What's in Twitter: I Know What Parties are Popular and Who You are Supporting Now!

Antoine Boutet
INRIA Rennes Bretagne Atlantique
Rennes, France
antoine.boutet@inria.fr

Hyoungshick Kim
University of British Columbia
Vancouver, Canada
hyoung@ece.ubc.ca

Eiko Yoneki
University of Cambridge
Cambridge, United Kingdom
eiko.yoneki@cl.cam.ac.uk

***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, Labour members were the most active and influential during the election while Conservative members were the most organized to promote their activities. Second, the websites and blogs that each political party's members supported are clearly different from those that all the other political parties' members supported. From these observations, we develop a simple and practical classification method which uses the number of Twitter messages referring to a particular political party. The experimental results showed that the proposed classification method achieved about 86% classification accuracy and outperforms other classification methods that require expensive costs for tuning classifier parameters and/or knowledge about network topology.**

I. 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 [1], [2].

In modern politics, political parties must have an online presence. In this context, monitoring social media can help parties and individual candidates to measure the success of their political campaigns and then refine their strategies. 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 Labour members were the most active and influential in Twitter during the election while Conservative members were the most organized 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 policital parties' members referred to.

Through this intensive analysis about the users with political interests, we develop a simple 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) is used to estimate the overall political leaning of users.

To demonstrate the effectiveness of the proposed heuristic model, 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 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 [3], [4], [5]), which require expensive costs for tuning of parameters to construct classifier and/or the knowledge about network topology. Although some classification algorithms based on network topology performed well, these may indeed be unacceptable or very expensive: crawling topology information is strictly limited in practice.

Our approach has three key advantages: (1) 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; (2) the proposed method does not require the knowledge about network topology unlike some classification methods based on community structure [6], [5]; (3) it also has potential: we can discover the temporal trends of a user's political views by analysing her political leaning over time.

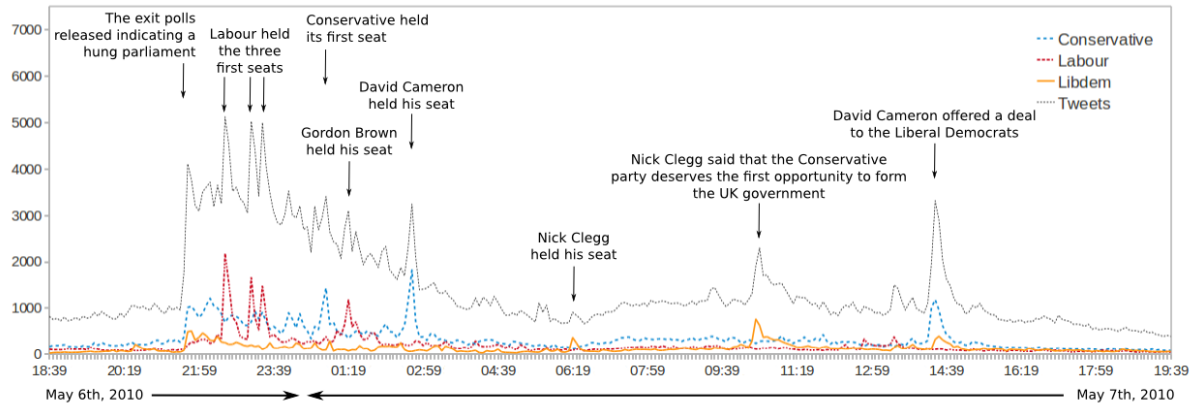


Fig. 1: *Tweets volume and references to party after the exit polls.*

II. Twitter Dataset for the UK General Election

The UK General Election took place on May 6th, 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 showed 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 check). 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 can collected about 42,000 users' location information including 27,000 users in UK from their profiles, too.

III. Party Characteristics

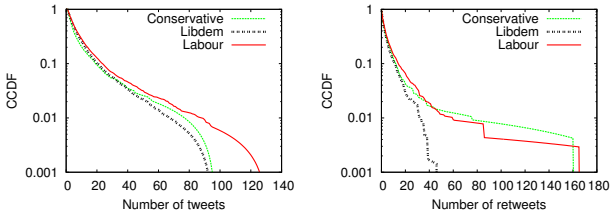
In this section we analysed the characteristics of the Labour, Conservative and LibDem party to find only the relevant features for user's party affiliation. To have a larger set of users to observe than the collected ground truth information, we first detected the communities associated to each political party. To achieve that, we used a well-known technique called label propagation method [6] on the retweets structure. This technique is very reasonable – people usually retweet tweets they like (i.e. tweets expressing a similar political opinion in our context), and thus form a highly clustered structure according to parties in a retweet graph. [7] recently verified this idea in politics on Twitter.

Here, the label propagation method spreads affiliations from ground truth users called seeds throughout the retweet graph – we label a user with the party affiliation according to seeds who have reached it. We performed the label propagation until the greatest propagation distance k which avoids tie-breaking case (i.e. multiple nearest nodes with different party memberships exist at the same time). 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) was used as the seed users and the rest (1,046) was 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 analyzed the following characteristics of each party: (*i*) activity, (*ii*) influence, (*ii*) structure/interaction, (*iv*) content and (*v*) sentiment features.

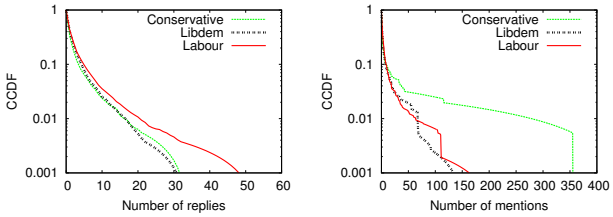
A. Activity

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 2 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.

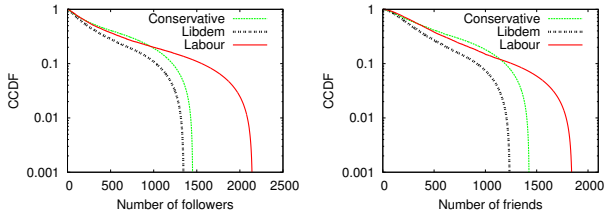


(a) Tweets (b) Retweets

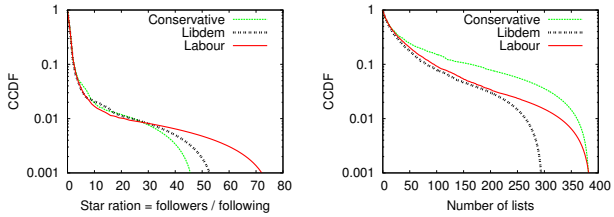


(c) Replies (d) Mentions

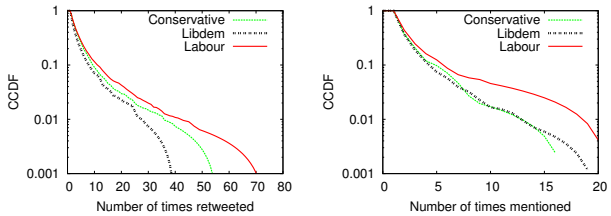
Fig. 2: *CCDF for the activity metrics.*



(a) Followers (b) Following



(c) Star (d) Listed



(e) Retweeted (f) Mentioned

Fig. 3: *CCDF for the influence metrics.*

B. 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 $\frac{followers}{following}$ 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 [1] in Twitter is used to measure the level of organization 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 Figure 3). Probably, the Labour party benefited from more content providers than Conservative and LibDem generating a large numbers of tweets (correlation with Figure 2a) which were widely followed, retweeted and mentioned. In another hand, Conservative members were those which frequently used the Twitter Lists feature and probably the more organized to promote their activities during the election.

C. Structure and Interaction

We also studied the differences between the political parties in network structure and interaction patterns. The structure and the interaction 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. Tables I 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. 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).

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 I: *Graph properties for each party.*

In addition to the following/followers graph, 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 (Figure 4).

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. We surmise that the suggested coalition between

[1]The Twitter Lists feature allows users to create groups or circles of people in order to provide only one feed gathering their activities.

Conservative and LibDem have generated more discussions among members of both parties than between Labour and LibDem.

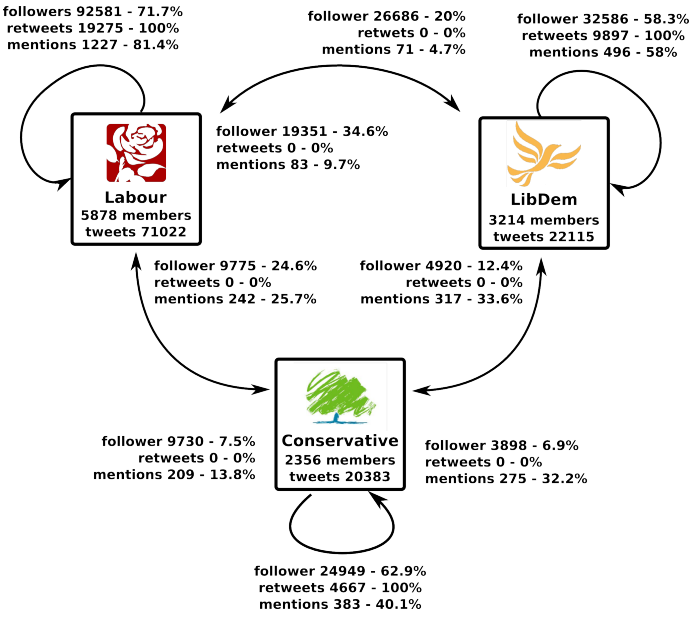


Fig. 4: *Exchanged messages between parties*

Finally, we analysed the correlation between social interaction and geographical distance in each party. Figure 5 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 kilometers).

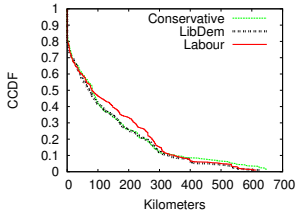
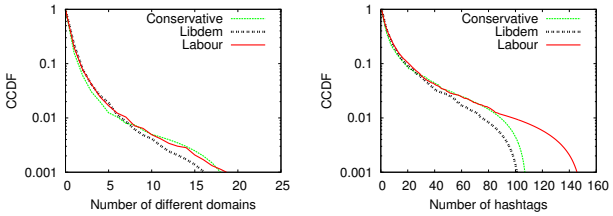


Fig. 5: *Interaction according to the location.*

D. Content

We analysed the contents of tweets by counting the number of hashtags and URLs used in tweets for each party (see Figure 6). 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 II 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



(a) URLs (b) Hashtags

Fig. 6: *CCDF for the content metrics.*

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.

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

TABLE II: *Ten most commonly used hashtags.*

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. Table III 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.

Party A	Party B	cos(A, B)
Labour	Labour	0.14
	LibDem	0.14
	Conservative	0.13
LibDem	Labour	0.15
	LibDem	0.18
	Conservative	0.18
Conservative	Labour	0.15
	LibDem	0.17
	Conservative	0.14

TABLE III: *Similarity of used hashtags according to parties.*

By analysing the URLs mentioned in tweets, we can identify the preferred websites of each party. Table IV shows the ten 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 (blogspot.com, livejournal.com, wordpress.com,

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

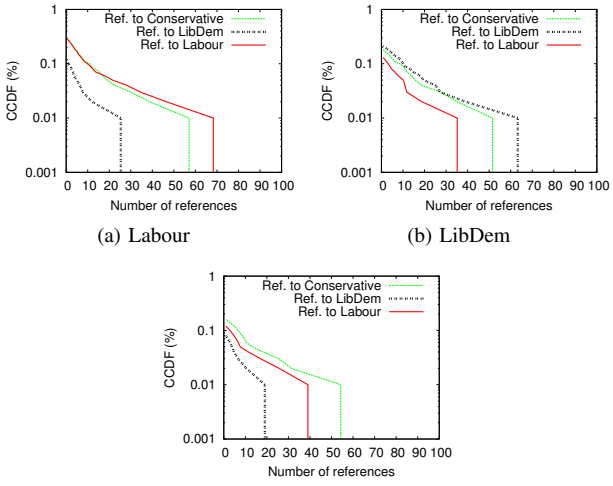
TABLE IV: *Ten most commonly used URLs.*

typad.com) have been taken into account. We compared the usage rates of these blogs between parties. Table V 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 [8].

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

TABLE V: *Three most cited blogs per party.*

Finally, 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 7 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 leader.



(a) Labour (b) LibDem

(c) Conservative

Fig. 7: *CCDF for the volume of references.*

E. Sentiment

We evaluated the sentiment of words used in tweets. To extract this information we used the Linguistic Inquiry Word Count[2]. LIWC is a dictionary of words used in everyday conversations which assess 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, good) and negative emotions (i.e. out, hate). Then, the sentiment for a given tweet was given by the sentiment score proposed by Kramer [9]: $Sentiment = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n}$ 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 VI shows the average sentiment score over 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.

Party	Reference to	average emotion score
Labour	Labour	**1.09**
	LibDem	0.03
	Conservative	0.32
LibDem	Labour	-0.21
	LibDem	**0.34**
	Conservative	0.00
Conservative	Labour	-0.08
	LibDem	-0.14
	Conservative	**1.36**

TABLE VI: *Sentiment on the references to party.*

IV. User Classification

In this section we present a new user classification approach based on the observations in the previous section. Our goal is to identify the party to which a user belongs. We particularly focus on developing a classification method without the knowledge about network topology. For this purpose, we propose an incremental Bayesian approach which requires only a user's tweet messages over time. We will show this approach performs well by evaluating the performance of the classification method.

A. Bayesian Classification

Without loss of generality, we assume that a sequence of tweet activities (e.g. retweets or references to a specific party/leader in tweets) 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 $1 \leq k \leq n$ 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

[2]An online version of LIWC is available at www.liwc.net

i after the kth time interval. We assume that all users should be included to one of parties; $\sum_i P(M_k^i(u)) = 1$. 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:

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

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 $M_k^i(u)$.

To calculate $P(A_k(u)|M_k^i(u))$, we consider the frequency of *referring to political parties in tweets* for $A_k(u)$ based on the observation in the previous section[3].

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:

$$P(A_k(u)|M_k^i(u)) = \frac{\sum_{t \in T} V_i(t)}{|T|}$$

where T is the tweets of the current user during the 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** to denote this Bayesian classification.

B. Evaluation

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 performance 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 referencing 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 a good emotion in the user's tweets for a party when the user prefers the party. We compute a user's sentiment scores of parties through the sentiment analysis of the user's tweets and then assigned the party with the best 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. This approach detects the communities of users using a label propagation method [6] 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 iteratively this process until all users' parties are classified.
- **Follower classifier**: The relationship of following and being followed in Twitter can reflect the political leanings of users as well [5]. 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.
- **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 [3], [10]: (i) the list of followers, (ii) the list of friends, (iii) the list of retweeted users, (iv) the list of used words in the user's tweets, (v) the list of used hashtags in the user's tweets, and (vi) the emotion over the user's tweets.

To show the performance of a classifier, we measured their *accuracy* for the self-identified users (1161). The classification accuracy is defined as the ratio between the number of correctly predicted samples; the results are shown in Table VII. Classifiers used tweets and relationships related to these self-identified users. These users published 27,696 tweets, 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, and SVM) require a training step used to learn the features determining political party membership and/or the knowledge about network topology. Training samples are composed of one-tenth of the ground truth users (115) to construct the classifiers and the rest (1,046) was reserved for out-of-sample testing.

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 when it leverages its incremental approach over time with 10 updates of the users' affiliation probabilities during the period (accuracy of 0.86).

[3]We have tested other potential alternatives, but given the space limitations, we describe this that led to the best classification performance.

Classifier	Accuracy
Volume	0.62
Sentiment	0.67
Follower	0.83
Retweet	0.81
SVM	0.77
Bayesian	**0.86**

TABLE VII: *Performance according with approach.*

Approach	Random	Most active	Most influent
Follower	0.80	0.77	**0.83**
Retweet	0.72	0.76	**0.81**
SVM	**0.80**	0.69	0.77

TABLE VIII: *Variation of the accuracy according to seeds.*

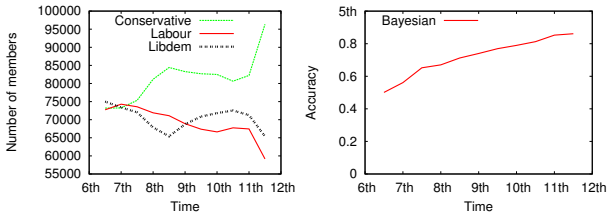
We used fixed time interval of 15 hours to periodically updates the users' affiliation probabilities according to their tweets in the associated interval. We note that this classification benefits from two advantages. Firstly, it requires to maintain only the affiliation probability of each user without massive training overheads and secondly, as the information about references to a party or a leader 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 should be used as an alternative when the conditions do not allow the use of Follower which requires the knowledge about network topology to achieve good results, which may indeed be unacceptable or very expensive: crawling topology information is strictly limited in practice. Unlike our expectations, SVM which involves an expensive tuning phase, did not outperform other algorithms.

In addition, we analysed the accuracy of these classifiers according to the set of training samples among (i) the most influential users with the highest number of followers, (ii) the most active users with the highest number of published tweets and, (iii) random users. Results are depicted in Table VIII, the training sample based on the most influential users provide the best accuracy for the Follower and the Retweet classifiers. Indeed, these classifiers require hubs or important users as seeds to start label or probability propagation. In contrast, as the SVM classifier aims to build a model reflecting the behavior of all users part of the same political party, accounting the behavior of various users is more useful than to select only the most active or influent ones.

We also analysed both how the number of partisans of each party evolves over all the 220,000 users of our dataset and how the accuracy of the proposed Bayesian classifier changes with time over the self-identified users. 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.

V. Related Work

The exponential growth and the ubiquitous trend of social media has attracted much attention.



(a) Numbers of members (b) Classification accuracy

Fig. 8: *Dynamic changes of the Bayesian classifier over time.*

A. Classification

Different approaches have been proposed for classifying users in many directions. [11] presented a semi-supervised algorithm for classifying political blogs. [4] also applied three semi-supervised algorithms for classifying political news articles and users, respectively. Their propagation algorithm 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. [5] presented a method that uses the follower connections in Twitter to compute political preferences. This method achived similar results than the label propagation method on the retweet graph in this paper.

[8] 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 [12]. They found that the blogosphere exhibits a politically segregated community structure with more limited connectivity between different communities. Recently, [7] observed a similar structure in a retweet graph of Twitter in politic context. Other classifications used machine learning methods to infer information on users. [3] 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. [10] used Gradient Boosted Decision Trees which is a machine learning technique for regression problems, which produces 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 [3], [4], [5] that require expensive costs for tuning of parameters to construct classifier and/or the knowledge about network topology. This is an extended paper of our preliminary work [13].

B. Characterization

Characterization aims to identify the main characteristics of population. Several studies have addressed to characterise user behaviour or personality in social networks [14], [15]. However few works have tried to study the characteristics of politic parties and the interaction structure between parties.

[16], [17] 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. [18] studied the usage patterns of tweets about the candidates in the 2010 U.S. midterm elections and showed stronger cohesiveness among Conservative and Tea party.

C. Prediction

Other studies have addressed the predictive power of the social media. [19] demonstrated how social media contents can be used to predict real-world outcomes and outperformed market-based predictor variables. In Politics, [18] has investigated the relation between the network structure and tweets and presented a forecast of the 2010 midterm elections in the US. [1] 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 [20] demonstrated that this result was not repeatable with the 2010 US congressional elections.

D. Sentiment analysis

[21] used sentiment analysis to compare Twitter streams with polls in different areas and showed the correlation on some points. [22] 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. [23] showed that tweets can be used to track real-time sentiment about candidates' performance during a televised debate. [24] 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.

VI. 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 changes 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 new user classification approach using Bayesian framework which can incrementally update the classification results with time. Moreover, this approach does not require the knowledge about network topology unlike the previous solutions [6], [5].

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: (*i*) the retweet graph presented a highly segregated partisan structure (*ii*) party members were more likely to make reference to their own party than another, and (*iii*) members were more likely to express more positive opinions when they referenced their own party. Through these party characteristics, we built 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 network topology information which make it a proper solution for real time classification.

VII. Acknowledgment

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

[1] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter : What 140 characters reveal aboutpolitical sentiment," in *ICWSM'10*, 2010.

[2] M. Cha, H. Haddadi, F. Benevenuto, and P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *ICWSM'10*, 2010.

[3] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: User classification in twitter," in *KDD'11*, 2011.

[4] D. X. Zhou, P. Resnick, and Q. Mei, "Classifying the political leaning of news articles and users from user votes," in *ICWSM'11*, 2011.

[5] J. Golbeck and D. Hansen, "Computing political preference among twitter followers," in *CHI '11*, 2011.

[6] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks." *Physical Review E*, 2007.

[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer, "Political polarization on twitter," in *ICWSM'11*, 2011.

[8] L. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *LinkKDD'05*, 2005.

[9] A. D. Kramer, "An unobtrusive behavioral model of "gross national happiness"," in *CHI'10*, 2010.

[10] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification," in *ICWSM'11*, 2011.

[11] F. Lin and W. W. Cohen, "The multirank bootstrap algorithm: Self-supervised political blog classification and ranking using semi-supervised link classification," in *ICWSM'08*, 2008.

[12] M. Hewstone and R. Brown, *Contact is not Enough: An Intergroup Perspective on the "Contact Hypothesis"*, 1986.

[13] A. Boutet, H. Kim, and E. Yoneki, "What's in your tweets? i know who you supported in the uk 2010 general election (poster paper)," in *ICWSM'12*.

[14] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *IMC'09*, 2009.

[15] D. Quercia, R. Lambiottez, D. Stillwell, M. Kosinski, and J. Crowcroft, "The personality of popular facebook users," in *CSCW'12*, 2012.

[16] Y. Sarita and B. Danah, "Dynamic debates: An analysis of group polarization over time on twitter," in *Bulletin of Science, Technology and Society*, 2010.

[17] J. An, M. Cha, K. Gummadi, and J. Crowcroft, "Media landscape in Twitter: A world of new conventions and political diversity," in *ICWSM'11*, 2011.

[18] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic, "The party is over here: Structure and content in the 2010 election," in *ICWSM'11*, 2011.

[19] S. Asur and B. A. Huberman, "Predicting the future with social media," in *WI-IAT '10*, 2010.

[20] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj, "Limits of electoral predictions using twitter," in *ICWSM'11*, 2011.

[21] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *ICWSM'10*, 2010.

[22] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "In the mood for being influential on Twitter," in *SocialCom'11*, 2011.

[23] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *CHI'10*, 2010.

[24] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "Tracking gross community happiness from tweets," in *CSCW'12*, 2012.