Sentiment Analysis and Opinion Mining of Twitter Data for Modeling Public Mood and Emotion on Demonetization in India

Ashwin. R. Jadhav[1]
Indian Institute of Technology
Vellore, India
ashwin.ravishankar01@gmail.com

B.K. Triparty[2]
Vellore Institute of Technology
Vellore, India
tripathybk@vit.ac.in

Shivansh Jagga[1]
Vellore Institute of Technology
Vellore, India
shivansh.jagga@gmail.com

***Abstract*—On 8th November 2016, Prime Minister Narendra Modi announced the demonetization of all notes of 500 rupees and 1000 rupees denominations in an unscheduled live televised address at 20:00 IST. The surprise nature of the announcement prompted widespread criticism from the opposition parties which caused debates in both houses of parliament. In this modern age where sharing information on social networks such as Twitter has become a part and parcel of the daily life, it motivated us to propose a sentiment analysis and opinion mining to extract public reaction. In this paper, we explain the process of data extraction and preprocessing and summarize the overall public sentiment and analyze the reaction times analytical tools such as word-clouds, time series analysis. Furthermore, a detailed sentiment analysis are also presented to provide a more sophisticated method to model the overall public mood and reactions to social and economic outcomes in order to determine the success or failure of the demonetization policy implemented by the Indian government.**

Keywords— Twitter Analysis, Twitter Classification, Sentiment Analysis, Time series analysis, Opinion mining.

I. INTRODUCTION

On 8th November 2016, Prime Minister of India Narendra Modi announced the demonetization of all notes of 500 rupees and 1000 rupees denominations in an unscheduled live televised address at 20:00 IST. The surprise nature of the announcement prompted widespread criticism from the opposition parties which caused debates in both houses of parliament. The government defended the demonetization policy amidst a major public outcry. The policy had its advantages such as: It was an attempt to decrease corruption in India. Secondly, it was done to curb black money and stop funds flowing to illegal activity. It also aimed to increase accountability among people and make them pay income tax. It was also supposed to force people to make cashless transactions and help establish a Digital India. In retrospect, this policy was actually in the grand scheme of things since the government went on to introduce another major policy called GST in July of 2017.

Although this policy aimed to improve the Indian economy on the long run, it failed miserably on the short term. Absence of cashless transactions immediately affected business and plunged the Indian economy causing major disruptions across the nation. Daily wage labors were left with no jobs and the supply of new 500 and 2000 notes was scarce. Due to this, they was a public debate to discuss the success and effectiveness of the policy. On one hand there were many supporters who didn't mind the temporary disruptions and on the other hand there was a public outcry about how this policy negatively affected the common man with people getting angry and impatient.

Social computing is an effective method to analyze and model socio-economic events in real time. It can be used to extract useful and critical information about these events (King, 2009) [1]. Twitter has become the most popular micro blogging service where users share short messages called tweets. These textual information also contain sentiments that can be extracted (Barahate, 2012) [2]. The user base is rapidly increasing (Krikorian, 2013) [3] generating a lot of data about diverse topics. These tweets are limited to 140 characters and this simplifies their analyses (Kasture and Bhilare, 2015) [4]. The contemporary sentiment analyses methods are used for disaster relief, political polls, scientific surveys, checking customer loyalty, unemployment rates, population health care (Dredze,2012) [5], advertising market and in education (Siemens and Long, 2012, Chen et.al 2014) [6,7].

This motivated us to extract the opinions and sentiments from Twitter in order to determine the overall public mood on this policy. Consequently, the twitter data was extracted from different times. First one during Nov 2016 when the policy had just been implemented. Second time, the data was extracted in the month of April 2017 after 6 months. Useful analytical tools such as word clouds, time series plotting and sentiment analysis are presented to accurately gauge the public emotion and opinion. With our analysis we attempt to identify a quantifiable relationship between overall public mood and reactions to social and economic outcomes in order to determine the success or failure of the demonetization policy implemented by the Indian government.

The rest of the paper is organized accordingly: Section 2 discusses other related works. Section 3 presents data collection technique and Section 4 describes the proposed methodology. Finally, Section 5 presents number analyses on the data and discusses the obtained results. Section 6 concludes the paper.

II. Related Works

Sentiment analysis started by being a document level classification task (Pang and Lee, 2004) [8] and then handled more complex tasks (Hu and Liu, 2004, Wilson et al., 2005[9, 10]). One of the earliest works on sentiments analysis was done by Go et al. (2009) [11] where distant learning methods were used to classify tweets ending with positive emoticons. Some of the other early works include (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010) [12,13]. Another important work on sentiment classification on Twitter data was by Barbosa and Feng (2010) [14]. It used polarity predictions from three websites as noisy labels to train a model and then uses 1000 manually labeled tweets for testing.

The number of empirical analyses of sentiment and mood are based on textual collections of data generated on social media has steadily increased. Some of the important mood surveys of communication are on Twitter (Thelwall et al. 2010) [15] and Myspace (Thelwall, Wilkinson, and Uppal 2009) [16]. Several studies also focused on response to a specific event such as Michael Jackson's death (Kim et al. 2009) [17] or a political election in Germany (Tumasjan et al. 2010) [18]. Other studies focused on much broader trends such as consumer confidence and political opinion (O'Connor et al. 2010) [19]. Socio-economic trends have also been studied by (Bollen, Mao, and Zeng 2010) [20] where change in Twitter mood is related with Wall Street fluctuations.

The results produced by the investigation of such collective mood aggregators are convincing and may even accurately indicate the overall public mood. Moreover, using openly available data such as Twitter data can help perform better sentiment analyses due to more outreach compared to normal surveys, reduced costs and reduced time to collect the data. In our case, we try to analyze the public mood regarding the demonetization policy implemented in India. Accurate opinions can be extracted using the Twitter data which can help economists, policy makers and government to make better decisions in the future.

III. Data Acquisition

Twitter is a social microblogging network where users post short messages, called Tweets. These short messages are usually limited to 140 characters and for this reason, people often tweet with many acronyms, special characters and emoticons to express their thoughts. Data collection mainly involved the use of "twitteR" package in R using the twitter API. This enabled collection of data using hash tags and stream real time data from Twitter. Our study is based on an important political event in India, the demonetization of 2016. The data was collected during two different times in order to see the effects at different times. Firstly, the data was collected right after the event in Nov 2016. Then, further data was collected for the month of April 2017 to understand how the sentiments had changed after a period of 6 months. Accordingly, a total of 15,251 tweets were collected including both the timelines. Many other feature such as retweet count of each of these tweets, Tweeter ID, created-at, text, statusSource, favorite-counts were also noted. The original tweets contained many elements such as stop words, user mentions, URLs, special characters, emoticons and hashtags. A sample of the collect tweets is given in Fig. 1.

RT @bhaiyyajispeaks: Here @sardesairajdeep struggling for one answer against #Demonetization, He should have taken some Congress leader lik…
RT @dhrumilpatel: Looks Ambit Capital, literally Choked with #demonetization, come-on U can't lower d GDP nos by huge margin. Perhaps, they…
RT @roshankar: Harvard's Larry Summers calls #Demonetization as poor policy with disproportionate negative impact on poor/trade.
RT @ModiBharosa: Huge support for PM @narendramodi 's #demonetization Move Across the Nation 80-86% people back demonetization: C-voter su…

Fig. 1. Sample Tweets

Finally, the collected raw data had to pre-process before doing any analysis with them. The data pre-processing step is discussed in detail in the next section.

IV. Proposed Method

Analysis on Twitter data can give important information about public mood and opinion on major socio-cultural events. This can help government, researchers take important decisions based on these sentiment analysis. The proposed model is given Fig. 2.

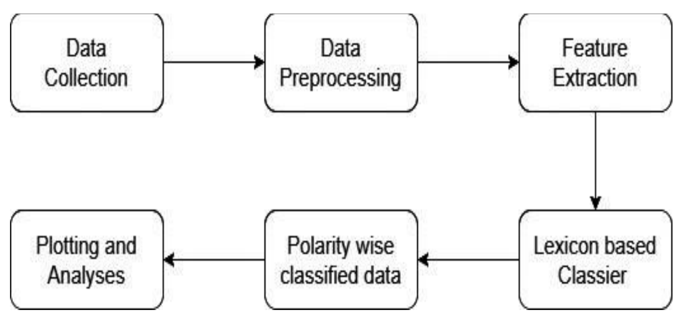

Fig. 2. Methodology Stages

The first step is to collect the data using various available techniques. Since this collected data is raw, it needs to be preprocessed so as to use it for further analyses. Hence, next step is to clean the data and processing it. Next, the features are extracted from the processed data and it prepared to be feed to the classifier. Next step is to use any pre-trained classifier to help derive the sentiments from the given data. Finally, many

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

analysis techniques can be used such as qualitative content analysis, word clouds, time series plotting and sentiment analysis in order to give us a clear overall picture. The pre-processing, feature extraction and sentiment analysis steps are discussed in detail.

A. Data Preprocessing

The data collected using the Twitter API is raw. It contains many extra elements such as usernames, hashtags, emoticons and stop words. In order to apply a classifier, these tweets need to be cleaned. The pre-processing task involves removing all these extra unnecessary elements in the data. The following pre-processing procedure was followed:

- Removal of common pointers such as username (@), hashtags (#) and retweets (RT).
- Removal of common stop words like are, is, am etc. Removing stop words helps compress the data with losing meaning of the sentences.
- Removal of the URLs and hyperlinks. Many tweets contain URLs in order to redirect to more content. But, in order to compress the data we can replace the all links by <URL>
- Removal of emoticons and replacing them with emotion keywords. Almost every tweet contains emoticons and they express strong emotions. Hence, it's important to consider them.
- Compression of the elongated words such as happyyy into happy and decompress of words such as g8, f9. Usually these slang words contain the extreme level of sentiments. So it is necessary to decompress them.

B. Feature Extraction

We first need to extract features from the cleaned data so that it can be used for sentiment classification. We have used a common feature extraction method to represent sentiments using a lexicon based technique. This techniques derives opinion based on lexicons from the text and then determines its polarity. The words are tagged with prior polarity such as neutral, negative or positive.

C. Sentiment and Opinion Mining

After features are extracted, we can use these to analyze the data. An estimate of positive, negative and neutral sentiments can be derived from the above extracted features. However, a deeper analyses can be done to extract detailed sentiments and opinions. This is achieved by using the NRC Sentiment and Emotion Lexicons. It is a collection of seven lexicons and each lexicons has a list of words associated with them. These categories include emotions such as joy, sadness, fear, anger, anticipation, disgust, surprise and trust. It can even classify sentiments such as positive or negative. This is exactly what is required in our case so as to track the exact public mood and opinions towards the political policy. Furthermore, we can help identify what evokes strong emotions in people. Thus, this can help in predicting the overall happiness of people with the demonetization policy and the popularity of the policy.

These options are mined for the twitter data in the month of November and April. Thus, we can analyze how the public mood has changed after 6 months and in what direction. If the policy is successful, then by the month of April we should find more positivity towards the demonetization topic. This sentiment analysis can help the government to take further steps in the right direction and help maintain the overall public happiness ad satisfaction with the governance.

V. PERFORMANCE EVALUTAION

After the data is collected and preprocessed, it is time for analyzing it. After building the intelligent model to determine the sentiments and opinions of public, we can gather critical information about the success of the demonetization policy. First, we analyze by generating a word cloud. Next we plot time series between the number of tweets and retweets against time of the day to get a clear picture of the activity periods of the people. Finally, we do sentiment and opinion mining to extract the public moods. The evolution of sentiments is plotted based on 3 different emotions: positive, negative and neutral.

A. Word Clouds

Word clouds are generated to get more details about the most significant topics in the collected data. The first word cloud is generated with the most significant topics that was tweeted during Nov 2016 as shown in Fig. 3. It can be observed that it contains terms like "Terrorists, RS Looted etc" which are not relevant to demonetization. This is because the implementation of the policy also coincided with other events. Hence we generate a second word cloud relevant to PM Narendra Modi to analyze the political trends. It contains more supportive and positive terms towards PM and it might be a sign of the success of the implemented policy.

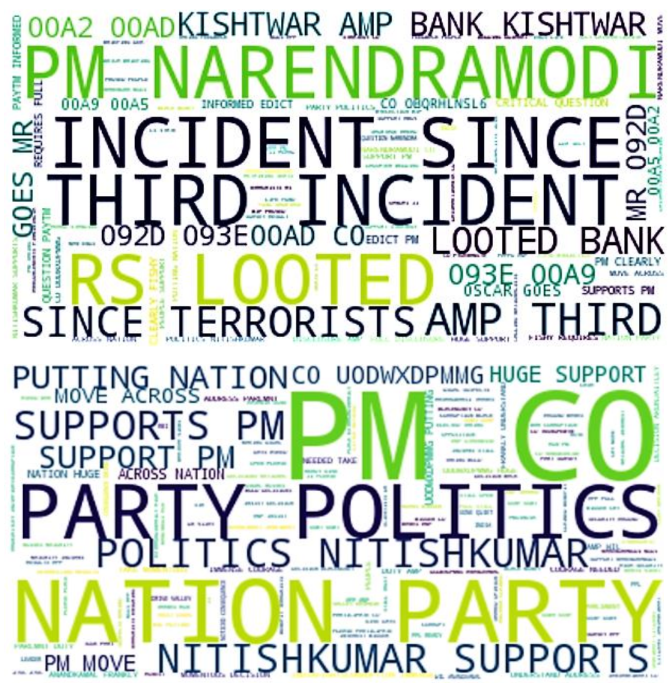

Fig. 3. Word Clouds

B. *Time Series Analysis*

A time series plot can help us identify more complex relationships in the data. Hence, two different time series plots are produced: one with number of tweets and the other with number of retweets against time of the day in hours. It is further plotted for Nov 2016 data and April 2017 data separately. By doing so, we can determine the change in activity at these two different time. Fig. 4 shows both the plots for the month of Nov 2016. The number of tweets shows only the unique tweets that were tweeted. Hence its count is lesser than retweet count which is in the order of thousands. These tweets were aggregated and averaged over hours of the day. Hence, it shows spike in activity during the day and subsides during night. The second graph shows the retweet count and it shows how widely the tweets were shared with retweet count crossing more than 200000 many times. Its magnitude is in tens of thousands and that shows that topic was very hot during this time.

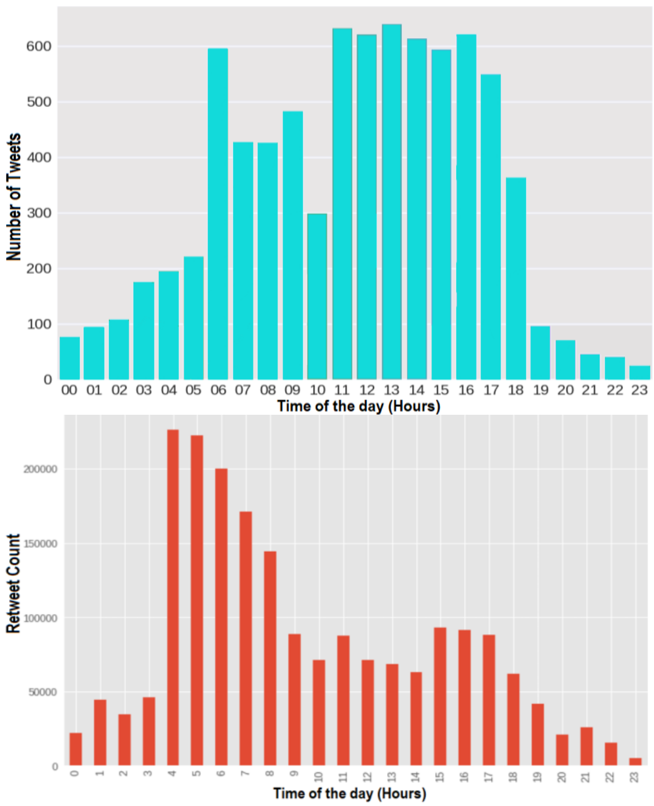

Fig. 4. Time series plots for November 2016

Now that we analyzed for November, we can look at the time series plots for April 2017 as shown in Fig. 5. Immediately, we can notice that the number of tweets and number of retweets have reduced when compared to the previous plot of Nov. This correlates to the real time events where most of the people were over the disruptions caused by the policy and the normal life had resumed. However, the activity did not fully decrease because people were still talking about further polices and how the demonetization was going to affect them along with GST policy. Now that we have a clearer idea about the rate of activity among the public regarding this policy, we should now try to determine the opinions and sentiments of these people to help model overall public mood.

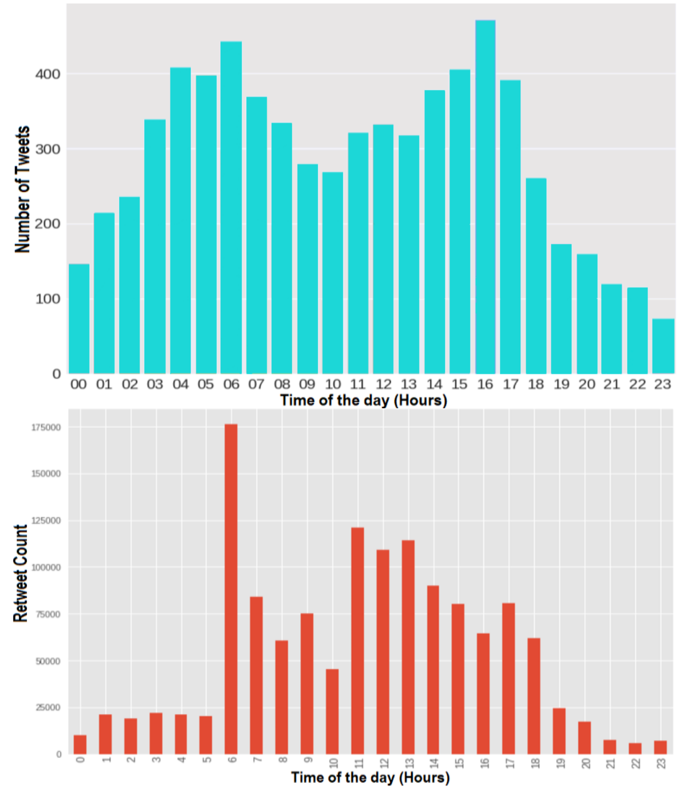

Fig. 5. Time series plots for April 2017

C. *Time Series Analysis*

In order to better understand and model public emotions, we need to do sentiment analysis and extract opinions. First, we tag each of the preprocessed tweets to positive, negative or neutral. Then we aggregate the mood vectors for a set of tweets submitted during that particular month m and average it out for every hour in a day. This way we can plot and track the positive responses, negative responses and the neutral responses from the public. This evolution of sentiments is plotted for the month of Nov 2016 and April 2017 separately.
Fig. 6 shows the evolution of sentiments during November and we can infer that while general positivity was present, it was also accompanied by significant negative response and almost equal neutral responses. This shows that while certain segments of society were happy, many other communities were really unhappy and remaining one were left in doubt. This correlate really well with the real life events that occurred because while many people supported the policy for the long run, many poor people were left jobless due to lack of running cash and a large segment of middle class families were waiting for weeks for withdrawing money causing the nation to come to a standstill for at least 2 weeks. Now, we can generate the evolution of sentiments plot for the month April 2017. One can immediate notice that the positive responses have significant majority. Although negative and neutral response still exist, they has significantly gone down. This is good sign for the policy makers because it shows that people

have accepted the demonetization policy after facing all the short term inconveniences. The evolution of sentiments plot can be seen in Fig. 7.

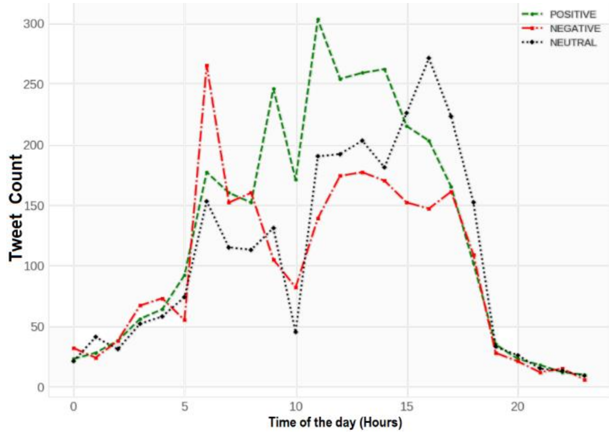

Fig. 6. Evolution of sentiments (Nov '16)

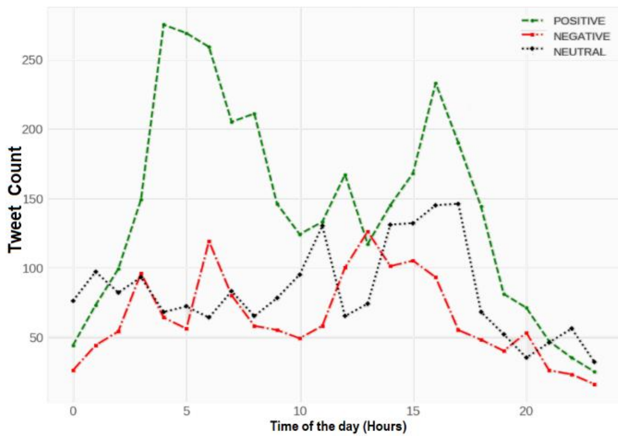

Fig. 7. Evolution of sentiments (Apr '17)

Furthermore, to get a more detailed analysis of the public moods, we can mine the opinions from the overall data. By using the NRC Sentiment and Emotion Lexicons we have categorized all the tweets into 8 different emotions and opinions. These include joy, sadness, fear, anger, anticipation, disgust, surprise and trust. Fig. 8 shows the plot of number of tweets with the respective opinion category.

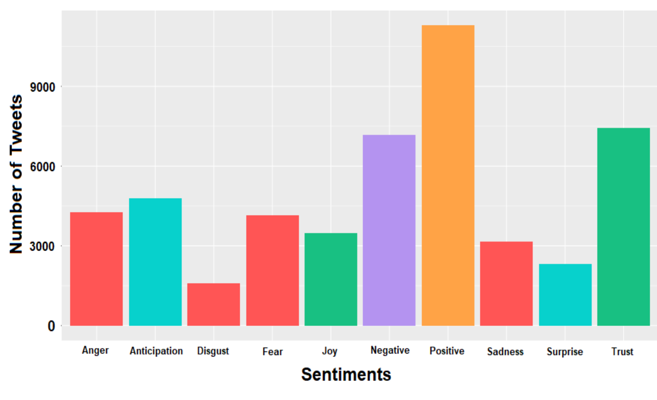

Fig. 8. Overall Public Opinion on Demonetization

We can notice that the trust category had the most number while other opinions like anger, joy, fear and surprise all have almost equal numbers. This shows that although people have reservations about the policy, they still trust the government and are hoping that in the long run this will have positive outcomes. Hence, the policy has had a positive opinion from the public overall. This is very useful for the government's policy makers since they are planning to extent this policy even further. They can now make better well planned policies later this year based on these types of analyses so as to decrease surprise and fear amongst the people.

VI. Conclusion

In this paper, we have attempted to model public opinion towards the demonetization policy in India to help analyze the political and social-economic relationships. This can help the policy makers and economists to make better and well planned decisions in the future. A thorough analysis was done with the help of word clouds, time series plotting as well as sentiment analysis and opinion mining. In the end, we can conclude that the overall public mood has been positive with the need of better information and planning from the government to help alleviate the public's concerns in the future.

References

[1] I. King, J. Li and K. T. Chan, "A Brief Survey of Computational Approaches in Social Computing", in Proc. of Int. Joint Conf. on Neural Network, 2009, pp. 2699-2706.

[2] S. R. Barahate and V. M. Shelake, "A Survey and Future Vision of Data mining in Educational Field", in Proc. 2nd Int. Conf. on Advanced Computing & Communication Technology, 2012, pp. 96-100.

[3] Krikorian, Raffi (VP, Platform Engineering, Twitter Inc.), "New Tweets per second record, and how!" Twitter Official Blog. August 16, 2013.[Online].

[4] N. Kasture and P. Bhilare, "An Approach for Sentiment analysis on social networking sites", Computing Communication Control and Automation (ICCUBEA), 2015, pp. 390-395.

[5] M. Dredze , "How Social Media Will Change Public Health", IEEE Intelligent Systems, 2012, pp. 1541-1672.

[6] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education", Educause Review, 2011, vol. 46, no. 5, pp. 30-32.

[7] X. Chen, M. Vorvoreanu and K. Madhavan, "Mining Social Media Data to Understand Students' Learning Experiences", IEEE Transaction, 2014, vol. 7, no. 3, pp. 246-259.

[8] B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.

[9] M Hu and B Liu. 2004. Mining and summarizing customer reviews. KDD.

[10] T.Wilson, J.Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. ACL.

[11] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[12] Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836.

[13] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[14] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[15] Thelwall, M.; Wilkinson, D.; and Uppal, S. 2009. Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology In press.

[16] Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; and Kappas, A. 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12):2544–2558.

[17] Kim, E.; Gilbert, S.; Edwards, M.; and Graeff, E. 2009. Detecting Sadness in 140 Characters: Sentiment Analysis of Mourning Michael Jackson on Twitter. Technical report, Web Ecology Project, Boston, MA.

[18] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Fourth International AAAI Conference on Weblogs and Social Media.

[19] O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In Fourth International AAAI Conference on Weblogs and Social Media.

[20] Bollen, J.; Mao, H.; and Zeng, X.-J. 2010. Twitter mood predicts the stock market. Journal of Computational Science 2(1):1–8.