9.3(English) (3) R Programming and (4) R Studio.

Data Collection

Data collection process started with the identification of the hashtags that were used for the tweets related to this act. While working on collecting tweets containing the identified hashtags, Twitter further provided us with suggestions of similar or related hashtags. Many hashtags related to demonetization in India were identified this way. However, the number of hashtags there were far too many to consider. To filter out the most powerful hashtags using the hashtag analysis a website (ritetag.com) was used. This analysis gave the weightage and total hashtag exposure of the various hashtags, which were useful in identifying the most informative and useful hashtags. (Hodeghatta 2013; Saif et al. 2016)



```
# authenticating R session with Twitter API
setup_twitter_oauth("MQlkr3e2FEU3BvEkjT74N8sKM", "goo7gtflousDY3g6zQAaB3HN4S2ZTTaMCD2B2ZZQhKdF7aEPfz", access_token=NULL, access_secret=NULL)

# collecting data from hashtags
tw1<- searchTwitter("demonetisation", n=10000,lang='en',  since='2016-11-7')
tw2<- searchTwitter("currencyban", n=10000,lang='en',  since='2016-11-7')
tw3<- searchTwitter("modifightscorruption", n=10000, lang='en', since='2016-11-7')
tw3<- searchTwitter("demonetisationresponse", n=10000, lang='en', since='2016-11-7')
tw5<- searchTwitter("arnabonblackmoney", n=10000, lang='en', since='2016-11-7')
tw6<- searchTwitter("iamwithmodi", n=10000, lang='en', since='2016-11-7')

# Stripping retweets from collected data
t1 <- strip_retweets(tw1)
t2 <- strip_retweets(tw2)
t3 <- strip_retweets(tw3)
t4 <- strip_retweets(tw4)
t5 <- strip_retweets(tw5)
t6 <- strip_retweets(tw6)

# Converting the data to Dataframe
df_temp1 <- twListToDF(t1)
df_temp2 <- twListToDF(t2)
df_temp3 <- twListToDF(t3)
df_temp4 <- twListToDF(t4)
df_temp5 <- twListToDF(t5)
df_temp6 <- twListToDF(t6)


# combining 6 dataframes into one
df_comb_temp1 <- rbind(df_temp1,df_temp2)
df_comb_temp2 <- rbind(df_temp3,df_temp4)
df_comb_temp3 <- rbind(df_temp5,df_temp6)
df1_temp <- rbind(df_comb_temp1,df_comb_temp2)
df1 <- rbind(df_temp1,df_comb_temp3)
```

Figure 1. R code to extract Twitter data

#Demonetization, #Currencyban, #Modifightscorruption, #Demonetizationresponse, #Arnabonbalckmoney are a few amongst the many hashtags used to identify tweets related to demonetization. Once the hashtags were confirmed we used R code (Figure 1) to extract the Twitter data and import it as a CSV file.

Data Preparation

The collected data needed to be cleansed to be used for the sentiment analysis process. Thus, R was used to clean up the data and keep only the necessary fields. A corpus containing all necessary fields was created and all unnecessary characters were removed leaving behind the numbers, alphabets and spaces. Later, the entire text was converted to lower case.

From a total of approximately 8000 twitter records randomly around 2000 records were picked for train data set. Those records were assigned a cluster after analyzing the sentiment manually. The assigned clusters are as follows - (1) P- Positive (2) S- Statement (3) N- Negative. The rest of the 6000 tweets (from the data-set) were placed in a different file to build a test data set.

Model Preparation

Model preparation is the process of assembling different techniques which will work as a single unit with an aim to produce an intended result. In this case, two models were prepared - one for the unsupervised method and the other for supervised method.

For the unsupervised model (Figure 2), the first node is the data set. This node contains the data in SAS data set format which is created from the actual CSV data set.

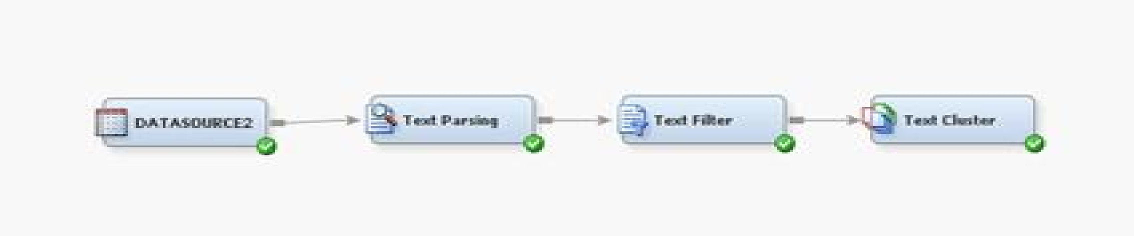


Figure 2. Model for unsupervised method

The next node is the Text Parsing node. This gives all statistical data about the terms from the data set. In addition to this, it also enables user to modify output set of parsed terms by dropping terms that are a certain parts of speech, type of entity, or attribute.

Third is the Text Filter node. It is used to filter out the duplication of terms and also the terms with similar meaning. The last node is the Test Cluster node. This is where the clustering of all the descriptive terms happens.

Next we constructed a Supervised model (Figure 3). The difference between the Unsupervised and the supervised method is that, in Supervised method we train the model with train data set containing the manually formed clusters. Once the model is trained we pass the actual test data to the model to form clusters based on its learnings from the train data set.

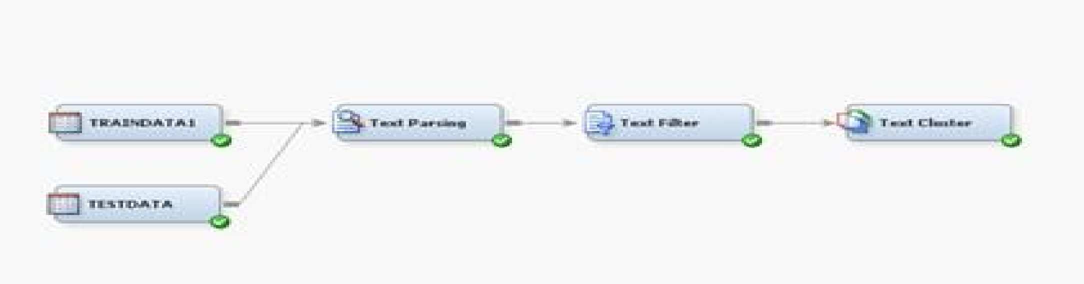


Figure 3. Model for Supervised Method

Cluster Analysis

Initially two clusters were created, positive and negative, to analyze the sentiments. Three different combinations of the settings were tried to improve the outcome and to get the best results containing the most consistent clusters. After closely analyzing the descriptive terms of the two clusters for the three different combinations, it was observed that combination 3 gave the most consistent clusters of descriptive terms with one cluster containing terms close to positive sentiment and the other containing terms close to the negative sentiment. (Table 1)

Table 1. Combinations Used for Two Cluster Unsupervised Method

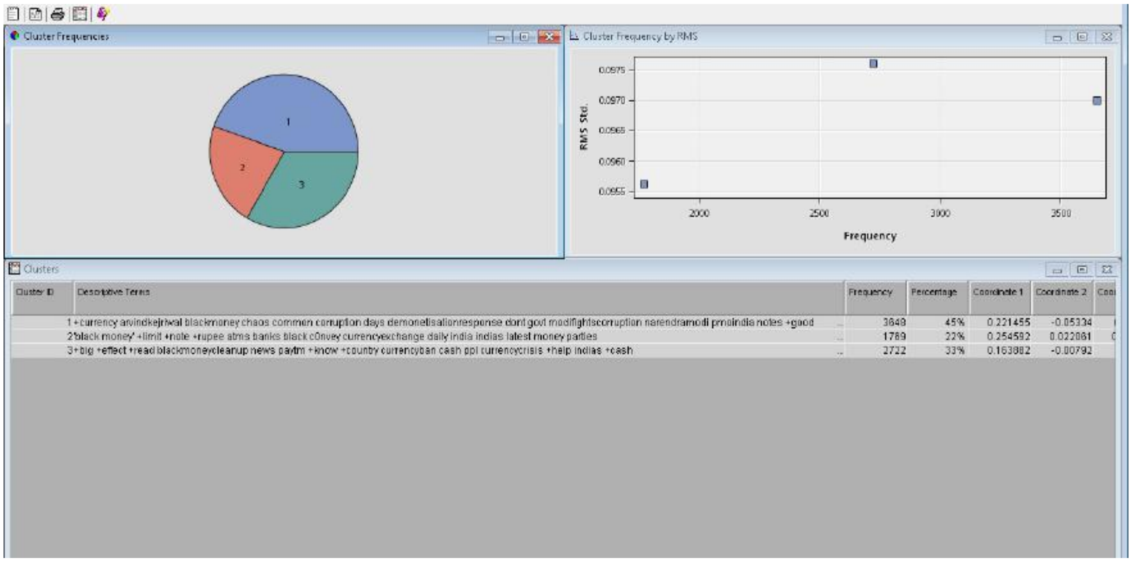
	Combination 1	**Combination 2**	**Combination 3 (Best)**
Frequency Weight	Log	Log	Binary
Term Weight	Entropy	Entropy	Entropy
SVD Resolution	Low	High	High
Clustering Algorithm	Expectation-Maximization	Expectation-Maximization	Expectation-Maximization
Exact or Minimum Number	Exact	Exact	Exact
Number of Clusters	2	2	2

Thus, the result of the combination 3 was selected and the records were analyzed for the correctness of the cluster number compared to the actual sentiment expressed by the text.

On analyzing the results, it was observed that there were lots of tweets those were mere statements and didn't exhibit either of the sentiments, i.e. positive or negative. Thus, the need for another cluster that contains neutral statements arose.
Thus, the focus shifted to the three-cluster method where we grouped the neutral statements. Configuration settings of combination 3 from two clusters method was used to perform the three-cluster method analysis.

The above method gave even more consistent clusters with each exhibiting one of the three sentiments (positive, negative and neutral). On further analysis of the results it was found that the correctness of tweet-cluster mapping was up to 55%. (Figure 4)

Further the supervised method was tried to achieve even higher percentage of correctness. In this method, only the three-cluster type with three different combinations of configuration settings was performed.



Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coo
1	+currency arvindkejriwal blackmoney chaos common corruption days demonetisationresponse dont govt modifightscorruption narendramodi pmoindia notes +good ...	3648	45%	0.221455	-0.05334	
2	'black money' +limit +note +rupee atms banks black c0nvey currencyexchange daily india indias latest money parties ...	1789	22%	0.254592	0.022081	
3	+big +effect +read blackmoneycleanup news paytm +know +country currencyban cash ppl currencycrisis +help indias +cash ...	2722	33%	0.163882	-0.00792	

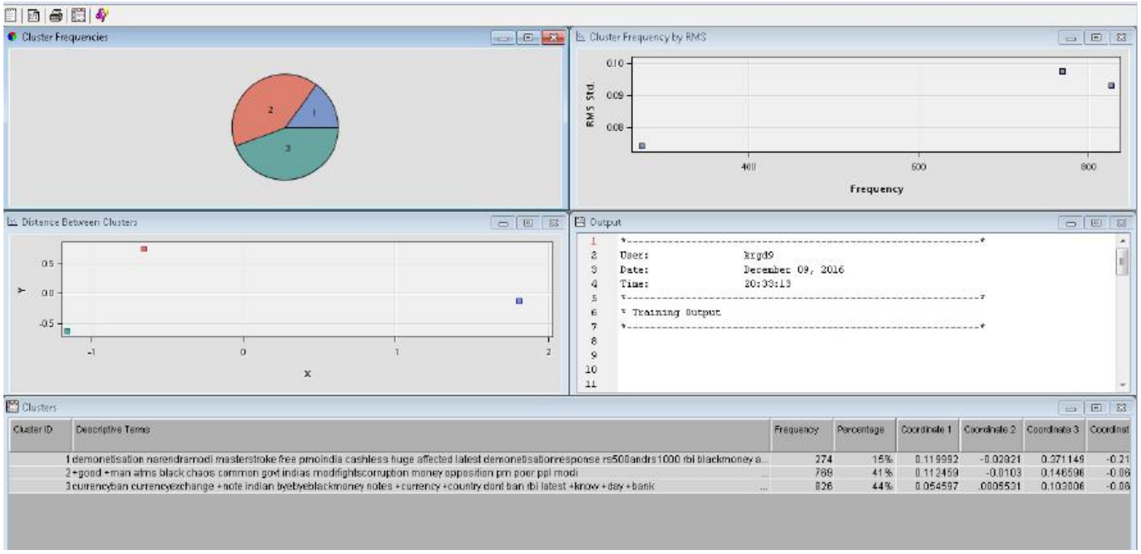
Figure 4. Descriptive Terms of Three Clusters Unsupervised Method

After closely analyzing the descriptive terms of the three clusters for the three different combinations, the second combination appeared to give the most consistent clusters of descriptive terms with each of the three clusters containing terms close to positive, neutral and negative sentiments respectively. (Table 2)

Table 2. Combinations Used for Three Cluster Supervised Method

	Combination 1	**Combination 2**	**Combination 3 (Best)**
Frequency Weight	Binary	Log	Binary
Term Weight	Mutual Info.	Mutual Info.	Entropy
SVD Resolution	High	High	High
Clustering Algorithm	Expectation-Maximization	Expectation-Maximization	Expectation-Maximization

On further analysis of the results it was found that the purity of the cluster was around 40 – 45%, which is significantly less than unsupervised three cluster method. Thus, we decided to analyze the sentiments related to demonetization in India, using the results of unsupervised three cluster method. (Figure 5)



```
*------------------------------------------------------------*
User:                  Brgd9
Date:                  December 09, 2016
Time:                  20:33:13
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinat
1	demonetisation narendramodi masterstroke free pmoindia cashless huge affected latest demonetisationresponse rs500andrs1000 rbi blackmoney a...	274	15%	0.119992	-0.02821	0.371149	-0.21
2	+good +man atms black chaos common govt indias modifightscorruption money opposition pm poor ppl modi ...	769	41%	0.112459	-0.0103	0.146596	-0.06
3	currencyban currencyexchange +note indian byebyeblackmoney notes +currency +country dont ban rbi latest +know +day +bank ...	826	44%	0.054597	.0005531	0.103006	-0.06

Figure 5. Descriptive Terms of Three Clusters Supervised Method

RESULTS AND CONCLUSIONS

The following results were obtained on analyzing the twitter data, in accordance to the demonetization in India (Nov 2016), for sentiments:

- Among all the tweets, approximately 45% of them were identified to be in support of this move from PM of India.

- Nearly, 22% of the tweets were identified to be neutral comments or mere statements on the effects of this act.

- With 33% of the tweets identified to be exhibiting negative sentiments most of them were identified against the execution of the act rather than the act itself.

While analyzing the sentiments of the results it was observed that there are many statements in the tweets for which the sentiments are neutral. In the remaining tweets the positive sentiment is on the higher side, approximately 50-55%. The following are a couple of example of tweets that bear positive sentiment:

- "thanks #narendramodi for #demonetisation of course v r in pain expectation is 4 higher goal n my family 100 supports u go on v r wid u"

- "modis demonetization drive has destroyed terror networks in Kashmir #blackmoney #demonetisation"

Many of the tweets found to be negative were not negative to the idea of Demonetization but to the execution process. The following are a couple of examples of such tweets:

- "govts #demonetization good idea but implementation badly planned…"

- "#currencyban #narendramodi situation is really bad no money in atms bad execution of such a strong good and right decision"

Sentiment analysis of social media or online texts is really important and of prime importance in many industries like online customer care service, spams filtering, contextual advertisement etc. This paper tries to capture the sentiment of the people on demonetization in India with the help of tweets. This can be helpful for the authorities to resolve the issues and make the transition smoother for people.

Some of the future works that can be done on this paper are:

- To improve the supervised model to get better efficiency as it is a proven fact that supervised models are capable for producing results with greater precision.
- To analyze the long term impact of demonetization in India and how far this act is successful in resolving the targeted problems.
- The same process can be followed to perform sentiment analysis on GST (Goods and Service Tax) in India, which has been rolled out on 1st July 2017 and is a great leap forward from the existing tax structure.

REFERENCES

Bae, Y., & Lee, H. (2012). Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the Association for Information Science and Technology*, *63*(12), 2521-2535.

Hodeghatta, U. R. (2013). Sentiment analysis of Hollywood movies on Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1401-1404). ACM.

Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., & Rahman, R. M. (2015). Localized twitter opinion mining using sentiment analysis. *Decision Analytics, 2*(1), 8.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research, 50,* 723-762.

Lipizzi, C., Iandoli, L., & Marquez, J. E. R. (2016). Combining structure, content and meaning in online social networks: The analysis of public's early reaction in social media to newly launched movies. *Technological Forecasting and Social Change, 109*, 35-49.

Meire, M., Ballings, M., & Van den Poel, D. (2016). The added value of auxiliary data in sentiment analysis of Facebook posts. *Decision Support Systems, 89,* 98-112.

Paltoglou, G. (2016). Sentiment-based event detection in Twitter. *Journal of the Association for Information Science and Technology, 67*(7), 1576-1587.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management, 52*(1), 5-19.

Sudhan, H. H., Garla, S., & Chakraborty, G. (2012). Analyzing sentiments in Tweets about Wal-Mart's gender discrimination lawsuit verdict using SAS® Text Miner. In *SAS Global Forum*, 306. Retrieved from http://support.sas.com/resources/papers/proceedings12/306-2012.pdf