Impact of Social Media on Democratic Decision Making

Eliott Chapuis, Arjun Chintapalli, Joy Kimmel, Gowri Nayar, M.S. Suraj, Jia Yi Yan
College of Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
{echapuis, arjun.ch, joykimmel, gnayar, mssuraj, jiayi.yan}@gatech.edu

***Abstract*—We analyzed Twitter and Reddit data through word frequency and natural language processing techniques in order to determine voter sentiment and predict results for both Brexit and the 2014 Indian General Elections.**

I. Introduction

In recent years, we have seen a prevalence of unexpected election results. Specifically in regards to the Indian election of 2014 and the British European Union referendum in June 23 2016, known as Brexit, the election results did not correspond to the portrayal of the situation by the media and the polling data. These circumstances have led us to question whether these traditional methods are the most useful in predicting results. It has been shown that social media is becoming a more prevalent platform to engage with political debates, and thus we will investigate whether social media data has more predictive power than traditional methods.

The failure of traditional polling methods to predict elections has led our group to analyze the predictive power of sentiment analysis to track public opinion and election outcomes, in regards to the Indian General Election 2014 and the Brexit Refferendum. We have sought to address this problem through the use of natural language processing on social media to predict election results and track public opinion.

II. Related Work

Our literature review consists of prior data analysis of the Indian election, Brexit election and current analysis methods.

A. Indian Parliamentary Election of May 2014

This election demonstrated an unprecedented use of social media amongst the Indian electorate to vent their viewpoints as well as for campaigning parties [23].

29 million Indians made 227 million social media actions regarding the Indian Lok Sabha elections on Facebook out of an estimated 110 million with internet connection. Out of these posts, 13 million people made 75 million interactions regarding Narendra Modi, the eventual winner of the election [17].

Prior Twitter LDA analysis reveals that the BJP replied to tweets and addressed concerns, while the other parties used Twitter for criticism [1]. Further evidence in support of BJP's Twitter strategy is found in the strong positive correlation between the proportion of first time electors and change in BJP's vote share [5]. The above groups just summarized raw data or manually categorized tweets into functional categories; our group will use sentiment analysis to further determine the predictive power of Twitter.

B. British Referendum of 2016 (Brexit)

Many studies discuss the declining usefulness of political polls. Though social media does not provide as broad representation as traditional polls, social media data shows the "trending" topics that are important to the public in upcoming elections [2]. Even simple analysis on the rate of tweets outperforms traditional polling methods as a predictive tool [3]. We often turn to social media as a news source, retweeting in agreement and mentioning in disagreement [6]. In regards to Brexit, two politically stratified groups emerged with minimal content crossover [7]. Using these ideas, we will analyze sentiment surrounding issues discussed on social media against polling data.

C. Current Methods in Text Analysis

Phrase and sentiment analysis is crucial in forming conclusions about public opinion [13]. A "gold standard" presents a set of topics and sentiments that organically arose, uninfluenced by campaign officials, as the topics gain traction in real-time [16]. We use these "gold standards" as reference to determine important phrases in the political spectrum. Reinforcement of in-group political beliefs occur because like-minded users cluster together, so analyzing the network of users that produces the stratification is crucial [11].

Many methods are used to classify social media data. Some argue that classifying agreement is more useful than classifying sentiment [8]. We can develop trends using support vector regressors to classify agreements as positive or negative. Another method combines a Twitter user's data with the user's retweets, used as a sign of endorsement [9]. We must also account for the activity of political "bots", the source of one-third of conservative Twitter traffic, compared to one-fifth of liberal Twitter traffic [15]. Particularly during Brexit, "bots" accounted for the majority of separation supporting tweets [12]. We should separate the automated tweets and analyze the influence of these bots.

Our analysis relies on keywords and sentiment analysis of these phrases. Tumasjan et al. argue the number of mentions for any political party is a good proxy for voting percentage, which we will use with analysis of sentiment words, as described by O'Connor et al., to increase the correlation to

80% [22], [18]. Taboada et al. describe methods to extract sentiment from a dictionary of words, including classifying intensity of feeling, methods we will use to classify tweets as liberal or conservative [21]. Similar techniques, performed manually, were used in classifying sentiment around the GMO debate to much success [20]. We will automate their process, as machine learning techniques in sentiment analysis are more accurate than human based computations [19]. However, Pang et al. found that machine learning tools performed better on topic classification than sentiment classification, which we will improve upon. We will classify both word sentiment and the combination of these words, as semantic context is important [14].

III. Method

Given the current state of research on both political analysis, our project differs in several ways. A list of the contributions made in this project include:

- A direct comparison between social media data from Twitter and Reddit and traditional polling data. This approach is new and innovative in that it attempts to combat the inadequacies of social media by combining both Reddit and Twitter, and surpasses polling data with more information from social media.
- An examination of the predictive power of social media across multiple cultures. By performing this analysis over different cultures, we can highlight any variations that may arise between the countries' user patterns, and in the future tailor the analysis to specific countries or even demographic groups.
- We are creating our own unique sentiment score created using a combination of natural language processing techniques. Further, by determining this score's correlation with the polling data for the two cases of Brexit and the India election, we will be able to gauge the event-specific and overall capability of the sentiment score to predict the election outcomes.

Given these specified contributions, we believe that our project is both meaningful and educational. In the following proposed work we first outline our data gathering and preprocessing methods followed by an examination of our text analysis process.

A. Data Gathering

We gathered data from Reddit, Twitter using the PRAW API [4] for Reddit data and Get Old Tweets API [10] for Twitter. Additionally, we gathered polling data from sources such as BBC, Wikipedia, and India Today. The scope of our data spanned the six months leading up to the Indian election and Brexit Refferendum. We acheived a total of 2,207,976 tweets and Reddit posts.

We used the PRAW API to search subreddits on Reddit based on key terms and timestamp ranges and compiled all submissions, comments, upvotes and author karma for future analysis. Our Reddit search using the Indian Election terms in the Appendix lead to 1,939 submisions and 45,439 comments. Simlarily for the Brexit election, we found 1506 submissions and 76,698 comments.

We collected a total of 2,097,096 tweets. This included 919,836 tweets for Brexit and 1,177,260 for the India election. Keywords used to gather the tweets were based on [13] and can be found in the Appendix.

B. Data Pre-Processing

After collecting all of the data, we converted the Twitter tweets and Reddit comments (collectively: 'the text') to vector representations by training Google's Word2Vec model using the text data [26]. This model uses multiple shallow neural networks to determine quantitative relationships between words located near each other. Effectively, it converts words to vectors, allowing us to perform interesting tasks such as clustering. For training our Word2Vec model, we used a feature length of 300, a 10 word context, a minimum word count of 20, and removed stop words. A different model was trained for each of the 4 datasets (twitter-brexit, twitter-india, reddit-brexit, reddit-india) that we collected.

C. Clustering

After converting all of the words in our text to vector representations, we conducted k-means clustering separately on each dataset and found the top 100 word clusters. After clustering all of the words in each corpus, we were able encode the text using a "bag of centroids" representation. This representation consists of a 100-length vector where each item is the count of words in that cluster found in the text. We then manually determined the relevent clusters. Using the bag of centroids representation and the relevant clusters, we were able to classify how relevant each tweet was based on how many words came from relevant clusters. We plot the relavency profiles for the different datasets below:

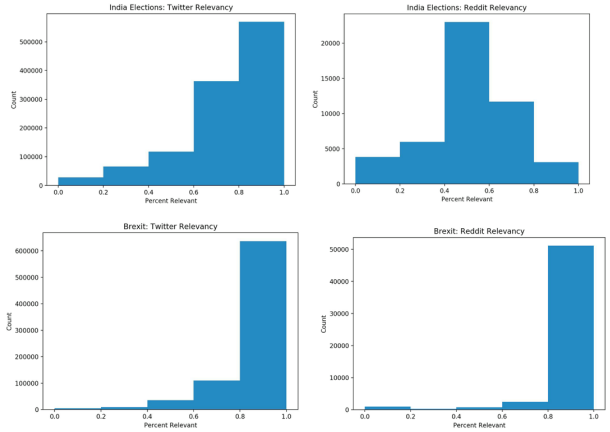

Fig. 1: Relevancy of Datasets

It can be seen in the above figure that Indian posts in general are less relevant to the topic than Brexit posts.

Using our relevancy metric, we were able to filter our data further, using the tweets/comments with the top 25% relevancy. Below, we show the word cloud results of our clusters, where we visualize the most frequent words of the top ten clusters in each dataset. The graphs below show the words proportionate in size to their count and colored by

the cluster they are classified in. We analyzed the percentage of words belonging to relevant clusters per tweet/Reddit comment, and observed that similar to before Indian words are less relevant to the topic than Brexit words.

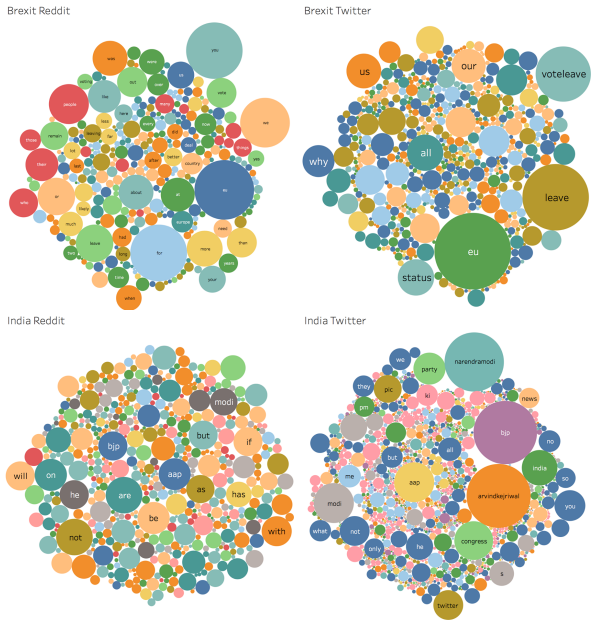

Fig. 2: Word Cloud Results

D. NLP Analysis

Our Natural Language Processing analysis involved Sentiment Analysis on Twitter data and Political Analysis on Reddit Data. After filtering using k-means clustering, we proceeded to derive sentiment scores for each post in order to categorize them.

For the Twitter data, we pulled the data based on key words and so we classified the key words into liberal, conservative, and neutral terms, and then used the Indico API [27] to get a score for each tweet. This score indicates the positive or negative sentiment in the tweet, so scores above 0.5 were considered positive and below, negative. Therefore, positive liberal and negative conservative were classified as liberal, favoring stay in the EU for Brexit and the Congress party for the Indian election. Negative liberal and positive conservative were classified as conservative, favoring leave the EU and the BJP party for the Indian election. We count the number of liberal and conservative tweets for each day and plot this over time.

For the Reddit data, we did not have the keywords available, and thus we performed political analysis, rather than sentiment analysis, using the Indico API. This returned a score for the percent the post is conservative and liberal, so we take the highest value and classify the post in that category. We categorized Reddit data into conservative and liberal over time for India and Brexit, where conservative relates to BJP party for India and leave for Brexit, liberal to Congress and stay. Again we count the number of conservative and liberal post and plot this over time.

IV. Experiments/Evaluation

In this section we have included the graphs that summarize the results of our Twitter, Reddit and polling data.

A. Polling Data Results

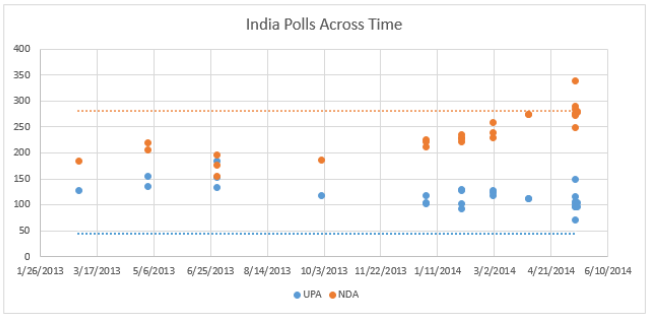

Fig. 3: Indian General Election Polling Data

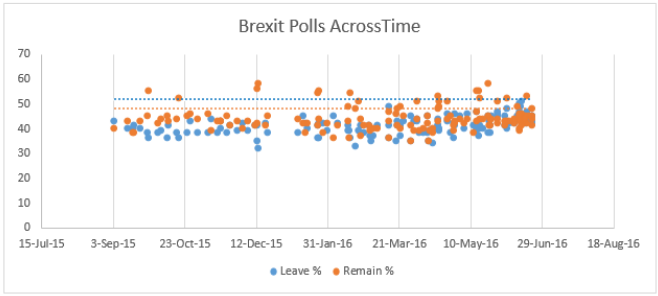

Fig. 4: Brexit Polling Data

We gathered polling data for the Brexit and the India General election to validate our methods and analysis. In these plots, the final electoral result for each election is provided as a dashed line across the time range. For the Indian election, polls for the two leading coalitions were provided, and the cluster of polls at the end are the exit polls. Through the length of the Brexit polling, the Stay vote was predicted to win whereas in actuality Leave won the referendum. The polling data displayed clearly demonstrate that the polls in both cases were very far from the final election outcome, which we hope to better predict through our models.

B. Brexit Results

In this section we provide the results of our Sentiment Analysis on the Brexit referendum using Reddit and Twitter data. We plot the sentiment leading upto the Brexit referendum date in order to determine the sentiment of voters going into the referendum.

The results from using Reddit Data to gauge sentiment for Brexit indicate that the result of the election would be to stay, whereas the actual result was ”Leave”. Additionally, the frequency of comments over time increased greatly as the Referendum date got closer.

In contrast, Twitter data indicated that the result of the

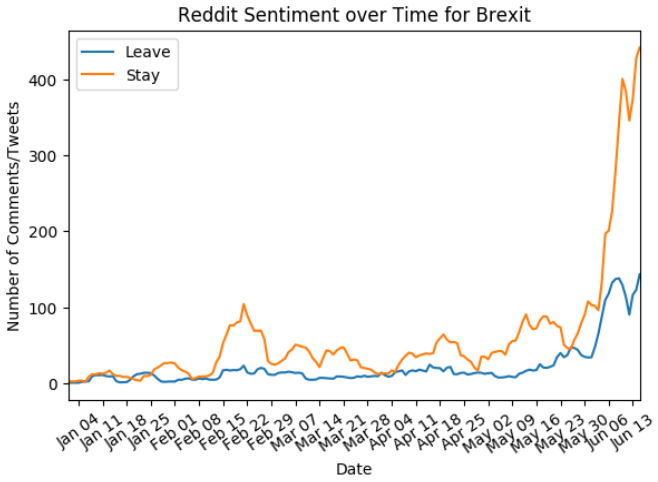

Fig. 5: Brexit Results using Reddit Data

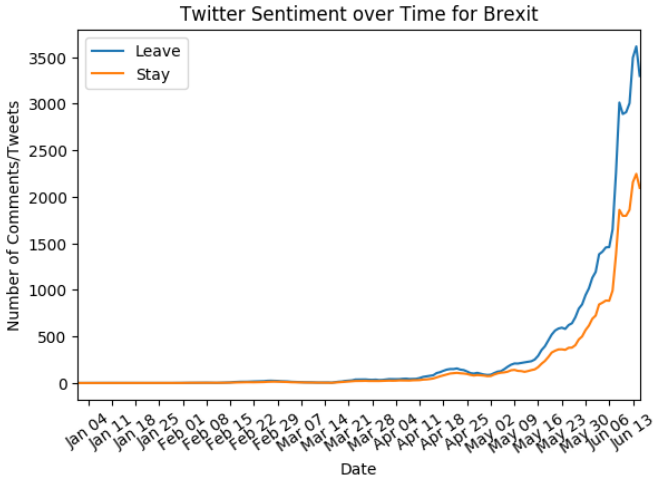

Fig. 6: Brexit Results using Twitter Data

referendum would be to vote ”Leave”, which was the actual outcome. In this way, using Twitter data, we were able to accurately predict the outcome of the British Referendum.

C. Indian General Election Results

This section includes the results of our Sentiment Analysis on the Indian General Elections. The two main parties for the election Bharitiya Janata Party and Congress Party were classified as ”Conservative” and ”Liberal” respectively, and so we plotted the liberal and conservative sentiment over time to indicate which party was favored most on social media. We plotted the 6 months leading up to the start of phase 1 of the Indian General election, which started in April 7, in order to determine the sentiment of voters going into the election.

The results from using Reddit Data to gauge sentiment for the Indian General Election indicate that the liberal Congress party was going to win the election, whereas the actual result was that the conservative BJP party won. Therefore, using Reddit data, we were unable to accurately predict the result

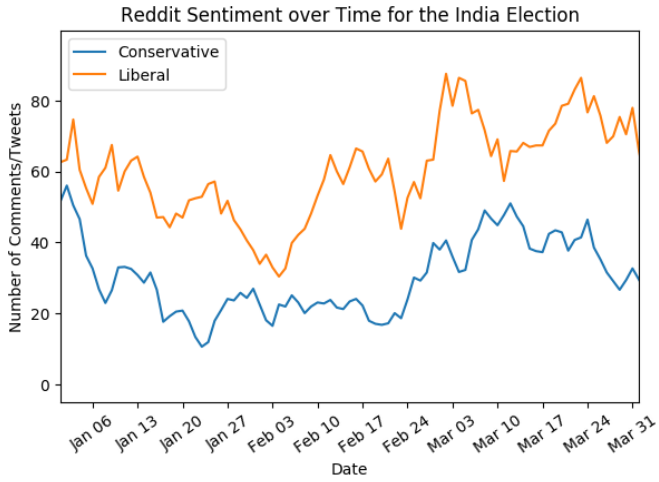

Fig. 7: Indian General Election Results using Reddit Data

of the Indian General Election.

The results from using Twitter Data to gauge sentiment

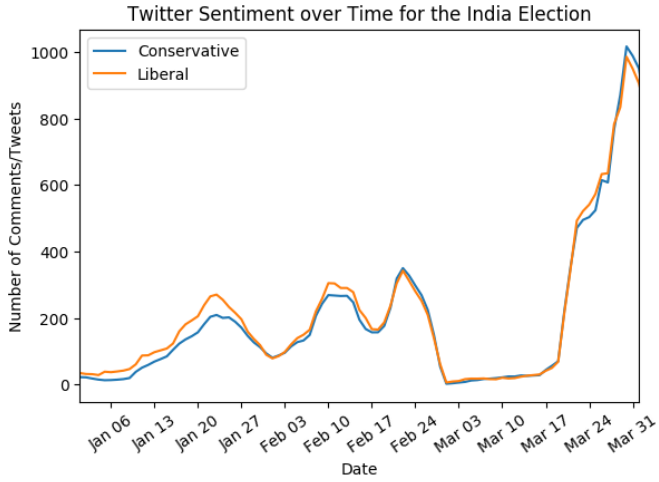

Fig. 8: Indian General Election Data using Twitter Data

for the Indian General Election indicate that conservative BJP party was going to win the election, which was the actual outcome of the election. Therefore, using Twitter data, we were able to accurately predict the result of the Indian General Election. The caveat though is that although the Twitter data shows an even split between the conservative and liberal parties, the conservative BJP party overwhelmingly won the elections.

Both the Reddit and Twitter datasets for the Indian election were dramatically smaller than for Brexit. This is easily noticed by comparing the y-axis scales for associated plots.

V. Conclusion

In this project, we have demonstrated that our sentiment and political analysis proves to be an alternate pathway to bet-

ter analyze and poll people's political opinions. Using Twitter data, we were able to accurately predict the Brexit outcome of "Leave" and the Indian General Election conservative win. In contrast, the Reddit data was not as reliable for gauging sentiment, which was likely due to the reduced amount of data available as well as liberal bias in Reddit users.

Current results show an overwhelming younger and liberal bias in social media, thereby demonstrating that the same pitfalls with polling data also apply to social media data. However, filtering by relevancy helped reduced this bias.

An additional limitation to our results is the limited penetration of Twitter and Reddit in India and Britain. At the time of the Indian election there were only 18 million Twitter users online as compared to the 800 million elgible voters [24]. In comparison, for Brexit there were 15.8 million Twitter users and 33.6 million voters. [25]. Moving forward, if statistics on the percent of Twitter and Reddit users that are liberal or conservative could be found, then we could weigh our results given the beliefs of the offline population to get more predictive results.

Further work is needed to get more representative datasets that include older and more conservative populations. Examples of such datasets could include Facebook and Linkedin. We could also alleviate this bias by using a weighting factor to shrink the weight of liberal groups and increase the weightage of conservative groups to be in sync with the wider population.

VI. Distribution of Effort

All team members have contributed an equal amount of effort towards this report.

References

[1] Ahmed, Saifuddin, Kokil Jaidka, and Jaeho Cho. "The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties." *Telematics and Informatics* 33.4 (2016): 1071-1087.

[2] Anstead, Nick, and Ben O'Loughlin. "Social media analysis and public opinion: The 2010 UK general election." *Journal of Computer-Mediated Communication* 20.2 (2015): 204-220.

[3] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media." *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2010. 492–499. ACM Digital Library. Web. 6 Mar. 2017. WI-IAT '10.

[4] Boe, Bryce. "PRAW: The Python Reddit API Wrapper — PRAW 4.4.1.dev0 Documentation." N.p., n.d. Web. 28 Apr. 2017.

[5] Chadha, Kalyani, and Pallavi Guha. "The Bharatiya Janata Party's Online Campaign and Citizen Involvement in India's 2014 Election." *International Journal of Communication* 10 (2016): 18.

[6] Conover, M. D.; Ratkiewicz, J.; et al. "Political Polarization on Twitter," Center for Complex Networks and Systems Research, 2011.

[7] Del Vicario, Michela et al. "The Anatomy of Brexit Debate on Facebook." arXiv:1610.06809 [cs] (2016): n. pag. arXiv.org. Web. 6 Mar. 2017.

[8] Fabio Celli, Evgeny A. Stepanov, Massimo Poesio, Giuseppe Riccardi. "Predicting Brexit: Classifying Agreement is Better than Sentiment and Pollsters". Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 110–118, Osaka, Japan, December 12 2016.

[9] Grčar M, Cherepnalkoski D, Mozetič I. The Hirsch index for Twitter: Influential proponents and opponents of Brexit. In: Proc. 5th Intl. Workshop on Complex Networks and their Applications. Studies in Computational Intelligence. Springer; 2016.

[10] Henrique, Jefferson. "Jefferson-Henrique/GetOldTweets-Python." GitHub. N.p., n.d. Web. 18 Apr. 2017.

[11] Himelboim, Itai; McCreery, Stephen; Smith, Marc. "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter." *Journal of Computer-Mediated Communication*, January 2013, Vol. 18, No. 2, 40-60.

[12] Howard, Philip N., and Bence Kollanyi. "Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum." arXiv:1606.06356 [physics] (2016): n. pag. arXiv.org. Web. 6 Mar. 2017.

[13] Hürlimann, Manuela et al. "A Twitter Sentiment Gold Standard for the Brexit Referendum." *Proceedings of the 12th International Conference on Semantic Systems*. New York, NY, USA: ACM, 2016. 193–196. ACM Digital Library. Web. 6 Mar. 2017. SEMANTiCS 2016.

[14] Kim, Soo-Min, and Eduard Hovy. "Determining the Sentiment of Opinions." *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. ACM Digital Library. Web. 6 Mar. 2017. COLING '04.

[15] Kollanyi, Bence, Philip N. Howard, and Samuel C. Woolley. Bots and automation over Twitter during the first US Presidential debate. COMPROP Data Memo, 2016.

[16] Llewellyn, Clare, and Laura Cram. "Brexit? Analyzing Opinion on the UK-EU Referendum within Twitter." *Tenth International AAAI Conference on Web and Social Media*. N.p., 2016. www.aaai.org. Web. 6 Mar. 2017.

[17] Narasimhamurthy, N. "Use and Rise of Social media as Election Campaign medium in India." (2014).

[18] O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." ICWSM 11.122-129 (2010): 1-2.

[19] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques." *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. 79–86. ACM Digital Library. Web. 6 Mar. 2017. EMNLP '02.

[20] Rivera, Ivan, Jim Warren, and James Curran. "Quantifying Mood, Content and Dynamics of Health Forums." *Proceedings of the Australasian Computer Science Week Multiconference*. New York, NY, USA: ACM, 2016. 67:1–67:10. ACM Digital Library. Web. 6 Mar. 2017. ACSW '16.

[21] Taboada, Maite et al. "Lexicon-Based Methods for Sentiment Analysis." *Comput. Linguist.* 37.2 (2011): 267–307. ACM Digital Library. Web.

[22] Tumasjan, Andranik, et al. "Predicting elections with Twitter: What 140 characters reveal about political sentiment." ICWSM 10.1 (2010): 178-185.

[23] Wani, Gayatri, and Nilesh Alone. "A Survey on Impact of Social Media on Election System." *International Journal of Computer Science and Information Technologies* 5.6 (2014): 7363-7366.

[24] India, P. T. (2014, May 28). India to have third-largest Twitter population by 2014: eMarketer. Retrieved April 27, 2017, from http://indianexpress.com/article/india/politics/india-to-have-third-largest-Twitter-population-by-2014-emarketer/

[25] UK Twitter users 2012-2018 — Forecast. Retrieved April 27, 2017, from https://www.statista.com/statistics/271350/twitter-users-in-the-united-kingdom-uk/

[26] "Google Code Archive - Long-Term Storage for Google Code Project Hosting." N.p., n.d. Web. 28 Apr. 2017.

[27] "Indico — Text and Image Analysis Powered by Machine Learning." N.p., n.d. Web. 28 Apr. 2017.

Appendix A
Reddit India Keywords

"Modi"; "BJP"; "Election"; "Congress";"NaMo"; "election"; "AAP"; "Lok Sabha"

Appendix B
Reddit Brexit Keywords

ukpolitics, europe, AskReddit

APPENDIX C
TWITTER BREXIT KEYWORDS AS DEFINED IN [13]

brexit, beleave, betteroffin, betteroffout, Bremain, Brexit, brexitfears, britainout, britin, EdEURef,eukay, eunegotiation, EUpol, EUpoll, euref, eureferendum, eureform, eurenegotiation, europeanunion, fudgeoff, grassrootsout, greenerin, ImagineEurope, LabourIn, leadnotleave, leafchaos, leaveeu, loveeuropeleaveeu, MyImageOfTheEU, no2eu, notoeu, projectfact, projectfear, ref, referendum, remain, remaineu, saferbritain, StrongerIn, theinvisableman, theknoweu, UKandEU, ukineu, UKRef, UKreferendum, votein, voteleave, voteout, voteremain, wrongthenwrongnow, yes2eu, yestoeu, Davidcameron, INtogether, TakeControl, euin , euout, NoEu, @vote_leave, @Vote_LeaveMedia, @StrongerIn, @StrongerInPress, @britinfluence, @lsebrexitvote, @eureferendum, @LeaveEUOfficial, @whatukthinks, @JuneExit, @EUinEUout, @Grassroots_out, @euromove, @UKandEU, @sayyes2europe, @Choice4Britain, @BrexitWatch

APPENDIX D
TWITTER INDIA KEYWORDS

Rahul Gandhi; Narendra Modi; BJP; Congress Party; UPA; Liberals; Conservative; Facism; Hindu; Muslim; elections; INC ; AAP; @AamAadmiParty's; @ArvindKejriwal; @BJP4India's; @NarendraModi; @INCIndia; YourVote2014; Lok Sabha Elections 2014 ; Elections2014; AAP; NaMo; BJP; PMPressMeet; Kejriwal; Modi; NaMoinGoa; QuitAAP; AAP-Drama; AICCMeet; RahulSpeaksToArnab; NaMoinKolkata; NaMoinMeetur; BhagodaKejri; AnarchistCongBJP; RGforSoldiers; RGinAssam; RGforEducation; NaxalAAP; AKinGujarat; AAPtards; AAPwedsAajTak; KattarSchoNahiYuvaJosh; ChaloVaranasi; GharGharModi; KejriwalinVaranasi