Soc. Netw. Anal. Min.
DOI 10.1007/s13278-013-0120-1

ORIGINAL ARTICLE

What's in Twitter, I know what parties are popular and who you are supporting now!

Antoine Boutet · Hyoungshick Kim · Eiko Yoneki

Received: 24 December 2012 / Revised: 22 May 2013 / Accepted: 3 June 2013

Abstract In modern politics, parties and individual candidates must have an online presence and usually have dedicated social media coordinators. In this context, we study the usefulness of analysing Twitter messages to identify both the characteristics of political parties and the political leaning of users. As a case study, we collected the main stream of Twitter related to the 2010 UK General Election during the associated period—gathering around 1,150,000 messages from about 220,000 users. We examined the characteristics of the three main parties in the election and highlighted the main differences between parties. First, the retweet structure is highly clustered according to political parties. Second, users are more likely to refer to their preferred party and use more positive affect words for the party compared with other parties. Finally, the self-description of users and the List feature can reflect the political orientation of users. From these observations, we develop both an incremental and practical classification method which uses the number of Twitter messages referring to a particular political party or retweets, and a classifier leveraging the valuable semantic content of the List feature to estimate the overall political leaning of users. The experimental results showed that the proposed incremental method achieved an accuracy of 86 % for classifying the users' political leanings and outperforms other classification methods that require expensive costs for tuning classifier parameters and/or knowledge about network topology. This advantage allows this approach to be a good candidate for social media analytics application in real time for political institution. The proposed method using List feature, in turn, achieved an accuracy of 92 %.

Keywords Twitter and politics · Party characteristics · User classification

1 Introduction

Social media such as Facebook and Twitter have revolutionised the way people communicate with each other. Users generate a constant stream of online messages through social media to share and discuss their activities, status, opinions, ideas and interesting news stories; social media might be an effective means to examine trends and popularity in topics ranging from economic, social, environmental to political issues (Tumasjan et al. 2010; Cha et al. 2010).

In modern politics, political parties must have an online presence to reach the users they want to influence. They also monitor social media to measure the success of their political campaigns and then refine their strategies (e.g. to help their candidates win in elections). This phenomenon creates a new opportunity for users to participate to democratic process and trends to increase civic participation from users and changes how people are engaging in politic debates (Zhang et al. 2009; Mcclurg 2003; Baumgartner and Morris 2010). The promise of social media starts to be exploited by government agencies from different countries as vehicles for consultation and for enhancing civic engagement (Downey and Jones 2012; Traunmüller 2010).

A. Boutet (✉)
INRIA Rennes Bretagne Atlantique, Rennes, France
e-mail: antoine.boutet@inria.fr

H. Kim
Sungkyunkwan University, Suwon, Korea
e-mail: hyoung@skku.edu

E. Yoneki
University of Cambridge, Cambridge, UK
e-mail: eiko.yoneki@cl.cam.ac.uk

Published online: 04 July 2013

We are particularly interested in this paper in how to identify the characteristics of political parties and the political leaning of users in social media. To illustrate the practicality of our analysis, we used a dataset formed of collected messages from Twitter, which is a popular social network and microblogging service that enables its users to broadcast and share information within posts of up to 140 characters, called tweets. We gathered around 1,150,000 messages from the main stream of Twitter related to the 2010 UK General Election between the 5th and the 12th of May from about 220,000 users in Twitter.

We first examined the characteristics of the three main parties (Labour, Conservative, Liberal Democrat) in the election and discussed the main differences between parties in term of activity, influence, structure, interaction, contents, mood and sentiment. Our results demonstrated that (a) the retweet structure is highly clustered according to political parties, (b) users are more likely to refer to their preferred party and use more positive affect words for the party compared with other parties, and (c) the self-description of users and the List feature can reflect the political orientation of users.

In addition, we analysed that Labour members were the most active and influential in Twitter during the election while Conservative members were the most organised to promote their activities. Also, the websites and blogs that each political party's members frequently referred to are clearly different from those that all the other political parties' members referred to. Furthermore, we observed that the level of interaction between members of different political parties can estimate the success to a coalition between their associated party.

Through this intensive analysis about the users with political interests, we develop a novel incremental and practical algorithm to identify the political leaning of users in the microblogging service (i.e. Twitter)—the messages expressing the user's political views (i.e. tweets referring to a particular political party and retweets from users with known political preferences) can be used to estimate the overall political leaning of users. Furthermore, we also propose a new classifier leveraging the valuable semantic content of the List feature to estimate the political orientation of users.

To demonstrate the effectiveness of the proposed heuristic models, we evaluated the performance of the proposed classification method based on a ground truth dataset composed of users who reported their political affiliation in their profile. The experimental results showed that our incremental method—which uses the number of tweets referring to a particular political party—achieved about 86 % classification accuracy using all trials, which outperforms the best known classification methods (see Pennacchiotti and Popescu 2011a; Zhou et al. 2011; Golbeck and Hansen 2011), which require expensive costs for tuning of parameters to construct classifier and/or the knowledge about network topology. Our method which uses the semantic content of the List feature, in turn, achieved about 0.92 % classification accuracy. Although some classification algorithms based on network topology or Lists feature performed well, these may indeed be unacceptable or very expensive: crawling topology information is strictly limited in practice.

Our incremental approach has the three following key advantages: (a) as we only process the messages relevant to a particular event rather than the whole dataset at one time, it dramatically reduces the computation costs of constructing a classifier compared with existing approaches—huge computational overhead for large training sets they impose are likely to be nontrivial, and they may indeed be unacceptable for online classification; (b) the proposed method does not require the knowledge about network topology unlike some classification methods based on community structure (Raghavan et al. 2007; Golbeck and Hansen 2011); (c) it also has potential: we can discover the temporal trends of a user's political views by analysing her political leaning over time. All these advantages allow this approach to address the emerging applications for political institutions to continuously monitor, analyse and summarise their impact and influence in social media.

2 Related work

The exponential growth and the ubiquitous trend of social media have attracted much attention. In particular, social media are increasingly used in political context (Wattal et al. 2010; Tumasjan et al. 2010; Aday et al. 2010). In this section, we review related works to social media in classification, characterisation, prediction, sentiment analysis and applications.

2.1 Classification

Different approaches have been proposed for classifying users in many directions. Lin and Cohen (2008) presented a semi-supervised algorithm for classifying political blogs. Zhou et al. (2011) also applied three semi-supervised algorithms for classifying political news articles and users, respectively. Their propagation algorithms particularly achieved the accuracy of 99 % which is higher than the accuracy results of this paper. This is because we used only 10 % of the dataset as initial seeds while they used 90 % of the dataset as initial seeds. Golbeck and Hansen (2011) presented a method that uses the follower connections in Twitter to identify users' political preferences. This method

achieved similar results to the label propagation method on the retweet graph in this paper.

Adamic and Glance (2005) studied the linkage patterns between political blogs and confirmed the hypothesis—the limited degree of contacts which may take place between the members of different social groups—which was suggested in Hewstone and Brown (1986). They found that the blogosphere exhibits a politically segregated community structure with more limited connectivity between different communities. Recently, Conover et al. (2011) observed a similar structure in a retweet graph of Twitter in political context. Other classifications used machine learning methods to infer information on users. Pennacchiotti and Popescu (2011a) demonstrated the possibility of user classification in Twitter with the three different classifications: political affiliation detection, ethnicity identification and detecting, affinity for a particular business. Their best algorithm achieved the accuracy of about 88.9 % for political affiliation. We note that their results might be overestimated compared with ours because the results were for binary-class classification. They also used Gradient Boosted Decision Trees (Pennacchiotti and Popescu 2011b) which is a machine learning technique for regression problems, which produce a prediction model in the form of an ensemble of decision trees.

In this paper, we tested several classification methods in order to demonstrate that our proposed method has a comparable performance to the best known classification methods (Pennacchiotti and Popescu 2011a; Zhou et al. 2011; Golbeck and Hansen 2011) that require expensive costs for tuning of parameters to construct classifiers and/or the knowledge about users or network topology. This is an extended paper of our preliminary work (Boutet et al. 2012a, b).

2.2 Characterisation

Characterisation aims to identify the main characteristics of population. Several studies have addressed to characterise user behaviour or personality in social networks and blogs (Benevenuto et al. 2009; Quercia et al. 2012; An et al. 2011; Agarwal et al. 2012). For instance, Sharma et al. (2012) have proposed to leverage the valuable semantic content of the Lists feature to infer attributes that characterise individual Twitter users. However, few works have tried to study the characteristics of politic parties and the interaction structure between parties. Some previous studies (Sarita and Danah 2010; An et al. 2011) showed that interactions between dislike-minded groups in social media expose people to multiple points of views and promote diversity and thus tend to reduce extreme behaviours. Livne et al. (2011) studied the usage patterns of tweets about the candidates in the 2010 US midterm election and showed stronger cohesiveness among Conservative and Tea party. Balasubramanyan et al. (2012) have proposed a model to predict how members of different political communities respond to the same news story.

2.3 Prediction

Other studies have addressed the predictive power of the social media. Asur and Huberman (2010) demonstrated how social media contents can be used to predict real-world outcomes and outperformed market-based predictor variables. In politics, Livne et al. (2011) have investigated the relation between the network structure and tweets and presented a forecast of the 2010 midterm election in the US. Tumasjan et al. (2010) claimed that Twitter can be considered as a valid indicator of political opinion and found that the mere number of messages mentioning a party reflects the election result through a case study of the German federal election. However, Gayo-Avello et al. (2011) demonstrated that this result was not repeatable with the 2010 US congressional election.

2.4 Sentiment analysis

O'Connor et al. (2010) used sentiment analysis to compare Twitter streams with polls in different areas and showed the correlation on some points. Quercia et al. (2011) studied the links between the degree of expressed sentiment and influence of users in Twitter and suggested that Twitter users are influenced by those who express negative emotions. Stieglitz and Dang-Xuan (2012) found that tweets containing words that reflect positive and negative emotions tend to be retweeted more often than those, which do not contain such words in political context. The same authors found similar results in political blogs where blog entries with either more positive or more negative overall sentiment tend to receive significantly more comments compared to sentiment-neutral or mixed-sentiment entries. Diakopoulos and Shamma (2010) showed that tweets can be used to track real-time sentiment about candidates' performance during a televised debate. Quercia et al. (2012) also analysed the correlation between the sentiment of tweets in a community and the community's socio-economic well-being. In addition, they proposed a machine learning technique to learn new positive and negative words for their dictionary of words reflecting people's emotional and cognitive perceptions.

Political blogs and mainstream journalists usually support their positions by criticising those of the opposite political figures. As shown in our sentiment analysis, users use more positive affect words for the party compared with other parties. However, some researches (Sarita and Danah 2010; An et al. 2011) showed that interactions between dislike-minded groups in social media expose people to multiple points of views and promotes diversity and thus

tend to reduce extreme behaviours. On the other hand, opinion formation in the social media and how political elites affect public opinion formation have been recently studied in Petersen et al. (2010), Sobkowicz et al. (2011), Kaschesky et al. (2011) and Druckman et al. (2013).

2.5 Applications

Open APIs provided by most of the social media to allow any developer to build application on the top of their platforms have received a great success. Consequently, plenty of applications appear (and disappear) every day. In period of the election, several Twitter applications can analyse each candidate's popularity and influence. Stieglitz and Dang-Xuan (2012) proposed a social media analytics framework in political context but not in real time unlike our solution. Government agencies also start developing applications to exploit social media to improve civic engagement (Loukis and Charalabidis 2012).

3 Twitter dataset for the UK General Election

The UK General Election took place on May 6, 2010, and was contested by the three major parties: the Labour party led by Gordon Brown, the Conservative Party led by David Cameron, and the Liberal Democrat (LibDem) party led by Nick Clegg. Although exit polls and initial results were released on the night of the 6th, the final outcome of the election, due to the UK parliamentary system, was not clear until the 11th of May, when Gordon Brown resigned and David Cameron became prime minister, announcing that he would attempt to form a coalition with the Liberal Democrats.

We collected all tweets published on the top trending topics related to the UK election between the 5th and 12th of May, and kept only the 419 topics which have over 10,000 tweets. The resulting dataset gathers more than 220,000 users for almost 1,150,000 tweets. Figure 1 shows how the volume of tweets referring to each party changed in response to the major events occurred over the election period.

The collected messages include about 168,000 mentions (direct messages to another user), 290,000 retweets (forward messages to its followers), 515,000 hashtags (tags used to define topics) and 25,000 distinct URLs. For these users, we also collected their profiles and about 79,000,000 following/follower relationships.

For some users, their profiles can be used to identify their political party affiliation (with manual checking). We called them self-identified members. We used the associated 633 Labour, 231 Conservative and 297 LibDem self-identified members as a ground truth dataset to evaluate the performance of classification methods. Furthermore, we collected the details about the 37,185 Lists subscribed by the ground truth users and about 42,000 users' location information including 27,000 users in UK from their profiles, too.

4 Party characteristics

In this section we analysed the characteristics of the Labour, Conservative and LibDem party to find only the relevant features for user's party affiliation. With the collected ground dataset (1.161 self-identified users), we found the communities of political parties from the Twitter network for a bigger set of users. Community detection in social networks has been largely addressed (Vasudevan and Deo 2012; Kashoob and Caverlee 2012; Branting2012). To achieve this here, we used a well-known technique called label propagation method (Raghavan et al. 2007) on the retweets structure. This technique is very reasonable—people usually retweet tweets which they prefer to be shared (i.e. tweets expressing a similar political opinion in our context), and thus form a highly clustered structure

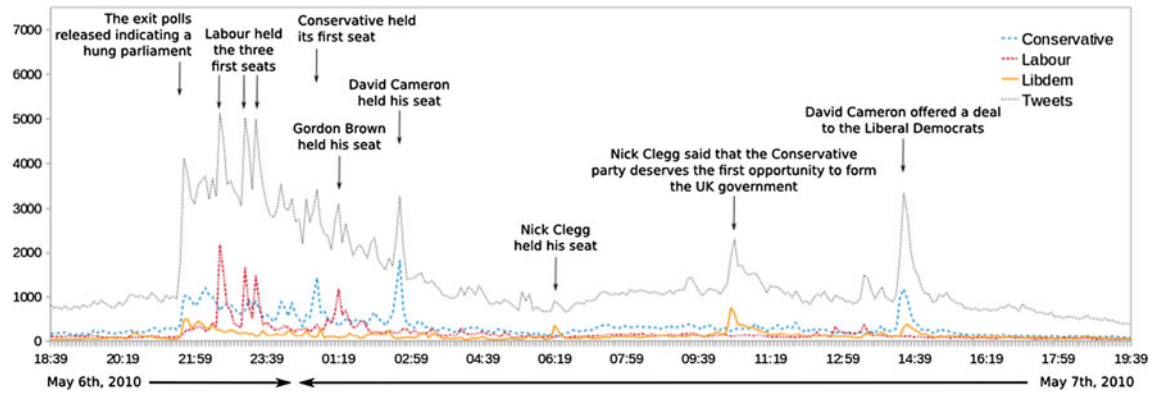


Fig. 1 Tweets volume and references to party after the exit polls

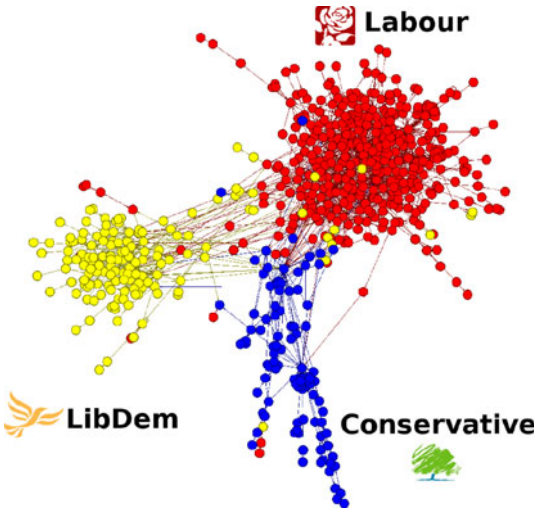


Fig. 2 Retweet graph among ground truth users

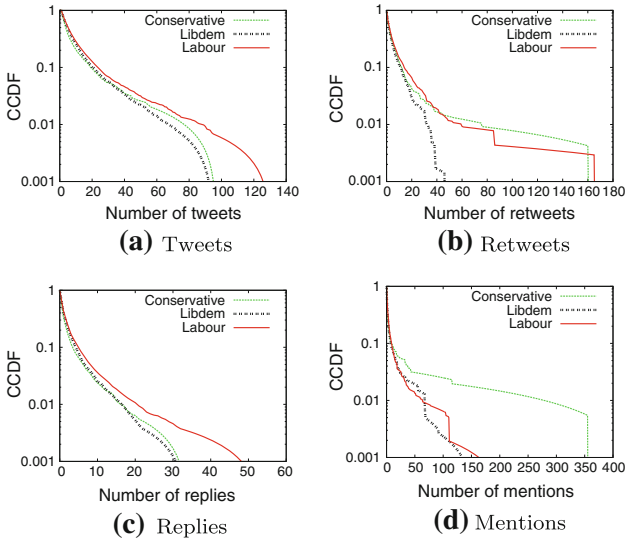


Fig. 3 CCDF for the activity metrics

according to parties in a retweet graph. Conover et al. (2011) recently verified this idea in politics on Twitter and Fig. 2, which shows the retweet structure among the ground truth users, confirms this assumption in our dataset as well (for visualisation, we used the Force Atlas layout of Gephi, Bastian et al. 2009).

Here, the label propagation method spreads affiliations from ground truth users called seeds throughout the retweet graph where edges are weighted according to the number of retweet—we labelled a user with the party affiliation according to seeds who have reached it with the largest number of retweets. We performed the label propagation until the propagation distance k to avoid tie-breaking cases (i.e. when there exist multiple nearest nodes with different party memberships and the same edge weight). It is achieved for $k = 2$ which permitted to detect 5,878 Labour, 3,214 LibDem and 2,356 Conservative candidates. We tested the performance of this heuristic by selecting one-tenth of the ground truth users, 115 were used as the seed users and the rest (1,046) were reserved for testing. This heuristic produced a high accuracy of 0.77, 0.78 and 0.90, respectively, for an average at 0.82. With these candidates, we analysed the following characteristics of each party: (a) activity, (b) influence, (c) structure and interaction, (d) content and (e) sentiment features.

4.1 Activity

The amount of messages about the political issues in Twitter can be used for measuring the activities of political parties. The activity level of parties can be measured in the different functions: the content generation is measured by the number of tweets; the content relay is quantified by the number of retweets; and the participation in political debates is evaluated by the number of replies and mentions. Figure 3 shows the Complementary Cumulative Distribution Function (CCDF) defined as $\bar{F}(x) = \mathrm{P}(X > x) = 1 - F(x)$ for these metrics where $F(x)$ is the cumulative distribution.

Interestingly, the Labour members generated more tweets and replies than those of the other parties while the Conservative members sent much more mentions than other parties. The LibDem party exhibited a relatively smaller activity for retweets.

4.2 Influence

The potential impact in term of visibility and information spread can be leveraged to evaluate the influence of each party. The numbers of following/followers are used to measure the size of the audience of members; the *star* metric defined by the ratio of $\frac{\text{followers}}{\text{following}}$ is used to evaluate the behaviour and the visibility of members in a party—information providers or stars tend to follow few while being followed by many (high *star* ratio), in contrast consumers tend to follow many while being followed by few people (low *star* ratio); the number of Lists, a feature in Twitter which allows users to create groups or circles of people in order to provide only one feed gathering their activities, is used to measure the level of organisation and promotion of the political parties; the numbers of times users of each party have been retweeted and mentioned are useful to evaluate the effective influence of parties.

Our analysis demonstrates that all metric values of the Labour members are significantly higher than those of the other two political parties except for the Lists (see Fig. 4). Probably, the Labour party benefited from more content providers than Conservative and LibDem generating a large numbers of tweets (correlation with Fig. 3a) which

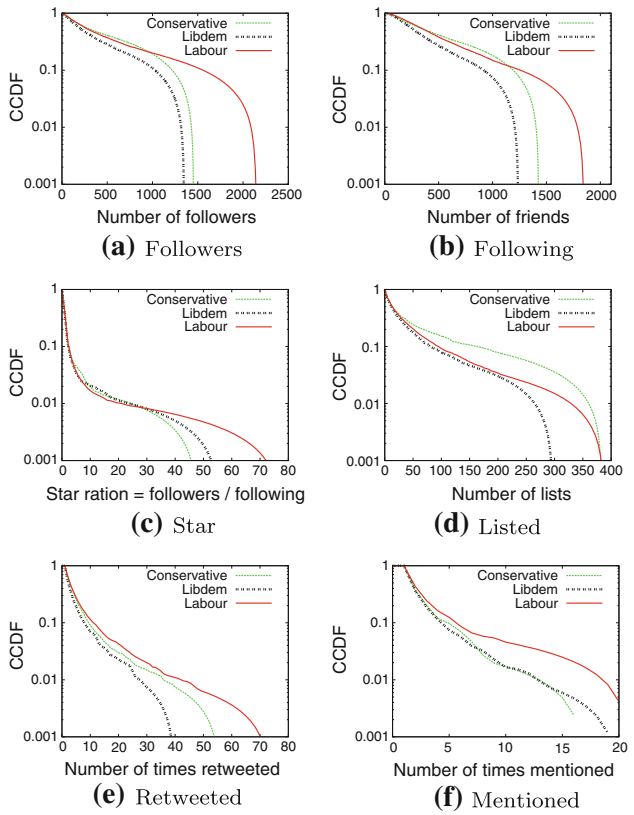


Fig. 4 CCDF for the influence metrics

were widely followed, retweeted and mentioned. In another hand, Conservative members were those which frequently used the Twitter Lists feature and probably the more organised to promote their activities during the election.

4.3 Structure and interaction

We also studied the differences between the political parties in network structure and interaction patterns. The structure patterns between members within a party reflect a level of party cohesion while the interaction patterns between different communities reflect the exchanges (i.e. conflict or collaboration) between them. Table 1 shows some properties (the average degree, the average Clustering Coefficient and size of the Largest Strongly Connected Components) of the following/followers graph for each party.

Table 1 Graph properties for each party

Dataset statistics	Labour	LibDem	Conservative
Nodes	**5,878**	3,214	2,356
Edges	**92,581**	32,586	24,949
Size in LSCC	**5,157**	2,418	2,183
Average degree	**31.5**	20.3	21.3
Average CC	0.2562	**0.3890**	0.3549

Table 2 Global properties for each party

Party	Followers	Tweets	Retweets	Mentions
Labour	128,997	71,022	19,275	1,507
LibDem	55,835	22,115	9,897	854
Conservative	17,644	20,383	4,667	942

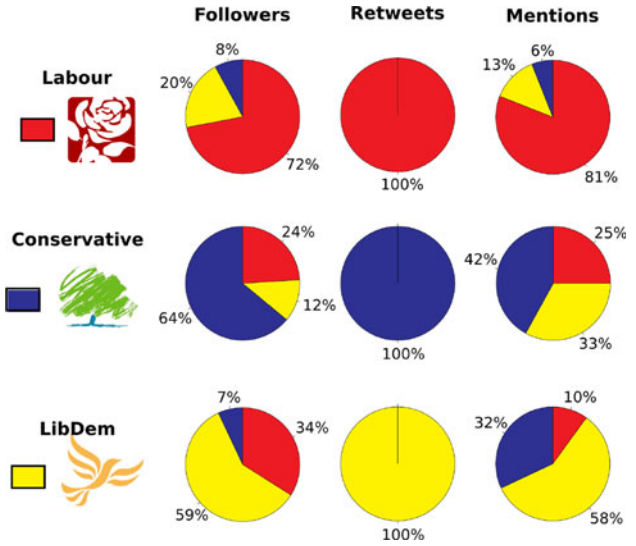


Fig. 5 Exchanged messages between parties

The Labour members formed a larger network structure and also had a high average degree compared with the other two parties. Interestingly, however, the structure of LibDem (0.3890) and Conservative (0.3549) members were much more clustered than that of Labour members (0.2562).

The relationship between following and follower in Twitter can be related to not only political preferences but other factors such as the same hobbies or locations. As a consequence, we also particularly observed the amount of interactions between political parties by counting the number of exchanged retweets and mentions between them during the election period (Table 2; Fig. 5).

According to the detected communities described above, we can see that there was no retweet exchanged between different political parties. In contrast, the mentions between different parties were more frequently used. We can also see that few interactions have been observed between the Labour and Libdem members, in opposition to the high rate of interactions between Conservative and both Labour and LibDem.

On May 7, several coalitions have been proposed. Clegg says that the Conservative party deserves the first opportunity to form a government while Brown raises possibility of talks with Liberal Democrats. Cameron, in turn, offers deal to the Liberal Democrats. We surmise that the

suggested coalition between Conservative and LibDem have generated more discussions among members of both parties than between Labour and LibDem. As the actual coalition involved Liberal Democrats and Conservative, this observation suggests that the success to a coalition between parties could be measured through the level of interaction between their members.

Finally, we analysed the correlation between social interaction and geographical distance in each party. Figure 6 shows the distribution of all interactions including retweets and mentions according to the distance between members in a party. All political parties had the similar behaviours, and mainly interacted with close users (around 50 % of the interactions was performed with users located at less than 50 km).

4.4 Content

By analysing the user-generated content, we can detect the trend in tweets and the habit of each political party. We first analysed the contents of tweets by counting the number of hashtags and URLs used in tweets for each party (see Fig. 7). We can see that the political parties showed a similar behaviour for the number of used URLs while Labour members used various hashtags in their tweets compared to the other parties.

Table 3 shows the ten most commonly used hashtags and their associated usage rates per party. The usage rates of neutral hashtags indicating the UK election remained at a similar level between all parties while non-neutral hashtags were more or less used depending on their underlying meaning. For instance, about 80 % of the hashtag *#imvotinglabour* and about 7 % of the hashtag *#imnotvotingconservative* were used by the Labour and Conservative members, respectively.

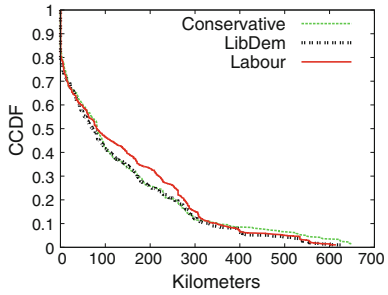
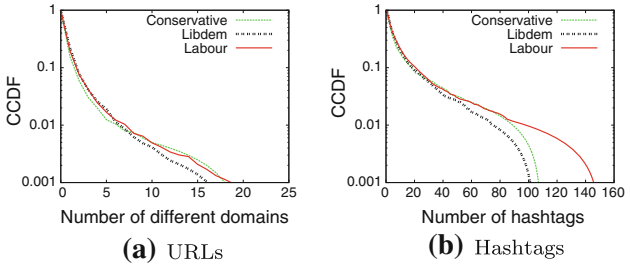


Fig. 6 Interaction according to the location



(a) URLs **(b)** Hashtags

Fig. 7 CCDF for the content metrics

Table 3 Ten most commonly used hashtags

Hashtags	Times	Labour	LibDem	Conserv.
#ge2010	39,742	0.34	0.36	0.28
#ukelection	13,506	0.31	0.27	0.40
#ukvote	6,332	0.35	0.34	0.29
#ge10	4,936	0.40	0.27	0.32
#GE2010	4,642	0.34	0.27	0.38
#imnotvotingconservative	1903	**0.50**	**0.41**	**0.07**
#electionday	1,586	0.36	0.27	0.36
#dontdoitnick	1,097	**0.63**	**0.25**	**0.10**
#imvotinglabour	904	**0.80**	**0.05**	**0.14**
#ukelection2010	795	0.40	0.26	0.32

We also analysed the hashtag similarity between users to evaluate the content homogeneity of each party. For a user, we define a vector containing the frequencies of hashtags used in the user's tweets and then we computed the cosine similarity between each pair of all users (users without hashtags have not been taken into consideration). Table 4 shows that the average similarity is overall low regardless of political party affiliation. That is, these results imply that Twitter users have heterogeneous behaviour in the use of hashtag.

By analysing the URLs mentioned in tweets, we can identify the preferred websites of each party. Table 5 shows the 10 most commonly used websites and their associated usage rates per party. We can see that the LibDem members more frequently referred to *Financial Times*, *The Independent* and *The BBC* compared with the other party members.

We also particularly observed the blogs which are usually more politically oriented. Only blogs using the most famous frameworks (http://blogspot.com, http://livejournal.com, http://wordpress.com, http://typad.com) have been taken into account. We compared the usage rates of these blogs

Table 4 Similarity of used hashtags according to parties

Party A ↔ Party B	cos(A, B)
Labour ↔ Labour	0.29
Labour ↔ LibDem	0.32
Labour ↔ Conservative	0.28
LibDem ↔ LibDem	0.35
LibDem ↔ Conservative	0.30
Conservative ↔ Conservative	0.28

Table 5 Ten most commonly used URLs

Websites	Times	Labour	LibDem	Conserv.
http://www.guardian.co.uk	532	0.37	0.34	0.28
http://www.youtube.com	484	0.30	0.31	0.37
http://twitpic.com	467	0.40	0.33	0.25
http://news.bbc.co.uk	314	0.26	**0.43**	0.25
http://yfrog.com	261	0.45	0.38	0.16
http://www.voterpower.org.uk	241	0.42	0.35	0.21
http://www.independent.co.uk	173	0.37	**0.51**	0.11
http://blogs.ft.com	137	0.24	**0.69**	0.05
http://sphotos.ak.fbcdn.net	115	0.27	**0.47**	0.24
http://www.telegraph.co.uk	83	0.38	0.32	0.28

Table 6 Three most cited blogs per party

Party	Blogs
Labour	http://thenewmrsbrown.wordpress.com
	http://newlyinterested.blogspot.com
	http://vonpip.wordpress.com
LibDem	http://lizw.livejournal.com
	http://cubiksrube.wordpress.com
	http://jeremyrowe1.wordpress.com
Conservative	http://dailyreferendum.blogspot.com
	http://conservativehome.blogs.com
	http://disenchanted-voter.blogspot.com

between parties. Table 6 shows the three most frequently referenced blogs per party. In addition, we observed very few overlaps of the referenced blogs between the parties. This result may confirm the high segregated structure of the blogosphere according to political parties reported in (Adamic and Glance 2005).

In addition, we measured the volume of references to a specific party included in tweets. We considered only the tweets referring to one name of party or its leader as such tweets are more likely to reflect the allegiance or interest of the users. Figure 8 illustrates the relative volumes of references to parties according to each party. These results clearly show that users were more likely to frequently refer to their own preferred party or candidate.

Finally, we analysed the most used words through the users description, the tweets and the description of the Lists subscribed by users. Due to the rate limit in the number of requests made with the Twitter API, only details of the 37,185 distinct Lists subscribed by the ground truth users have been taken into account. Table 7 depicts the most used words through the user description, the tweets and the Lists description for members of each party. We can see that both the user description and the description of the Lists subscribed by users contain valuable semantic content

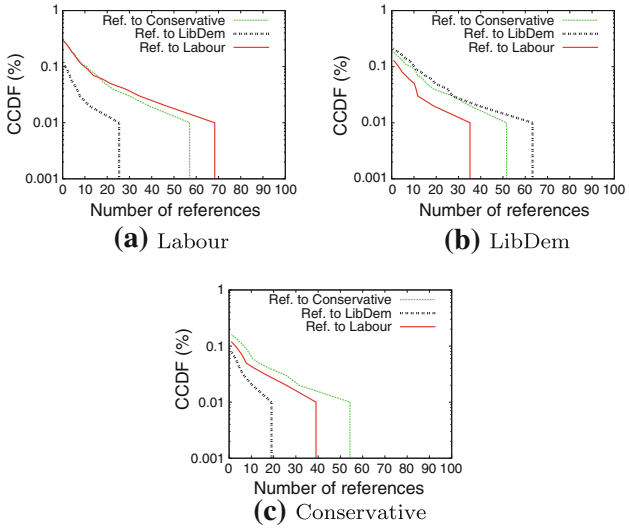


Fig. 8 CCDF for the volume of references

Table 7 Most used words by users

Party	User description	Tweets	Lists description
Labour	labour,	labour,	politics,
	party,	tories,	labour,
	member,	vote,	people,
	councillor,	tory,	political,
	activist	cameron	list
LibDem	liberal,	labour,	politics,
	democrat,	tories,	people,
	councillor,	clegg,	liberal,
	democrats,	vote,	libdems,
	candidate	tory	dems
Conservative	conservative,	labour,	politics,
	politics,	cameron,	people,
	activist,	clegg,	conservative,
	blogger,	tories,	conservatives,
	party	david	political

to characterise the political orientation of users. In contrast, the content of the tweets does not reflect polarisation.

4.5 Sentiment

We evaluated the sentiment of words used in tweets. To extract this information we used the Linguistic Inquiry Word Count[1]. LIWC is a dictionary of words used in everyday conversations, which assesses the emotional, cognitive and structural components of a text sample. After removing the URLs and hashtags from the collected tweets, LIWC makes the words matching for positive (i.e. happy,

[1] An online version of LIWC is available at http://www.liwc.net.

good) and negative emotions (i.e. out, hate). Then, the sentiment for a given tweet was given by the sentiment score proposed by (Kramer 2010):

$$\text{Sentiment} = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n} \quad (1)$$

where p_i (n_i) is the fraction of positive (negative) words for user i; μ_p (μ_n) is the average fraction of positive (negative) across all users; and σ_p (σ_n) is the corresponding standard deviation.

Table 8 shows the average sentiment scores over tweets referring to a party. Due to the limit in the number of using Twitter API, we analysed the only 66 % of the 110,308 tweets referring to a party. It is clearly shown that better sentiment was expressed in tweets when users referred to their own preferred party or leader in the tweets.

We also evaluated the sentiment of words in tweets through other directions: the self-focus (i.e. I, my, me), cognitive (i.e. cause, know, ought) and social (i.e. she, their, them). However, none party exhibits a particular sentiment behaviour.

5 User classification

In this section we present a novel user classification approach based on the observations in the previous section. Our goal is to identify the party to which a user belongs (or prefers) to. For this purpose, we propose an incremental Bayesian approach which divides the whole period into subsequences where the affiliation of users evolves according to their tweet activities as shown in Fig. 9. This solution has the advantage to require only a user's tweet messages over time. We will show this approach performs well by evaluating the performance of the classification method. In addition, we also propose a Bayesian classifier using the retweets and a classifier based on the description of the Lists subscribed by users.

Table 8 Sentiment on the references to party

Party	Reference to	Average emotion score
Labour	Labour	**2.26**
	LibDem	0.08
	Conservative	0.14
LibDem	Labour	0.17
	LibDem	**0.85**
	Conservative	0.10
Conservative	Labour	−0.09
	LibDem	0.10
	Conservative	**1.22**

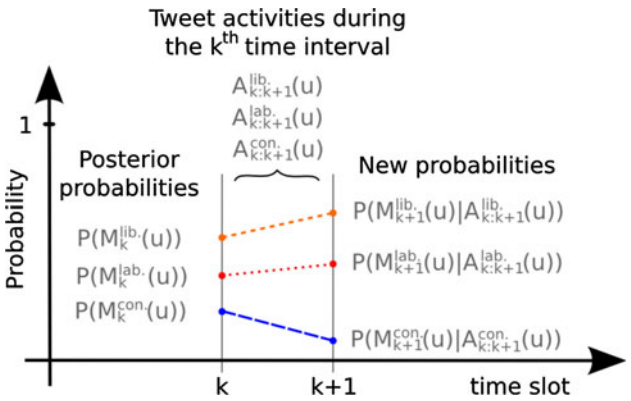


Fig. 9 Principe of the Bayesian classifier

5.1 Bayesian classification

Without loss of generality, we assume that a sequence of tweet activities (e.g. retweets or references to a specific party/politician in tweets) conducted by a user is divided into n subsequences, where the kth subsequence corresponds to the tweet activities during the kth time interval. For a user u, we use $A_k(u)$ and $M_k^i(u)$ to denote the kth subsequence (i.e., the tweet activities performed by the user u during the kth time interval) and the 0–1 binary variable indicating user u's membership for the party i after the kth time interval (i.e., $M_k^i(u) = 1$ when u is a member of the party i), respectively where $1 \le k \le n$ and $i \in \{labour, libdem, conservative\}$. We also use $P(M_k^i(u))$ to denote the probability of user u to be a member of the party i after the kth time interval. We assume that all users should be included to one of parties to satisfy the condition of $\sum_i P(M_k^i(u)) = 1$. After the nth time interval, we classify the user u as a member of the party j where $P(M_n^j(u)) = \max_i \{P(M_n^i(u))\}$. For example, when the affiliation probability distribution for the user u after the nth time interval is given as [0.7, 0.2, 0.1], we classify the user u as a member of the Labour party. We randomly choose the user u's party in case of equiprobability distribution.

We now focus on how to compute $P(M_k^i(u))$. At each time interval, for each $i \in \{labour, libdem, conservative\}$, $P(M_k^i(u))$ is updated stochastically according to its probability distribution relying on the user's tweet activities during the time interval.

Before the first inference step, the initial prior affiliation probability of the user u is set uniformly: $P(M_0^i(u)) = \frac{1}{3}, \forall i$. After the kth time interval, $P(M_k^i(u)|A_k(u))$ can be calculated by using Bayes' theorem as follows:

$$P(M_k^i(u)|A_k(u)) = \frac{P(A_k(u)|M_k^i(u))P(M_k^i(u))}{\sum_j P(A_k(u)|M_k^j(u))P(M_k^j(u))} \quad (2)$$

where $P(M_k^i(u)|A_k(u))$ is the posterior of user u, the uncertainty of $M_k^i(u)$ after $A_k(u)$ is observed; $P(M_k^i(u))$ is the

prior, the uncertainty of $M_k^i(u)$ before $A_k(u)$ is observed ; and $\frac{P(A_k(u)|M_k^i(u))}{P(A_k(u))}$ is a factor representing the impact of $A_k(u)$ on the uncertainty of $M_k^i(u)$.

We define two alternative conditional probabilities (evidences) to calculate $P(A_k(u)|M_k^i(u))$. The first one is defined as the frequency of *referring to political parties in tweets* for $A_k(u)$ based on the observation in the previous section. We can see that a user u more frequently generates tweet messages referring to the political party (or party leader) that the user u is supporting. For this activity, we assume $P(A_k(u)|M_k^i(u))$ can be calculated as follows:

$$P(A_k(u)|M_k^i(u)) \approx \frac{\sum_{t \in T(k)} V_i(t)}{|T(k)|} \quad (3)$$

where $T(k)$ is the tweets of the current user during the observed period and $V_i(t)$ is equal to 1 if the tweet t does a reference to the political party i, 0 otherwise. We use **Bayesian-Volume** to denote this Bayesian classification. This classifier has the advantages to only use the content of tweets without the information about users or network topology.

The second one is based on the pattern of retweets and highlights the fact that people of similar political persuasion might often retweet the same or similar things. Retweets can be easily identified in the tweets stream thanks to the keyword *RT* followed by the associated user-name of the source of information. This conditional probability is defined as the average of the affiliation probability of both people retweeted by the user or people retweeting the user (called retweet interactions) during the period $[k-1, k]$.

$$P(A_k(u)|M_k^i(u)) \approx \frac{\sum_{j \in RT(k)} P_j(M_n^i)}{|RT(k)|} \quad (4)$$

where $RT(k)$ is the mapping of the user's identity of each retweet interaction of the current user during the observed period, and $P_j(M_n^i)$ is the prior probability of user j to be a member of the party i. We use **Bayesian-Retweet** to denote this Bayesian classification. This classifier requires to identify the political affiliation of some users in order to start the propagation of the probabilities.

5.2 Evaluation setup

The aim of our experiment was to demonstrate feasibility and effectiveness of the proposed classification approach compared with the other popularly used classification methods. For comparison, we also tested the classification accuracy of the following classification methods:

- *Volume classifier* As we observed, the volume of reference to a specific party can reflect the political leaning of the user. We simply counted the frequencies of referencing to parties (or party leaders) in a user's tweets and then assigned the most frequently referenced party to the user's political party.
- *Sentiment classifier* As we observed, a user is more likely to express positive emotions in the user's tweets when the user posts tweets her preferred political party. We compute the user's sentiment scores of parties through the sentiment analysis of the user's tweets and then assign the party with the largest average emotion score to the user's political party.
- *Retweet classifier* As the retweet structure is highly segregated according to the party, the retweet graph can be used to predict users' affiliation. The representative method detects the communities of users using a label propagation method (Raghavan et al. 2007) on the retweet graph. In the label propagation process, each user's party is classified with the majority party in the user's neighbours. Ties can be broken according to the volume of references to party. From the initial seed users (self-identified members), we iterate this process until all users' parties are uniquely labelled.
- *Follower classifier* The relationship of following and being followed in Twitter can reflect the political leanings of users as well (Golbeck and Hansen 2011). Compared to the previous classifier, this one uses the followers graph to propagate the probability to be members of a certain political party from the selected ground truth users. The inferred probabilities are computed as the average probabilities for all people he or she follows.
- *Lists classifier* Sharma et al. (2012) have proposed to leverage the valuable semantic content of the Lists feature to infer attributes that characterise individual Twitter users. As we observed, the words used in the description of users and/or the description of the subscribed Lists of users can be used to estimate their political orientation. However, as the ground truth users have been manually selected according to the semantic content of their self-description, we have only taken into account the description of subscribed Lists. This classifier measures the similarity between the ten most used words in descriptions—given a user u, we first find the most similar user v in the training set and then assign the user v's party to the user u's political party.
- *SVM classifier* Support Vector Machine (SVM) is known as one of the best supervised learning techniques for solving classification problems with high dimensional feature space and small training set size. We constructed a SVM classifier using the following six features of a user proposed in Pennacchiotti and Popescu (2011a, b): (a) the list of followers (SOC-FOLL), (b) the list of friends (SOC-FRIE), (c) the list of retweeted users (SOC-RET), (d) the list of used words in the user's tweets (LING-WORD), (e) the list

Table 9 Performance according with approach

Approach	Accuracy
Volume	0.62
Sentiment	0.67
Follower	0.80
Retweet	0.72
Lists	**0.92**
SVM	
SOC-FOLLOWER	0.82
SOC-FRIEND	0.75
SOC-RETWEET	0.65
LING-WORD	0.58
LING-HASH	0.59
LING-SENTIMENT	0.54
LING-OPOSITION	0.61
All	0.80
Bayesian	
Volume	**0.86**
Retweet	0.64

of used hashtags in the user's tweets (LING-HASH), and (f) the emotion over the user's tweets (LING-SENT).

To show the performance of a classifier, we measured their *accuracy* only for the self-identified users (1161 from the ground dataset). The classification accuracy is defined as the ratio between the number of correctly predicted samples; the results are shown in Table 9. Classifiers used tweets and relationships related to these self-identified users. These users published 27,696 tweets; they formed a followers graph of 135,786 users for 7, 113, 860 edges and a retweet structure composed of 89,942 users for 286,614 retweets.

Some classifiers (Follower, Retweet, Lists, SVM and Bayesian-Retweet) require a training step used to learn the features determining political party membership and/or the knowledge about network topology or additional information relative to users. Training samples are composed of one-tenth of the ground truth users (115) picked at random to construct the classifiers and the rest (1,046) was reserved for out-of-sample testing. The impact of varying the composition of the training sample is discussed in Sect. 5.3.

5.3 Experiments

Unlike our expectations, results show that approaches using following/follower relationship produce better accuracy than the ones using the retweet structure. Also, SVM which involves an expensive tuning phase, did not outperform other algorithms. The best accuracies are given by the Lists (0.92) and the Bayesian-Volume (0.86) classifiers.

Although the performance of the Bayesian method computed only once at the end of the period is not as strong as some other candidates (accuracy of 0.64 in this case), it outperforms all classification methods except for the Lists classifier when it leverages its incremental approach over time with 10 updates of the users' affiliation probabilities during the period (accuracy of 0.86 as shown on Fig. 10). We used fixed time interval of 15 h to periodically updates the users' affiliation probabilities according to their tweets in the associated interval.

We note that the Bayesian classification method has two advantages compared with the other methods including the Lists classifier. First, it requires to maintain only the affiliation probability of each user without massive training overheads and second, as the information about references to political parties (or politicians) in tweets is only needed, incremental computation is significantly faster. These important advantages make it possible to use this solution in real time. Therefore, we recommend that Bayesian-Volume should be used as an alternative when the conditions do not allow the use of Follower or Lists which requires the knowledge about network topology or additional information to users to achieve good results, which may indeed be unacceptable or very expensive: crawling the information about users and network topology is strictly limited in practice.

In addition, we analysed the accuracy of these classifiers according to the set of training samples among (a) the most influential users with the highest number of followers, (b) the most active users with the highest number of published tweets and, (c) random users. Results are depicted in Table 10, the training sample set based on the most influential users provide the best accuracy for the Follower and the Retweet classifiers. Indeed, these classifiers require to use hubs or important users as seeds in order to start label or probability propagation. In contrast, as the SVM and the Lists classifier aim to build a generalised model reflecting the behaviour of average users, the uniform selection of users might be better than the selection of the most active or influent users. The Bayesian-Retweet, in turn, is not sensitive to the choice of the initial seeds.

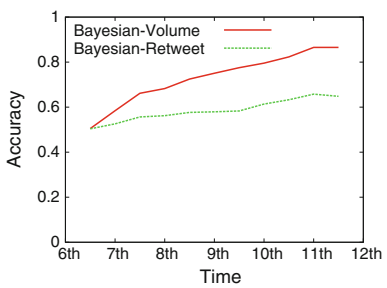


Fig. 10 Dynamic changes of the Bayesian classifiers accuracy over time

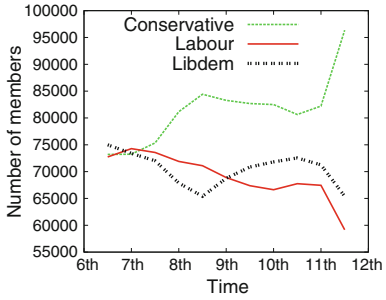


Fig. 11 Dynamic changes of the numbers of members over time

Table 10 Variation of the accuracy according to seeds

Approach	Random	Most active	Most influent
Follower	0.80	0.77	**0.83**
Retweet	0.72	0.76	**0.81**
SVM-All	**0.80**	0.69	0.77
Lists	**0.92**	0.89	0.90
Bayesian-Retweet	**0.64**	0.64	0.62

We also analysed both how the accuracy of the proposed Bayesian classifier changes with time over the self-identified users (Fig. 10) and how the number of partisans of each party evolves over all the 220,000 users of our dataset (Fig. 11). We can see that the Conservative members outnumbers the Labour and LibDem members at the end of the election. Inherently, the accuracy of Bayesian starts at $\frac{1}{2}$ (equiprobability), continuously increases with time, and achieved at 0.86. These results imply that the proposed Bayesian approach is proper to understand users' political leaning over time.

6 Conclusion

The existing classification methods are generally based on the assumption that the data conforms to a stationary distribution. Since the statistical characteristics of the real-world data continuously change over time, this assumption may lead to degrade the predictive performance of a classification model when the characteristics of dataset are dynamically changed. To address this weakness, we proposed a novel user classification approach using Bayesian framework which can incrementally update the classification results over time. Moreover, this approach does not require the knowledge about users or network topology unlike the previous solutions (Raghavan et al. 2007; Golbeck and Hansen 2011).

As a case study, we first analysed the characteristics of the political parties in Twitter during the 2010 UK General Election and identified three main ways to differentiate political parties: (a) the retweet graph presented a highly segregated partisan structure; (b) party members were more likely to make reference to their own party than the other parties; and (c) users were more likely to express more positive opinions when they referenced to their preferred party. With these party characteristics, we developed a classification algorithm based on Bayesian framework to compute political preferences of users. The experimental results showed that the proposed classification method is capable of achieving an accuracy of 86 % without any training and the network topology information which makes it a proper solution for a social media analytics application processing real time classification.

While our sentiment analysis evaluation incorporating linguistic information has not exhibited a particular sentiment behaviour for each party, using advanced solutions to dynamically capt the sentiment of users and the political parties as well as the dynamics of political opinion formation in response to the major events in the election campaign is an interesting perspective for future works.

Acknowledgments This research is part-funded by the EU grants for the RECOGNITION project (FP7-ICT 257756), the EPSRC DDEPI Project, EP/H003959 and by the ERC Starting Grant GOSSPLE number 204742.

References

Adamic L, Glance N (2005) The political blogosphere and the 2004 U.S. election: divided they blog. In: International workshop on link discovery. Chicago, Illinois, pp 36–43

Aday S, Farrell H, Lynch M, Sides J, Kelly J, Zuckerman E (2010) Blogs and bullets: new media in contentious politics. Tech Rep. Issue no 65

Agarwal N, Liu H, Tang L, Yu PS (2012) Modeling blogger influence in a community. Soc Netw Anal Min 2(2):139–162

An J, Cha M, Gummadi K, Crowcroft J (2011) Media landscape in Twitter: a world of new conventions and political diversity. In: International conference on weblogs and social media. AAAI, Barcelona, Spain

Asur S, Huberman BA (2010) Predicting the future with social media. In: International conference on web intelligence and intelligent agent technology. vol 1, Washington, DC, USA, pp 492–499

Balasubramanyan R, Cohen WW, Pierce D, Redlawsk DP (2012) Modeling polarizing topics: when do different political communities respond differently to the same news? In: International conference on weblogs and social media. IAAA, Dublin, Ireland

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: International conference on weblogs and social media. IAAA, San Jose, California, USA

Baumgartner JC, Morris JS (2010) MyFaceTube politics: social networking web sites and political engagement of young adults. Soc Sci Comput Rev 28(1):24–44

Benevenuto F, Rodrigues T, Cha M, Almeida V (2009) Characterizing user behavior in online social networks. In: Conference on internet measurement conference. Chicago, Illinois, USA

Boutet A, Kim H, Yoneki E (2012a) What's in your tweets? i know who you supported in the UK 2010 general election. In: International conference on weblogs and social media. IAAA, Dublin, Irlande

Boutet A, Yoneki E, Hyoungshick K (2012b) What's in twitter: i know what parties are popular and who you are supporting now! In: International conference on advances in social networks. Istanbul, Turkey

Branting K (2012) Context-sensitive detection of local community structure. Soc Netw Anal Min 2(3):279–289

Cha M, Haddadi H, Benevenuto F, Gummadi P (2010) Measuring user influence in Twitter: the million follower fallacy. In: International conference on weblogs and social media. IAAA, Washington, DC, USA

Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Flammini A, Menczer F (2011) Political polarization on twitter. In: International conference on weblogs and social media. IAAA, Barcelona, Spain

Diakopoulos NA, Shamma DA (2010) Characterizing debate performance via aggregated twitter sentiment. In: Conference on human factors in computing systems. Atlanta, Georgia, USA, pp 1195–1198

Downey E, Jones MA (2012) Public service, governance and web 2.0 technologies: future trends in social media. IGI Publishing

Druckman JN, Peterson E, Slothuus R (2013) How elite partisan polarization affects public opinion formation. Am Political Sci Rev 107:57–79

Gayo-Avello D, Metaxas PT, Mustafaraj E (2011) Limits of electoral predictions using twitter. In: International conference on weblogs and social media. IAAA, Barcelona, Spain

Golbeck J, Hansen D (2011) Computing political preference among twitter followers. In: conference on human factors in computing systems. Vancouver, BC, Canada, pp 1105–1108

Hewstone M, Brown R (1986) Contact is not enough: an intergroup perspective on the "Contact Hypothesis". In: Contact and Conflict in Intergroup Relations. Blackwell, UK

Kaschesky M, Sobkowicz P, Bouchard G (2011) Opinion mining in social media: modeling, simulating, and visualizing political opinion formation in the web. In: International Conference on Digital Government Research. College Park, MD, USA

Kashoob S, Caverlee J (2012) Temporal dynamics of communities in social bookmarking systems. Soc Netw Anal Min 2(4):387–404

Kramer AD (2010) An unobtrusive behavioral model of "gross national happiness". In: Conference on human factors in computing systems. Atlanta, Georgia, USA, pp 287–290

Lin F, Cohen WW (2008) The multirank bootstrap algorithm: Self-supervised political blog classification and ranking using semi-supervised link classification. In: International conference on weblogs and social media. IAAA, Seattle, WA, USA

Livne A, Simmons MP, Adar E, Adamic LA (2011) The party is over here: structure and content in the 2010 election. In: International conference on weblogs and social media. IAAA, Barcelona, Spain

Loukis E, Charalabidis Y (2012) Participative public policy making through multiple social media platforms utilization. Int J Electron Gov Res 8(3):78–97

Mcclurg SD (2003) Social networks and political participation: the role of social interaction in explaining political participation. Political Res Q 56:449–464

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: International conference on weblogs and social media. IAAA, Washington, DC, USA

Pennacchiotti M, Popescu AM (2011a) Democrats, republicans and starbucks afficionados: User classification in twitter. In: International conference on knowledge discovery and data mining. San Diego, California, USA, pp 430–438

Pennacchiotti M, Popescu AM (2011b) A machine learning approach to twitter user classification. In: International conference on weblogs and social media. IAAA, Barcelona, Spain

Petersen MB, Slothuus R, Togeby L (2010) Political parties and value consistency in public opinion formation. Public Opin Q 74(3): 530–550

Quercia D, Ellis J, Capra L, Crowcroft J (2011) In the mood for being influential on Twitter. In: International conference on social computing. Boston, MA, USA, pp 307–314

Quercia D, Ellis J, Capra L, Crowcroft J (2012) Tracking gross community happiness from tweets. In: Conference on computer supported cooperative work. Seattle, Washington, USA, pp 965–968

Quercia D, Lambiottez R, Stillwell D, Kosinski M, Crowcroft J (2012) The personality of popular facebook users. In: Conference on computer supported cooperative work. Seattle, Washington, USA, pp 955–964

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106

Sarita Y, Danah B (2010) Dynamic debates: an analysis of group polarization over time on twitter. Bull Sci Technol Soc 30(5):316–327

Sharma NK, Ghosh S, Benevenuto F, Ganguly N, Gummadi K (2012) Inferring who-is-who in the twitter social network. In: Workshop on online social networks. Helsinki, Finland, pp 55–60

Sobkowicz P, Kaschesky M, Bouchard G (2011) Opinion formation in the social web: agent-based simulations of opinion convergence and divergence. In: Agents and data mining interaction. vol 7103. Taipei, Taiwan, pp 288–303

Stieglitz S, Dang-Xuan L (2012) Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In: Hawaii international conference on system sciences. IEEE Computer Society, Grand Wailea, Maui, Hawaii, pp 3500–3509

Stieglitz S, Dang-Xuan L (2012) Social media and political communication: a social media analytics framework. Social Network Analysis and Mining

Traunmüller R (2010) Web 2.0 creates a new government. In: Electronic Government and the Information Systems Perspective. Bilbao, Spain, pp 77–83

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Election forecasts with twitter: How 140 characters reflect the political landscape. Soc Sci Comput Rev 29(4):402–418

Vasudevan M, Deo N (2012) Efficient community identification in complex networks. Soc Netw Anal Min 2(4):345–359

Wattal S, Schuff D, Mandviwalla M, Williams CB (2010) Web 2.0 and politics: the 2008 U.S. presidential election and an e-politics research agenda. MIS Q 34(4):669–688

Zhang W, Johnson TJ, Seltzer T, Bichard SL (2009) The Revolution Will be Networked: The Influence of Social Networking Sites on Political Attitudes and Behavior. Soc Sci Comput Rev 28(1):75–92

Zhou DX, Resnick P, Mei Q (2011) Classifying the political leaning of news articles and users from user votes. In: International conference on weblogs and social media. July, IAAA, Barcelona, Spain