What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election

Antoine Boutet
INRIA Rennes Bretagne Atlantique
France

Hyoungshick Kim and **Eiko Yoneki**
Computer Laboratory,
University of Cambridge
UK

Abstract

Nowadays, the use of social media such as Twitter is necessary to monitor trends of people on political issues. As a case study, we collected the main stream of Twitter related to the 2010 UK general election during the associated period. We analyse the characteristics of the three main parties in the election. Also, we propose a simple and practical algorithm to identify the political leaning of users using the amount of Twitter messages which seem related to political parties. The experimental results showed that the best-performing classification method – which uses the number of Twitter messages referring to a particular political party – achieved about 86% classification accuracy without any training phase.

Introduction

We are interested in how to measure the authority of political parties and the political leaning of users from social media. To illustrate the practicality of our analysis, we used a dataset formed of collected messages from Twitter related to the 2010 UK general election which took place on May 6th, 2010.

We examined the characteristics of the three main parties (Labour, Conservative, Liberal Democrat) and discussed the main differences between parties in term of structure, interaction, and contents.

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 Twitter – the messages expressing the user's political views (e.g. tweets referring to a particular political party and retweets from users with known political preferences) are 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. The experimental results showed that the proposed classification method – which uses the number of tweets referring to a particular political party – achieved about 86% classification accuracy using all trials without any training phase, outperforming existing heuristics (Pennacchiotti and Popescu 2011a; Zhou, Resnick, and Mei 2011) that require expensive costs for tuning of parameters to construct classifier.

Our approach has two key advantages: (1) as we only process the messages relevant to a particular event rather than the whole dataset at one time, it drastically reduces the computation costs of constructing a classifier compared with existing approaches which may indeed be unacceptable for online classification; (2) it also has potential: we can discover the temporal trends of a user's political views by analysing her political leaning over time.

Party Characteristics

To analyse the characteristics of the Labour, Conservative and Liberal Democrat (LibDem) parties to identify the relevant features for user's party affiliation, 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. For these users, we also collected their profiles and about 79,000,000 following/follower relationships. Some user profiles can be used to identify their political party affiliation. We manually identified the 356 Labour, 159 Conservative and 169 LibDem self-identified members as a ground truth dataset.

With this ground truth dataset, we detected the communities associated to each political party using a well-known technique called label propagation method (Raghavan, Albert, and Kumara 2007). 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 need to set the *maximum propagation distance* to k to avoid tie-breaking cases (i.e. multiple nearest nodes with different party memberships exist at the same time). We performed the label propagation until the propagation distance is greater than k. When $k = 2$, we detected 5,878 Labour, 3,214 LibDem and 2,356 Conservative candidates with 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: (*i*) structure/interaction and (*ii*) content features.

Structure and Interaction We studied the differences between the political parties in network structure and inter-

action patterns. The interaction patterns between members within a party reflects a level of party cohesion while the interaction patterns between different communities reflect the exchanges (i.e. conflict or collaboration) between them.

We particularly observed the amount of interactions between the political parties by counting the number of exchanged retweets (forward messages to its followers) and mentions (direct messages to another user) between them during the election period.

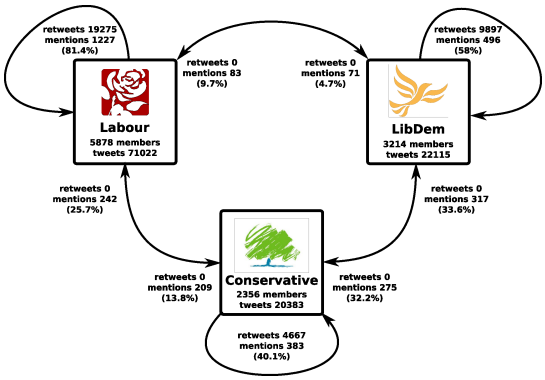


Figure 1: Exchanged messages between parties

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

Content We analysed the contents of tweets by counting the number of hashtags (tags used to define topics) and URLs used in tweets for each party. 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. 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.

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. 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. LibDem members more frequently referred to *Financial Times*, *The Independent* and *The BBC* compared with the other party members. We also observed the blogs which are usually more politically oriented. 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).

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. The analysis clearly shows that users were more likely to frequently refer to their own preferred party or leader.

User Classification

For user classification, our goal is to the party to which a user belongs. We particularly focused on developing a classification method to process the dynamically updated statistics on users; user's tweet activities are sequentially observed over time.

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 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 two tweet activities for $A_k(u)$, respectively, based on the observation in the previous section: (1) *retweeting the messages from the members of each political party* and (2) *referring to political parties in tweets*.

We can see that tweets generated by users supporting the same political party were more frequently retweeted. For this activity, we assume $P(A_k(u)|M_k^i(u))$ can be calculated as follows:

$$P(A_k(u)|M_k^i(u)) = \frac{\sum_{v \in RT} P(M_{(k-1)}^i(v))}{|RT|} \quad (1)$$

where RT is the set of the users included in retweets as the source or the destination of information. We use **Bayesian-Retweet** to denote the Bayesian classification where $P(A_k(u)|M_k^i(u))$ is defined in (1).

The other important tweet activity is to generate a tweet referring to the political party (or party leader) that the user u will support after the kth time interval since party members are more likely to make reference to their own party than another. For this activity, we assume $P(A_k(u)|M_k^i(u))$ can be calculated as follows:

$$P(A_k(u)|M_k^i(u)) = \frac{\sum_{t \in T} V_i(t)}{|T|} \quad (2)$$

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-Volume** to denote the Bayesian classification where $P(A_k(u)|M_k^i(u))$ is defined in (2).

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**: We 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.
- **Retweet classifier**: This approach detects the communities of users using a label propagation method (Raghavan, Albert, and Kumara 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 iteratively this process until all users' parties are classified.
- **SVM classifier**: Support Vector Machine (SVM) is known as one of the best supervised learning techniques for solving classification problems. We constructed a SVM classifier using the following six features of a user proposed in (Pennacchiotti and Popescu 2011a; 2011b): (i) the number of followers, (ii) the number of replied users, (iii) the number of retweeted users, (iv) the number of used words in the user's tweets, (v) the number of used hashtags in the user's tweets, and (vi) the average emotion over the user's tweets.

To show the performance of a classifier, we measured the *accuracy* of the classifier for the self-identified users. The classification accuracy is defined as the ratio between the number of correctly predicted samples and the total number of testing samples.

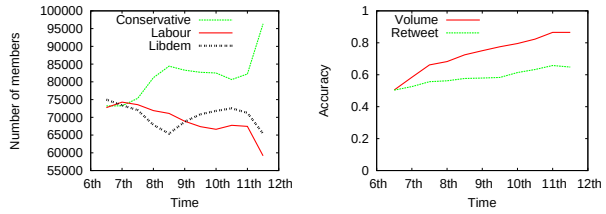
For the classifiers requiring the training samples (Retweet, SVM, and Bayesian-Retweet), one-tenth of the ground truth users was used to construct the classifiers and the rest was reserved for out-of-sample testing. We have seen that the accuracy of these classifiers can be changed with the set of training samples. We used the most influential users with the highest number of followers since these training samples provide the best accuracy compared to the most active users with the highest number of generated tweets or random users.

Classifier	Accuracy
Volume	0.60
Retweet	0.73
SVM	0.63
Bayesian-Retweet	0.64
Bayesian-Volume	**0.86**

Table 1: Performance according with approach.

The Bayesian-Volume produced the best results: the measure accuracy (0.86) is significantly higher compared to other classification methods. In addition, 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. Unlike our expectations, SVM which involves an expensive tuning phase, did not outperform other algorithms.

We also analysed how the number of partisans of each party and the accuracy of the proposed Bayesian classifiers, respectively, changes with time. The results are shown in Figure 2. Conservative members outnumbers the Labour and LibDem members at the end of the election. Inherently, the accuracy of Bayesian-Volume and Bayesian-Retweet starts at 1/2 (equiprobability), continuously increases with time, and achieved at 0.86 and 0.64, respectively. These results imply that the proposed Bayesian approach is proper to understand users' political leaning over time. However, the accuracy for Bayesian-Retweet increases more slowly than Bayesian-Volume. Even if a retweet graph generally presents a high segregated structure, latency might be expected to sufficiently propagate retweets over the graph. In contrast, Bayesian-Volume, which uses only the tweets published by users during the current time interval, achieves accurate prediction without latency caused by message propagation between users.



(a) Numbers of members (b) Classification accuracy

Figure 2: Dynamic changes of the Bayesian classifiers.

Related Work

The exponential growth of social media has attracted much attention. 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, Resnick, and Mei 2011) also applied three semi-supervised algorithms for classifying political news articles and users, respectively. On the other hand, (Adamic and Glance 2005) studied the linkage patterns between political blogs and 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 politic 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. (Pennacchiotti and Popescu 2011b) 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.

Several studies have addressed to characterise user behaviour or personality in social networks (Benevenuto et al. 2009). However few works have tried to study the characteristics of politic parties and the interaction structure between parties. (Livne et al. 2011) 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.

Other studies have addressed the predictive power of the social media. (Livne et al. 2011) has investigated the relation between the network structure and tweets and presented a forecast of the 2010 midterm elections in the US, and (Tumasjan et al. 2010; Gayo-Avello, Metaxas, and Mustafaraj 2011) discussed the relevance of Twitter as a valid indicator of political opinion.

(O'Connor et al. 2010) used sentiment analysis to compare Twitter streams with polls in different areas and showed the correlation on some points. (Diakopoulos and Shamma 2010) showed that tweets can be used to track real-time sentiment about candidates' performance during a debate.

Conclusion

As a case study, we first analysed the characteristics of the political parties in Twitter during the 2010 UK General Election and identified the two main ways to differentiate political parties: (*i*) the retweet graph presented a highly segregated partisan structure, and (*ii*) party members were more likely to make reference to their own party than another. Through these party characteristics, we built two classification algorithms based on Bayesian framework. The experimental results showed that the proposed classification method is capable of achieving an accuracy of 86% without any training which make it a perfect solution for real time classification.

References

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

Benevenuto, F.; Rodrigues, T.; Cha, M.; and Almeida, V. 2009. Characterizing user behavior in online social networks. In *IMC'09*.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Political polarization on twitter. In *ICWSM'11*.

Diakopoulos, N. A., and Shamma, D. A. 2010. Characterizing debate performance via aggregated twitter sentiment. In *CHI'10*.

Gayo-Avello, D.; Metaxas, P. T.; and Mustafaraj, E. 2011. Limits of electoral predictions using twitter. In *ICWSM'11*.

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

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

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

Pennacchiotti, M., and Popescu, A.-M. 2011a. Democrats, republicans and starbucks afficionados: User classification in twitter. In *KDD'11*.

Pennacchiotti, M., and Popescu, A.-M. 2011b. A machine learning approach to twitter user classification. In *ICWSM'11*.

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

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

Zhou, D. X.; Resnick, P.; and Mei, Q. 2011. Classifying the political leaning of news articles and users from user votes. In *ICWSM'11*.