Member Classification and Party Characteristics in Twitter during UK Election

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

Eiko Yoneki
University of Cambridge, UK
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, real time members classification and party characterization taking into account the dynamic nature of social media are essential to highlight the main differences between parties and to monitor their activities, influences, structures, contents and mood. This paper summarizes a case study of member classification and party characteristics in Twitter during the UK election.

Keywords — Social Media, User Classification

1 Introduction

Using social media for political discourse is becoming common practice, especially around election time [4]. Different off-line approaches have been proposed for classifying the political leaning of users [3, 2]. However, inferring the political orientation of members in real time needs (1) to avoid collecting and using a prohibitive quantity of data for classification and (2) to consider the highly dynamic nature of social media including changes in term of active users, content or social structure.

On the other hand, while several studies have addressed the characterization of user in social networks [1], few works have tried to characterize politic parties. This characterization permits to highlight the main differences between parties, their impact in term of activity and influence, their preferred content, their mood and sentiment, their structure and how they exchange between each other over time. For instance, this information can be useful for political parties or candidates to monitor their impact and to assess and adapt their strategy.

This paper summarizes a case study of member classification and party characterization in Twitter during the 2010 UK general election. By looking at temporal changes in activity, Twitter is shown to react instantly to UK election events. To highlight the main differences between parties, we first characterize the self-identified users as member of a specific party according to several categories. Then, we present a real-time method for deducing user's party affiliation using only basic textual and semantic analysis of the public stream of Twitter. Finally, we discuss key parameters to address the dynamic nature of social media in online classification.

2 Party Characteristics

To highlight the main differences between parties, we first characterize the self-identified populations (users self-identified as members of specific parties in their Twitter user description) according to several categories: (1) the activity reflects the commitment of the parties to send tweets and to take place in political debates ; (2) the influence gives an indication on the potential impact of each party ; (3) the social structure reflects both the cohesiveness of the party and the level of exchange and debate between them ; (4) the content shows how the party used hashtags and urls, and the volume of references did to a specific party or political figure ; (5) the sentiment analysis evaluates the sentiment of words in tweets through different directions (self-focus, cognitive, social, positive and negative).

We saw that only few characteristics differ one party to another and made the following observations:

- **Activity.** Except for a small fraction of users with particular behavior in term of generated tweets, retweets or mentions, the majority of members from all parties exhibited a similar activity.
- **Influence.** Labour members had a better influence, however, Conservative members were better organized to promote their party.
- **Structure.** The retweet graph presents a highly segregated partisan structure where very few links connected different parties. In contrast, the mention graph formed a more heterogeneous structure where no particular clusters were observed.
- **Content.** The usage of hashtags depended on the underlying neutrality of the tag. In addition, each party was more likely to cite certains websites than others especially for blogs which were more political orientated. The similarity in terms of hashtags used showed a high heterogeneity between users in all parties. Moreover, we demonstrated that party members were more likely to make reference to

their own party than another.

- **Sentiment.** No particular sentiment direction permitted the differentiation of parties. However we clearly saw that members were more likely to express more positive opinion when they referenced its own party.

3 User Classification

The goal of this classification is to identify the party to which each person belongs and their evolution from the pattern of tweets. We consider the observed tweet process by events $E \in \{E_n\}$ representing the current state of beliefs of the political affiliation of members. Each affiliation probability is represented by an event, $Am \in \{labour, libdem, conservative\}$ for each user where $labour + libdem + conservative = 1$, containing elementary events $A_m \cap E_n$. The conditional probabilities or evidences $P(E_n|A_m)$ are specified to define the affiliation probability. $P(A_m)$ is the uncertainty of model A_m. Before the first inference step, the initial prior probability is set uniformly: $\{P(A)\} = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$. For each affiliation probability Am, $P(A_m)$ is updated to $P(A_m|E_n)$. From Bayes' theorem:

$$P(A_m|E_n) = \frac{P(E_n|A_m)P(A_m)}{P(E_n)} \tag{1}$$

where $P(A_m|E_n)$ is the posterior, the uncertainty of A_m after E_n is observed ; $P(A_m)$ is the prior, the uncertainty of A_m before E_n is observed ; and $\frac{P(E_n|A_m)}{P(E_n)}$ is a factor representing the impact of E_n on the uncertainty of A_m.

We define two evidences based on the main characters identified in the party charateristics analysis. The first is based on the pattern of retweets and highlights the fact that people of similar political persuasion retweet roughly the same thing (highly segregated partisan structure of the retweet graph). Retweets can be easily identified in the tweets stream thanks to the keyword *RT*. This evidence is defined as the average affiliation probability of both people retweeted by the user or people retweeting the user during the period $[E_{n-1}, E_n]$:

$$P(E_n|A_m = a) = \frac{\sum_{R \in Retweets} P(A = a)_R}{\textit{Number of retweets}} \tag{2}$$

The second exploits the fact that party members are more likely to make reference to their own party than another. This second evidence is defined as the ratio of tweets referencing one party over the total number of references during the period $[E_{n-1}, E_n]$:

$$P(E_n|A_m = a) = \frac{Volume(A_m = a)}{\sum Volume(A_m)} \tag{3}$$

Approach		Accuracy
Bayesian	Retweet	0.70
	Volume	0.80

Table 1: Performance of our user classification

Finally we define the political affiliation of user according to the best probability in A_m (i.e. $[0.7, 0.2, 0.1]$ refers a person who is probably Labour). Ties are broken randomly in case of equiprobabilities.

To evaluate our classification algorithm, initial seeds are chosen from the 10% most active users in term of generated tweets from those which are self-identified as member of specific party. We then measure the accuracy by calculating the number of correct predictions of the remaining 90% over the total number of predictions. Table 1 shows the high accuracy produced by our member classification.

4 Discussions and Conclusions

We summarize here a case study of member classification and party characteristics from Twitter during the 2010 UK general election. The three main differences identified between political parties are (1) the retweet graph presented a highly segregated partisan structure where very few links connect different parties, (2) party members were more likely to make reference to their own party than another, and (3) members were more likely to express more positive opinion when they referenced their own party.

To be able to act in real time, a classification system must not require a prohibitive amount of data. Furthermore, due to the dynamic nature of social media, active users and content can change quickly. In this context, collecting the social structure of every new user or performing linguistic analysis for each new hashtag is inconceivable. The proposed member classification exhibits good performance, can be easily used in real time and fits perfectly to the dynamic nature of social media.

References

[1] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 49–62, New York, NY, USA, 2009. ACM.

[2] Qiaozhu Mei Daniel Xiaodan Zhou, Paul Resnick. Classifying the political leaning of news articles and users from user votes. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[3] Ana-Maria Popescu Marco Pennacchiotti. Democrats, republicans and starbucks afficionados: User classification in twitter. In *KDD*, 2011.

[4] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, pages 178–185, 2010.