Big social data and political sentiment: the tweet stream during the UK General Election 2015 campaign

The 8th IEEE International Conference on Social Computing and Networking (SocialCom 2015), Chengdu, China, Dec. 19-21, 2015. (in press)

Big Social Data and Political Sentiment: the Tweet Stream during the UK General Election 2015 Campaign

Giuseppe Di Fatta
School of Systems Engineering
University of Reading
Reading, Berkshire, RG6 6AH, United Kingdom
Email: G.DiFatta@reading.ac.uk

J. James Reade
School of Politics, Economics and International Relations
University of Reading
Reading, Berkshire, RG6 6AH, United Kingdom
Email: j.j.reade@reading.ac.uk

Sylvia Jaworska
Department of English Language and Applied Linguistics
University of Reading
Reading, Berkshire, RG6 6AH, United Kingdom
Email: s.jaworska@reading.ac.uk

Anupam Nanda
Henley Business School
University of Reading
Reading, Berkshire, RG6 6AH, United Kingdom
Email: a.nanda@reading.ac.uk

Abstract

The General Election for the 56th United Kingdom Parliament was held on 7 May 2015. Tweets related to UK politics, not only those with the specific hashtag "#GE2015", have been collected in the period between March 1 and May 31, 2015. The resulting dataset contains over 28 million tweets for a total of 118 GB in uncompressed format or 15 GB in compressed format. This study describes the method that was used to collect the tweets and presents some analysis, including a political sentiment index, and outlines interesting research directions on Big Social Data based on Twitter microblogging.

Index Terms

Big social data, Twitter, microblogging, political sentiment index, #GE2015

I. Introduction

Big social data analytics arises from a big data approach [1], [2] to online social media [3] and offers the opportunity to social scientists to understand much more about social phenomena and their impact that until now have been difficult or impossible to measure. The most salient source of big data in the social sciences is social media data: electronic, online (hence readily recorded) versions of social (and more formal) networks that have always existed. The correspondence between offline and online social networks is, of course, not one-to-one; not least, the ability to publicise information to a vast, global audience is a feature not common to offline networks. That is, however, an additional appealing characteristic for research in the social sciences where the role of information and its spread is of particular importance in understanding outcomes and possible impacts. As such big social data analytics offer a promising research opportunity for social sciences.

In particular, Twitter is a microblogging service that was launched in 2006 and has been recently reported to have over 300 million monthly active users worldwide. In the UK it is estimated that there are 14/15 million monthly active users, which is about 22% of the UK population.

In this paper we document a method for data collection from Twitter for both real-time and ex post analysis surrounding the UK General Election 2015. A political sentiment index is adopted to focus specifically on Twitter data generated during the main TV political debates. We were specifically interested in measuring the public attitudes or sentiments in response to the messages voiced by the individual party leaders during the debates to identify key moments for further analysis.

We used a data feed using the Twitter streaming API to collect tweets satisfying a set of criteria that were specified to capture tweets with UK political content. Due to the general nature of a number of the search terms (e.g. 'Labour', 'Green'), we added context checks to ensure the tweets we collected were related to the general election. However, the adopted method is easily applicable to many other domains and events one may wish to track and analyse.

In total we collected about 28 million (28,473,893) tweets from Twitter associated to UK politics and the General Election 2015. We anticipate numerous further research strands will emerge from this dataset. In the social sciences, for example, it is well appreciated now that economic decisions such as willingness to buy are influenced by social networks and the information transmitted within them, and hence understanding how such networks form, and their characteristics, is very important. Furthermore, many events such as general elections have profound impacts on economic variables, and as such the potential for better, quicker and more frequently updated predictions of such outcomes using social media data is very

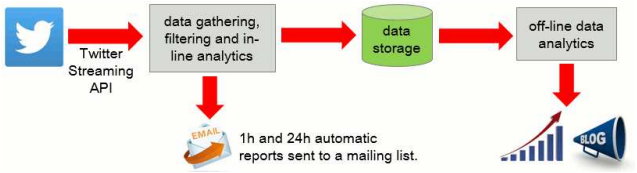


Fig. 1: Online and offline analytics process on Twitter data streams.

important. It is for this reason we make our UK General Election 2015 Twitter dataset available [4] to all who wish to use it for their research.

The rest of the paper is organised as follows. Section II presents a brief overview of related work. Section III provides a description of the method used to gather the data and to generate a political sentiment index. Section IV presents the dataset and the results of the analysis that was carried out. In Section V a brief discussion of possible use cases in other domains is provided. Finally, Section VI presents some conclusions and directions of future work.

II. Related Work

The study in [5] investigated the use of Twitter in the 2009 German federal election and, in particular, attempted to verify if tweets validly mirror offline political sentiment. The analysis was based on a dataset of over 100,000 messages containing a reference to either a political party or a politician.

Authors in [6] collected 1,150,000 tweets on the top trending topics from about 220,000 users related to the 2010 UK General Election and between the 5th and 12th of May 2010. They analysed these Twitter messages to identify both the characteristics of political parties and the political leaning of users.

An analysis of sentiment in Twitter messages with political content was studied in [7]. The work is based on a data set of 64,431 tweets related to two political elections in Germany in 2011. In particular, the work investigates the feature of 'retweeting' as a simple mechanism for information and opinion diffusion, which may help to increase political participation.

Authors in [8] used sentiment detection and tweet classification to predict election results of the Pakistan 2013 General Election. The work is based on 612,802 tweets associated to names of political parties and political celebrities.

The study in [9] analysed 460K tweets over three years for 687 candidates running for national House, Senate, or state governor seats in the 3.5 years leading to the elections. The data was augmented with over 690k documents by crawling outgoing links referred to by candidate tweets. The work analysed the differences in the usage patterns of social media and built a model to predict candidate victory.

Televised election debates are an interesting case in point, as they have been shown to have a decisive impact on voting participation and behaviour [10]. The increased use of social media sites, microblogging in particular, seems to reinforce the impact. For example, the study [5] on the 2009 German election campaign demonstrates that Twitter data can be, in fact, a good and reliable predictor of election outcomes. To our knowledge, there is no study that looked at the connection between microblogging and political debates in the UK. This is probably due to the fact that televised election debates are a rather new phenomenon on the British political scene.

III. Methodology

A. The Data Collection Process

An ad-hoc application in Java was developed to manage the retrieval of relevant tweets associated to UK politics during the three-month period of interest preceding and following the UK General Elections. The Twitter streaming API [11] was adopted to monitor any tweet related to UK politics in real time. A combination of tracked terms and ad-hoc filters for a political context check were used to identify the 'political' tweets of interests, which were a superset of the tweets containing the hashtag "#GE2015". The relevant terms were chosen by domain experts and were divided in four categories: terms with unambiguous reference to UK politics (e.g., ge2015, uklabour, scottishlabour, votelabour, ukip, voteukip, etc.), ambiguous terms (e.g., labour, greens, Cameron[1], etc.), terms for context check and rejected terms (e.g., Clinton, USA, Canada, America, TCOT). Tweets were tracked if they contain any of the unambiguous terms, or any of the ambiguous terms and at least one of the context terms. Tracked tweets were rejected if they contain any of the rejected terms in order to reduce the noise in the data generated by unrelated tweets containing tracked terms. During the campaign special TV events were broadcast. In the days when these TV events were scheduled, the official hashtag of the event was also tracked (e.g., #battlefornumber10). Especially during the first days of the tracked period the lists of terms were heuristically fine tuned with the help of domain

[1]Cameron Dallas is an 18-year-old celebrity of Vine, a short video sharing service and microblogging website. At the time of this study, he has over 4.6 million followers on Twitter; while the UK Prime Minister David Cameron has less than a million.

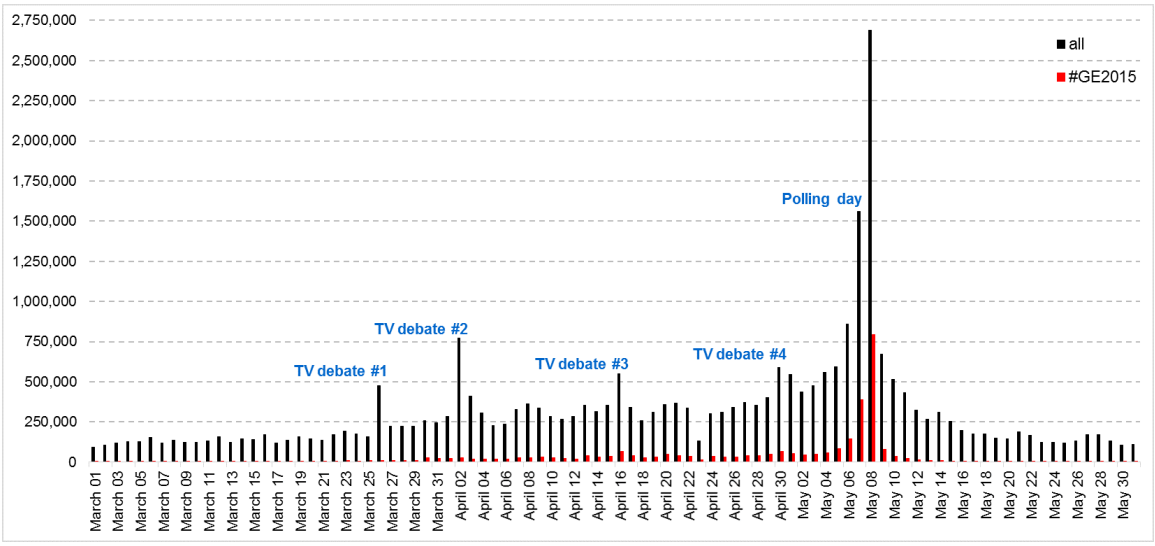


Fig. 2: Number of tweets per day from March 01 to May 31, 2015.

experts to find a good trade-off between precision and recall. On average the process adopted about 30 unambiguous terms, no more than 10 ambiguous terms, about 50 context terms and no more than 10 rejected terms.

Figure 1 shows a typical high level diagram of this process. The ad-hoc application for the retrieval process of relevant tweets has three concurrent threads of execution: a Consumer, a Controller and the Observer. The Consumer manages the stream of tweets for the tracked terms, receives and process tweets from Twitter in real time and stores them to a secondary memory. The Controller controls that the tweets Consumer is working properly and, if not, it starts a new Consumer. The Observer generates and sends periodic summaries (online analytics) by email to a list of project followers. Further analytics is generated off-line by additional processing, such as generation of counts, word clouds, co-occurrence of terms and sentiment index. During the campaign our findings were discussed and presented in a dedicated blog [4].

The specific sentiment index used in this study is briefly described in the following section.

B. A Political Sentiment Index

In order to assess the public attitude or mode, we created a specific political sentiment index that we applied to the collected data. We also used particular party-specific search terms to associate a particular tweet with one of the parties in the election, enabling us to analyse online moods [12], [13] surrounding that particular party at that point in time. The index, alongside the frequency of tweets collected, enabled us to identify salient moments during those events. This is of particular interest for predicting event outcomes such as elections.

Most studies analysing public sentiments adopt external list of words or dictionaries that are based on general language use and do not distinguish well between the positive or negative meanings that one word can have. Studies in pragmatics have shown that one word form can change it meaning from positive or negative or vice-versa depending on the context and the purpose for which it is used. A good and recent example from the political discourse is the term 'flamboyant', which in general language use will be assigned a positive score. When used in the political domain however, it is often associated with negative events such as 'flamboyant expenses' and thus, it acquires a negative meaning. Our index is based on evaluative words (mainly adjectives) that we retrieved from the data using the parser Penn Treebank[2] which automatically assigns a part-of-speech to each word in the text. We focus on adjectives, because these are evaluative words ('good', 'bad', 'positive', 'negative', 'happy' etc.) most likely to indicate mood. Once all adjectives were retrieved, the lists were scanned manually and each item was assigned a score: +1 for positive meanings, -1 for negative meanings and 0 for neutral. When the meaning was unclear, the tweets were examined to disambiguate the meaning. In this way and in contrast to general sentiment indices, our index was specifically data and context-driven.

IV. Collected Data and Analysis

A. The Dataset

The process described in Section III-A was used to collect tweets related to UK politics, not only those with the specific hashtag "#GE2015", in the period between March 1 and May 31, 2015. The resulting dataset contains over 28 million

[2]https://www.cis.upenn.edu/ treebank/

TABLE I: Number of recorded and missed tweets due to the track limit.

date	*recorded tweets*	*missed tweets*	*missed (%)*
April 2	772,763	175,959	18.5
April 16	548335	2,752	0.5
May 7	1,559,604	94,162	5.7
May 8	2,689,062	78,978	2.9

(28,473,893) tweets for a total of 118 GB in uncompressed format or about 15 GB in compressed format. Figure 2 shows the number of all tweets collected per day and those with hashtag #GE2015 over the entire period. The Twitter streaming API does not guarantee to report all tweets containing the tracked terms: it caps the traffic to 1% of the global traffic. However, it does report the number of tweets which are excluded from the live stream. We did experience this limit in some occasions and mostly for a negligible amount. In a few cases the number of missed tweets (i.e., that should have been reported and were not) was significant and these are reported in Table I.

During the campaign there were four live television programmes featuring the main political party leaders:

1) ”Cameron & Miliband: The Battle for Number 10” (#battlefornumber10) on March 26,
2) ”Leaders’ debate” (#leadersdebate) on April 2,
3) ”Challengers’ debate” (#challengersdebate) on April 16 and
4) ”Question Time special” (#bbcqt) on April 30.

These TV events, the polling day and the day after are clearly visible in the chart of Figure 2 because of the larger number of tweets they have induced. In particular, up to now the general election event (polling day and the day after) is the public event in the UK that has seen the largest participation on Twitter by far: the two days have accounted for 4,248,666 tweets, which is 15% of the total number of tweets collected over the entire period of three months.

In the next section we show the analysis of the tweet stream for one of these TV events by means of the political sentiment index described in Section III-B and tag clouds.

B. Sentiment Analysis during a TV Debate

The TV political debates seem to engage Twitter users. We recorded a massive rise in Twitter activity during these debates. In this section we provide an example of analysis carried out on the data for one of these debates and, specifically, for the TV debate (#2) ”Leaders’ debate” (02/04/2015, 20:00-22:00 BST)[3].

The leaders’ debate involved leaders of seven British political parties including the major and ’fringe’ parties. Nick Clegg (Liberal Democrats), Ed Miliband (Labour Party) and the current Prime Minister David Cameron (Conservative Party) represented the main political forces. The leaders of the fringe parties were Natalie Bennett (Green Party of England and Wales), Nigel Farage (UKIP), Leanne Wood (Plaid Cymru) and Nicola Sturgeon, First Minister of Scotland (Scottish National Party, SNP). The debate was divided into four slots each dedicated to a different topic including: 1) budget deficit, 2) National Health Service (NHS), 3) immigration and 4) future of young people. The leaders were invited to present their stance and future policies in relation to each of the theme.

The total count of ’political’ tweets, that is, tweets including specific references to tracked terms and produced on the day of the debate was about $800,000$, of which nearly 80% were generated between 7pm and midnight , obviously associated to the live TV event.

Figure 3 shows the number of ’political’ tweets recorded in that day, those containing ’#leadersdebate’ and the total number of tweets we estimated when considering the tweets that were missed due the streaming track limit. This confirms a high involvement of the public in the leaders’ debate.

Subsequently, a sentiment score was assigned to the political tweets generated during the debate and these were averaged over time intervals of one minute. The graph in Figure 4 shows this 1-minute average of the political sentiment index for each of the six parties (Labour, Tories, UKip, LibDems, Greens, SNP and Plaid Cymru) to provide a representation of the Twitter moods in relation to political parties as the debate evolved.

We were particularly interested in the public sentiment in relation to the views and policies that each leader expressed regarding the main themes. In other words, we tried to identify the party policy which received most positive or most negative responses. The sentiment towards a party fluctuated depending on the topic discussed. Whereas the Liberal Democrats and the Labour party received more endorsement for their NHS policies, their stance towards immigration seemed to win less support. In contrast, immigration appeared to improve the score of UKip, the essentially anti-Europe and anti-immigration party. There was also one leader who seemed to be positively valued for most of the debate. Nicola Sturgeon, the leader of the SNP, emerged as the ’winner’ of the debate. Especially, her call for a rational debate on immigration and a plea for free education were endorsed by Twitter users generating the highest positive sentiment during the debate. In this way, our analysis revealed

[3]https://www.youtube.com//watch?v=7Sv2AOQBd_s

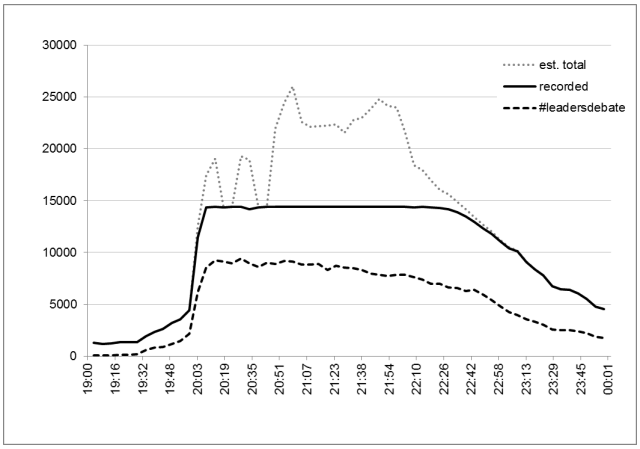


Fig. 3: Number of tweets recorded in 5’ intervals during the TV debate on April 2, 2015.

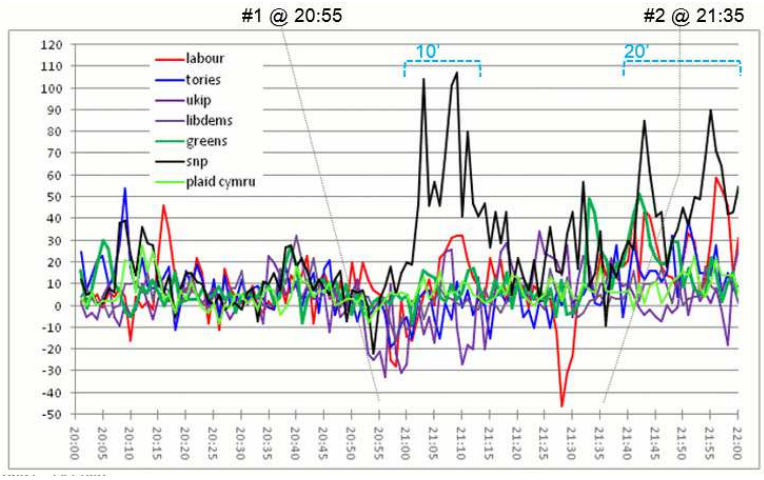


Fig. 4: Political sentiment index during the TV debate on April 2, 2015.

not just a general attitude or sentiment towards the leaders and the parties they were representing; it also pointed specifically to the issues that matter to the public.

In order to investigate further the positive response to SNP during these two key moments of the debate we generated tag clouds from the tweets following these moments.

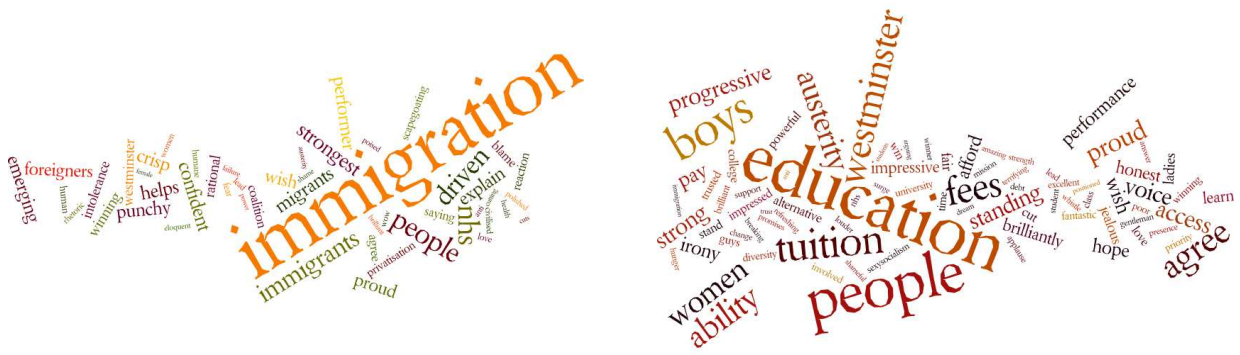
SNP’s leader Nicola Sturgeon’s appeal for a rational debate on immigration (20:55 BST) and her personal statement about free education that enabled her to be where she is (21:35 BST) won massive support, as does her final statement, in which she outlined SNP as an alternative to Westminster.

The two word clouds in Figure 5 have been generated with the frequent words found in the tweets associated with SNP during the two main periods, respectively, of 10’ and 20’ of Twitter popularity as indicated in the figure. The two word clouds clearly represent these two messages that appeared to be the winners of the leaders’ debate.

V. Use cases

With all the caveats as discussed above and practicalities of finding separating signals from noises in social media [14], social media adds a layer of information that may alleviate common estimation biases (e.g. omitted variable bias) in causal relationships across several fields. The above-mentioned empirical framework and data gathering exercise can be applied to several other research areas as potential ’use cases’, some of which are listed below.

- It is well established that many economic decisions such as willingness to buy goods and services can be influenced by the dynamics of social networks and the information transmitted within them, and hence understanding how such networks form, and their characteristics, can enhance analytical rigour in terms of devising targeted brand equity building, product positioning and marketing strategies.



(a) 10' interval after key moment at 20:55

(b) 20' interval after key moment at 21:35

Fig. 5: Word clouds for two selected time intervals during the TV debate on April 02, 2015.

- Many large-scale public events such as general elections have profound impacts across economic indicators and their relationships, and as such, the potential for better, quicker and more frequently updated analysis of the changing economic and social dynamics using social media data can be undertaken. It is for this reason we make our UK General Election 2015 dataset available to all who wish to use it for their research. Simple real-time analysis as well as more complex off-line analytics can provide interesting insights.
- Public and social policy making process is notorious in terms of being heavily constrained by a lack of real time, granular information around individuals and their reactions to policy instruments. Social media can potentially provide a mirror through which reflections of the public feedbacks can be ascertained. Such approach can also facilitate analysis of online feedback systems in other sectors e.g. in tourism domain [15], rural life [16].
- Across several sectors, one of the key challenges is to collate information across various touchpoints in the supply chain, especially when touchpoints are spatially fixed and separated. Social media having engaged user networks can provide insights into propagation of information on key incidents through various touchpoints in the supply chains.

VI. Conclusions

This study has presented a method used to collect tweets about UK politics for three months, from March to May 2015, in correspondence of the UK General Election 2015. The method allowed capturing a more comprehensive set of 'political' tweets than it would have been possible by tracking only obvious terms, such as #GE2015.

Simple real-time analysis and more complex off-line analytics have provided interesting insights. It was shown how a political sentiment index is useful to identify key moments in public events (e.g., in TV debates) and this was used to guide the execution of other data analytics techniques, such as a descriptive approach based on the visualisation of summaries by means of tag clouds.

The dataset is publicly available and can be used to test research ideas on text mining, data visualisation, complex social networks, economics and politics.

Future research directions include algorithmic and implementation improvements, and domain-specific research investigations. In particular, the method for data collection can be extended to include both automatic and interactive management of the tracked terms to adapt to concept drift and changes of interests in the social media.

References

[1] D. Bollier, *The promise and peril of big data.* Washington, DC: Communications and Society Program, Aspen Institute, 2010.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., May 2011.

[3] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59 – 68, 2010.

[4] UK General Election 2015 Blog, University of Reading. [Online]. Available: http://blogs.reading.ac.uk/reading-general-election-blog/

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proc. of the Fourth Int'l AAAI Conference on Weblogs and Social Media (ICWSM'10)*, Washington, DC, May 2010, pp. 178–185.

[6] A. Boutet, K. Hyoungshick, and E. Yoneki, "What's in twitter: I know what parties are popular and who you are supporting now!" in *Proc. of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'12)*, Istanbul, Turkey, Aug. 26–29, 2012, pp. 132–139.

[7] S. Stieglitz and L. Dang-Xuan, "Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior," in *Proc. of the 45th Hawaii International Conference on System Sciences (HICSS'12)*, Jan. 4–7, 2012, pp. 3500–3509.

[8] M. A. Razzaq, A. M. Qamar, and H. Bilal, "Prediction and analysis of pakistan election 2013 based on sentiment analysis," in *Proc. of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'14)*, Beijing, China, Aug. 17–20, 2014, pp. 700–703.

[9] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic, "The party is over here: Structure and content in the 2010 election," in *Proc. of the Fourth Int'l AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, Jul. 17–21, 2011, pp. 201–208.
[10] D. Weaver and D. Drew, "Voter learning and interest in the 2000 presidential election: Did the media matter?" *Journalism and Mass Communication Quarterly*, vol. 78 (4), no. 1, pp. 787–798, 2001.
[11] Twitter API. [Online]. Available: https://dev.twitter.com
[12] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proc. of the Fourth Int'l AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, Jul. 17–21, 2011.
[13] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11. Association for Computational Linguistics, 2011, pp. 30–38.
[14] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. ACM, 2008, pp. 183–194.
[15] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," *Tourism Management*, vol. 31, no. 2, pp. 179–188, 2010.
[16] E. Gilbert, K. Karahalios, and C. Sandvig, "The network in the garden: An empirical analysis of social media in rural life," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. ACM, 2008, pp. 1603–1612.