International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 10 || October 2017 || PP. 68-74

Sentimental Analysis on Political Aspects from Tweets and Retweets

Sabarmathi.N[1], Premalatha.A[2]
[1](M.Sc, Shri SakthiKailassh Women's Arts& Science College/Periyar University, India)
[2](MCA,M.Phil,B.Ed, Shri SakthiKailassh Women's Arts & Science College/Periyar University, India)

Abstract: Swelling number of Internet users and Online Social Media (OSM) users turned these unconventional media platforms into key medium in these elections; that could affect 3-4% of urban population votes as per a report of IAMAI (Internet & Mobile Association of India). Political parties making use of Google+ Hangout to interact with people and party workers, posting campaigning photos on Instagram and videos on YouTube, debating on Twitter and Facebook were strong indicators of the impact of the OSM on the General Elections. With hardly any political leader or party not having his account on the micro blogging site twitter and the surge in the political conversations on Twitter, inspired us to take the opportunity to study and analyze this huge ocean of elections data. We analyzed the complete dataset to find interesting patterns in it and also to verify if the trivial things were also evident in the data collected. We found that the activity on Twitter peaked during important events related to elections. It was evident from our data that the political behaviour of the politicians affected their followers count and thus popularity on Twitter. Yet another aim of our work was to find an efficient way to classify the political orientation of the users on Twitter.
Keywords: Tweets, Retweets, Online Social Media

I. INTRODUCTION

Social Media has recently evolved into a source of social, political and real time information. In addition to this it is also a great means of communication and marketing. People have been sharing information on social networks through the use of status updates , blogging, sharing multimedia content like images and videos as well as interacting together thereby forming groups and communities on social networks. Monitoring and analyzing this information can lead to valuable insights that might otherwise be hard to get using conventional methods and media sources. The social networking sites such as Facebook, Twitter and Flickr provide a new way to share the information among them and get frequent updates. In addition to this, the sites also allow sharing of additional information which can be important in analysing the contents, e.g. location etc. The social media has an advantage over conventional media sources as it is managed by the users. Conventional media only allowed users to gain information that was provided to them. The flow of information was only one-sided from the media to user. With social networks, however, the users now have the ability to respond to the news and events around them and provide their opinion on the as well as share them. This leads to the evolution of a multi-way mode of information dissemination in which the users post information along with other information like links images and videos. As a result, a user generated model of information is generated. The social graph of users and their connections on the social networks plays an important role in analysing this information model in order to obtain meaningful data from the vast amount of "user generated content" that is created every day. Since, the micro-blogging sites like Facebook, Twitter and Flickr allow users to share short messages and multimedia, they have become an instant source of information through which users from all around the world can remain connected and get to know about the information from several sources.

Mining social media sites, Twitter in particular, has been the focus of numerous recent studies, with a broad range of focus: analyzing social media platforms to study how these platforms can facilitate smoking cessation, mining public health information, detecting influenza epidemics, predicting election voting results, and studying global mood patterns. Twitter messages have also mined for the detection of drug-related adverse events, the assessment of the adequacy of gender identification terms on medical intake forms, and the analysis of U.S. weekly trends in work stress and emotion. Social media provide new data sources that significantly expand the range of what can easily be measured, and thus facilitate computational knowledge discovery.

The current phase on the internet is witnessing a tremendous growth of social networks and huge amounts of new data are being created every second. With the advent of social networks, it has also become possible to disseminate this information at very fast rates. Millions of new user posts everyday are being created on social networking sites like Facebook, Twitter, Wordpress and Flickr. In this section, we present a brief introduction about social networks with a special focus on twitter. Twitter is not only a fantastic real-time social

networking tool; it also acts as a great source of rich information for data mining. On an average, the users on twitter produce more than 140 million tweets per day.

There was a significant change in the General Elections 2014 from the General Elections 2009; this was the change in the role played by the social media during the elections. It has been observed worldwide that the democracies have been engaging in dialogues with the public over the social media. Twitter has played an outsized role in a 2016 presidential election that continues to test the electorate. Despite Twitter's ongoing business problems, the ability of a single tweet to shape political conversation and drive media coverage has never been greater. A marked contrast exists between Twitter's business acumen (or lack thereof) and the sometimes seemingly unintentional influence it wields on the current election.

Twitter's effect on the 2016 presidential election cycle will have lasting reverberations. Yet the company could potential be out of business by the time Americans head to the polls in 2020.The current election also reinforces the idea that Twitter is becoming less of a social network and more of a news-making medium with a social bent, Twitter's future is uncertain, but the format it introduced will have a lasting impact on politic.

II. RELATED WORK

Twitter has served as a platform for information dissemination, banter, breaking news, spreading rumors and many other purposes. This provides good opportunities to researchers to study the real-time events, like Sakaki etal. did earthquake detection on the basis of tweets [24] and Aramaki et. al. used it for influenza detection [1]. These real-time events can also be studied for the credibility in the tweets as done by Gupta et. al. [13]. One such real time event that has always made news is Elections.

Prediction of Elections Using Twitter

Tumasjan et. al. [26] who analyzed 100,000 tweets, in their paper claimed that a mere mention or volume analysis of the tweets related to elections was enough to predict the results. They also said that the people's sentiments in real world are closely related to the tweets' sentiments. Till date, this paper has had the most successful attempt in election predictions. The first attempt at mood detection is known to be done by Bollen et. al. [3], where they classified the tweets to be belonging to 6 different categories of moods namely tension, depression, anger, vigor, fatigue and confusion. Jungherr et. al. [17] however challenged the results of Tumasjan et. al [26] by saying that the results were time dependent and that all the parties were not taken into account.

A paper in counter response was also published by Tumasjan et. al. [27] in which they toned down their claim. We then had papers that started raising doubts over the predictive powers of Twitter for elections. First in this series was the paper by Metaxas et. al. [19]. They concluded the success in prediction of elections as a chance and cautioned that one must look at the demographics of the population before predicting elections. Yet another doubt that was raised on the prediction concerned the credibility of tweets by Castillo et. al. [7] and on the users tweeting about elections by Mustafaraj et. al [21].

Election Data Analysis

Skoric et. al. [25] in their research studied the elections data during the Singapore General Elections. They found out that though the predictive power of Twitter for elections cannot be claimed to be as good as in the Germany elections by Tumasjan et. al. [26], but it is still better than chance. Their mean absolute error was higher than that of previous studies and concluded that Twitter is indicative of the public opinion. Gayo-Avello[11] in his paper studied the US Presidential Elections 2008 and shows how analysis of twitter data failed to predict Obama's win even in Texas. He claimed that the twitter data is biased and cannot be used as a representative sample. He also challenged the sentiment analysis used in the earlier papers. As far as India General Elections 2014 were concerned, twitter data was also analyzed by Simplify. They prepared both short summary as well as a detailed report on the elections data. They prepared a Simplify Social Index (SSI) to calculate the popularity of the politicians. Awareness, spread, prominence and favorability were 4 parameters they used for the calculation of their index. The analysis by the NExT Center at the National University of Singapore was a weekly analysis.2 They would find out the statistics about the three major parties AAP, BJP and Congress from the weekly data and also report the political events that took place that week. Their last section had some political reviews about the three main candidates. Kno.e.sis, a research group at Wright State University also analyzed India General Elections 2014 with the help of Twitris+, a semantic social web application.3 The portal showed the hopefulness for the three major parties. The hopefulness was calculated taking the number of mentions and sentiments of the tweets as parameters. Apart from this, there were several portals by news media houses that showed some portion about the kind of activities going on on Online Social Media (OSM), for e.g., the pages by IBN and TOI .

III. SYSTEM ANALYSIS

Lexicon-based Approach

The lexicon-based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document. Dictionaries for lexicon-based approaches can be created manually, as authors describe in this article or automatically, using seed words to expand the list of words. Much of the lexicon-based research has focused on using adjectives as indicators of the semantic orientation of text. First, a list of adjectives and corresponding Sentiment Orientation (SO) values is compiled into a dictionary. According to previous study, adjectives are good indicators of SO. Then, for any given text, all adjectives are extracted and annotated with their SO value, using the dictionary scores. The SO scores are in turn aggregated into a single score for the text. However, although an isolated adjective may indicate subjectivity, there may be insufficient context to determine semantic orientation. As pointed out by Turney P., the adjective "brutal , „insane may have negative orientationin ape review, in a phrase such as "brutal and insane breed of dog", but it could have largely a positive orientation in a movie review, in a phrase like "brutal and insane action sequence". Therefore the algorithm extracts two consecutive words. The first member is an adverb or an adjective while the second word provides the context.

A) Dictionary-based approach

Dictionary-Based approach involves using a dictionary which contains synonyms and antonyms of a word. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specifically, this method works as follows: A small set of sentiment words (seeds) with known positive or negative orientations is first collected manually, which is very easy. The algorithm then grows this set by searching in any online available dictionary for their synonyms and antonyms. The seed list will be added with the new found words. The process iteratively keeps on adding the words until no more new words are found. Manual inspection can be used to clean up the list at last.

B) Corpus Based Approach

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. There are two methods in the corpus based approach:

C) Statistical Approach

If the word appears intermittently amid positive texts, then its polarity is positive. If it appears frequently among negative texts, then its polarity can be considered as negative. If it has equal frequencies, then it can be considered as neutral word. Seed opinion words can be found using statistical techniques. Most state of the art methods are based on the observation that similar opinion words mostly appear together in a corpus. Thus, if two words appear together frequently within the same context, then there is high probability that they have same polarity. Therefore, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word. This could be done using Point wise Mutual Information (PMI)as in example suggested by [11], SO of a given phrase is calculated by comparing its similarity to a positive word ("Awesome") And its similarity with negative word ("Awful"). More explicitly, a phrase is given a numerical rating by taking the mutual information between the given phrase and the positive reference word "Awesome" and subtracting the mutual information between the given phrase and the negative reference word "Awful". Using part-of-speech (POS) patterns, this technique then classifies the text by extracting the bigrams. PMI is then calculated by using the polarity score for each bigram.

D) Semantic approach

This principle assigns similar sentiment values to semantically close words. These Semantically close words can be obtained by getting the list of sentiment words and iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word.

Combining Lexicon and Learning Based Approaches For Concept-Level Sentiment Analysis

A concept-level sentiment analysis system called pSenti which combines lexicon based and learning based approaches. It measures and reports the overall sentiment of a review through a score that can be positive, negative or neutral or 1/5 stars classification. The main advantages and main interests of this article are the lexicon/learning symbiosis, the detection and measurement of sentiments at the concept level and the lesser sensitivity to changes in topic domain. It operates in four parts. First, the pre-processing of the review where the

noise (idioms and emoticons) is removed and each word is tagged and stored by the method Part Of Speech (POS). Second, the aspects and views are extracted to generate a list of top 100 aspect groups and top 100 views. The aspects are identified as nouns and noun phrases, and the views as sentiment words, adjectives and known sentiment words which occur near an aspect. Then the lexicon-based approach is used to give a \ sentiment value" to any sentiment word and generates features for the supervised machine learning algorithm.

To evaluate this method, experiments were conducted on two datasets: software reviews (more than 10,000) and movie reviews (7,000). Software reviews were separated into two categories: software editor reviews and customer software reviews. As a result, pSenti's accuracy was proved close to the pure learning-based system and higher than the pure lexicon-based method. It was also shown that the performance was not as good on customer software reviews as on software editor reviews because customer software reviews are usually much \noisier" (with comments that are irrelevant for the subject) than professional software editor reviews. Its accuracy was also affected by a large number of reviews for which it failed to detect any sentiment or assigned neutral score. However, the sentiment seperability in movie reviews was much lower than in software reviews. One of the reasons is that many movie reviews contain plots description and many quotes from the movie where words are identified as sentiments by the system.

IV. EXPERIMENTAL ANALYSIS

We collected data from Twitter with the help of Twitter API. We made use of both Twitter REST API v1.1 as well as Streaming API for different kinds of data collections. We used the Twitter's Streaming API for the collection of tweets. The method statuses/filter returns the public statuses containing one or more filter predicates.These predicates were a lists of keywords related to elections. These lists were manually prepared by us and we tried to be as exhaustive as possible.

Hashtag	# Tweets	Hashtag	# Tweets	Hashtag	# Tweets
#election2016	422,952	*#imwithher*	1,359,332	*#trump*	3,290,636
#elections2016	21,482	*#trump2016*	1,276,703	*#garyjohnson*	58,832
#tcot	1,027,036	*#nevertrump*	746,430	*#jillstein*	51,831
#p2	327,204	*#neverhillary*	1,063,545	*#jillnothill*	100,720
#hillaryclinton	962,054	*#trumppence16*	939,432	*#debatenight*	2,204,127
#donaldtrump	489,835	*#hillary*	1,516,318	*#debates*	870,744
#presidentialdebate	110,992	*#trumpwon*	349,118	*#VPDebate*	1,031,972
#debates2016	370,040	*#debate*	513,368		

Table 1. List of search terms we continuously monitored via the *Twitter Search API*

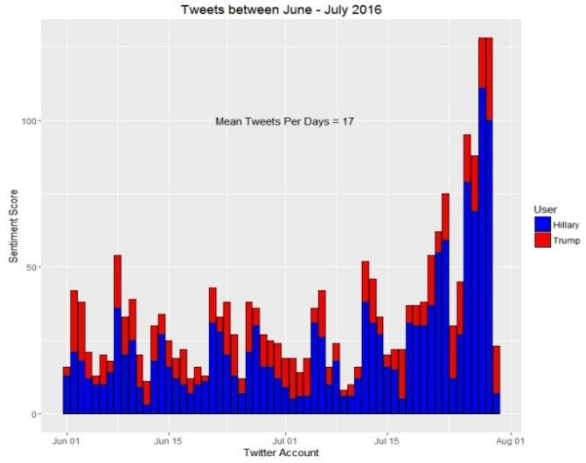
Collecting Data about Trending Politicians

We wanted to keep a track of the twitter accounts of the most trending politicians and the official accounts of their parties, if any. We used the twitter. Look up method of the Streaming API to collect this data. This method returned the JSON object for the users and provided information such as screen name, location, profile image url, followers count, friends count, statuses count and many other useful pieces of information about that profile. We handpicked 130 legitimate twitter profiles of some important politicians and parties. This list included those profiles which were actively using twitter or were of extreme importance in the national politics. We tried to prepare the list as exhaustively as we could.

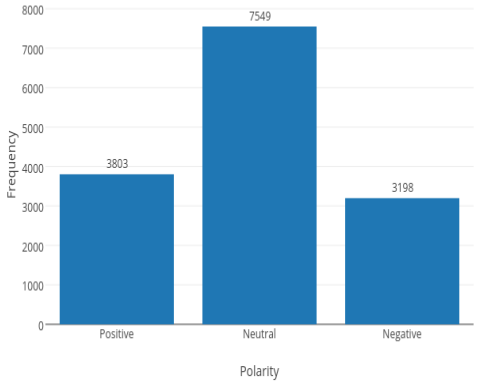
Data Analysis

The data collection had been on since September 2013. Apart from the daily analysis, we wanted to analyze the overall collection of data to see if there were any interesting patterns in them. As per the definition of data mining, we wanted to be able to give it a structure that is more understandable. To this effect, we used various tools to plot the graphs and draw information from the data. With the help of the graphs and other infographics, we wanted to be able to answer some of the research questions. We developed graphs to monitor the daily and hourly volume analysis. We wanted to find out that which parties got maximum mentions over the

period of time. Who were the people who were mentioned the most and how did the popularity of some leaders rise over the period of time were some of the questions we tried answering. We then related all these graphs with the political time line and drew certain inferences that have been discussed.



US election 2016 Sentiment Analysis



V. CONCLUSION

Scoring individuals by their political leaning is a fundamental research question in computational political science. From roll calls to newspapers, and then to blogs and microblogs, researchers have been exploring ways to use. Runtime of our algorithm with increasing N bigger and bigger data for political leaning inference. But new challenges arise in how one can exploit the structure of the data, because bigger often means noisier and sparser. Here, we assume: (a) Twitter users tend to tweet and retweet consistently, and (b) similar Twitter users tend to be retweeted by similar sets of audience, to develop a convex optimization-based political leaning inference technique that is simple, efficient and intuitive. Our method is evaluated on a large dataset of 119 million U.S. election-related tweets collected over seven months, and using manually constructed ground truth labels, we found it to outperform many baseline algorithms. With its reliability validated, we applied it to quantify a set of prominent retweet sources, and then propagated their political leaning to a larger set of ordinary Twitter users and hashtags. The temporal dynamics of political leaning and polarization were also studied.

We believe this is the first systematic step in this type of approaches in quantifying Twitter users' behavior. The Retweet matrix and retweet average scores can be used to develop new models and algorithms to analyze more complex tweet-and-retweet features. Our optimization framework can readily be adapted to incorporate other types of information. The y vector does not need to be computed from sentiment analysis of tweets, but can be built from exogenous information (e.g., poll results) to match the opinions of the retweet population. Similarly, the A matrix, currently built with each row corresponding to one event, can be made to correspond to other groupings of tweets, such as by economic or diplomatic issues. The W matrix can be constructed from other types of network data or similarity measures. Our methodology is also applicable to other OSNs with retweet-like endorsement mechanisms, such as Facebook and YouTube with "like" functionality.

REFERENCES

[1]. Aramaki, E., Maskawa, S., and Morita, M. Twitter catches the flu: detecting influenza epidemics using twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2011), Association for Computational Linguistics, pp. 1568-1576.

[2]. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, 10 (2008), P10008.

[3]. Bollen, J., Mao, H., and Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In ICWSM (2011).

[4]. Boyd, D., Golder, S., and Lotan, G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (2010), IEEE, pp. 1-10.

[5]. Bruns, A., and Burgess, J. E. The use of twitter hashtags in the formation of ad hoc publics.

[6]. Carletta, J. Assessing agreement on classification tasks: the kappa statistic. Computational linguistics 22, 2 (1996), 249-254.

[7]. Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (2011), ACM, pp. 675-684.

[8]. Cohen, R., and Ruths, D. Classifying political orientation on twitter: Its not easy! In Proceedings of the 7th International Conference on Weblogs and Social Media (2013).

[9]. Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. Predicting the political alignment of twitter users. In Privacy, security, risk and trust (pas-sat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (2011), IEEE, pp. 192-199.

[10]. Fruchterman, T. M., and Reingold, E. M. Graph drawing by force-directed placement. Software: Practice and experience 21, 11 (1991), 1129-1164.

[11]. Gayo-Avello, D. Don't turn social media into another'literary digest'poll. Communications of the ACM 54, 10 (2011), 121-128.

[12]. Golbeck, J., and Hansen, D. Computing political preference among twitter followers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011), ACM, pp. 1105-1108.

[13]. Gupta, A., and Kumaraguru, P. Credibility ranking of tweets during high impact events. In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (2012), ACM, p. 2.

[14]. Himelboim, I., McCreery, S., and Smith, M. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. Journal of Computer-Mediated Communication 18, 2 (2013), 40-60.

[15]. Honey, C., and Herring, S. C. Beyond microblogging: Conversation and collaboration via twitter. In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on (2009), IEEE, pp. 1-10.

[16]. Hong, S., and Nadler, D. Does the early bird move the polls?: the use of the social media tool'twitter'by us politicians and its impact on public opinion. In Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (2011), ACM, pp. 182-186.

[17]. Jungherr, A., Jürgens, P., and Schoen, H. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im predicting elections with twitter: What 140 characters reveal about political sentiment. Social Science Computer Review 30, 2 (2012), 229-234.

[18]. Krenn, B., Evert, S., and Zinsmeister, H. Determining intercoder agreement for a collocation identification task. In Proceedings of KONVENS (2004), pp. 89-96.

[19]. Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. How (not) to predict elections. In Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom) (2011), IEEE, pp. 165-171.

[20]. Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. Is the sample good enough? comparing data from twitters streaming api with twitters firehose. Proceedings of ICWSM (2013).

[21]. Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P. T. Vocal minority versus silent majority: Discovering the opionions of the long tail. In Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (2011), IEEE, pp. 103-110.

[22]. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. Physical review E 74, 3 (2006), 036104.

[23]. Nowak, A., Szamrej, J., and Latané, B. From private attitude to public opinion: A dynamic theory of social impact. Psychological Review 97, 3 (1990), 362.

[24]. Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (2010), ACM, pp. 851-860.

[25]. Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., and Jiang, J. Tweets and votes: A study of the 2011 singapore general election. In System Science (HICSS), 2012 45th Hawaii International Conference on (2012), IEEE, pp. 2583-2591.

[26]. Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM 10 (2010), 178-185.

[27]. Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Election forecasts with twitter how 140 characters reflect the political landscape. Social Science Computer Review 29, 4 (2011), 402-418.

[28]. Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L. Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) 26, 3 (2008), 13.