Analysing Political Events on Twitter: Topic Modelling and User Community Classification

Anjie Fang
Glasgow, UK
a.fang.1@research.gla.ac.uk

Abstract

Recently, political events, such as elections, have raised a lot of discussions on social media networks, in particular, Twitter. This brings new opportunities for social scientists to address social science tasks, such as understanding what communities said or identifying whether a community has an influence on another. However, identifying these communities and extracting what they said from social media data are challenging and non-trivial tasks.

We aim to make progress towards understanding 'who' (i.e. communities) said 'what' (i.e. discussed topics) and 'when' (i.e. time) during political events on Twitter. While identifying the 'who' can benefit from Twitter user community classification approaches, 'what' they said and 'when' can be effectively addressed on Twitter by extracting their discussed topics using topic modelling approaches that also account for the importance of time on Twitter. To evaluate the quality of these topics, it is necessary to investigate how coherent these topics are to humans. Accordingly, we propose a series of approaches in this thesis.

First, we investigate how to effectively evaluate the coherence of the topics generated using a topic modelling approach. The topic coherence metric evaluates the topical coherence by examining the semantic similarity among words in a topic. We argue that the semantic similarity of words in tweets can be effectively captured by using word embeddings trained using a Twitter background dataset. Through a user study, we demonstrate that our proposed word embedding-based topic coherence metric can assess the coherence of topics like humans [1, 2]. In addition, inspired by the *precision at k* metric, we propose to evaluate the coherence of a topic model (containing many topics) by averaging the top-ranked topics within the topic model [3]. Our proposed metrics can not only evaluate the coherence of topics and topic models, but also can help users to choose the most coherent topics.

Second, we aim to extract topics with a high coherence from Twitter data. Such topics can be easily interpreted by humans and they can assist to examine 'what' has been discussed and 'when'. Indeed, we argue that topics can be discussed in different time periods (see [4]) and therefore can be effectively identified and distinguished by considering their time periods. Hence, we propose an effective time-sensitive topic modelling approach by integrating the time dimension of tweets (i.e. 'when') [5]. We show that the time dimension helps to generate topics with a high coherence. Hence, we argue that 'what' has been discussed and 'when' can be effectively addressed by our proposed time-sensitive topic modelling approach.

Next, to identify 'who' participated in the topic discussions, we propose approaches to identify the community affiliations of Twitter users, including automatic ground-truth generation approaches and a user community classification approach. We show that the mentioned hashtags and entities in the users' tweets can indicate which community a Twitter user belongs to. Hence, we argue that they can be used to generate the ground-truth data for classifying users into communities. On the other hand, we argue that different communities favour different topic discussions and their community affiliations can be identified by leveraging the discussed topics. Accordingly, we propose a Topic-Based Naive Bayes (TBNB) classification approach to classify Twitter users based on their words and discussed topics [6]. We demonstrate that our TBNB classifier together with the ground-truth generation approaches can effectively identify the community affiliations of Twitter users.

Finally, to show the generalisation of our approaches, we apply our approaches to analyse 3.6 million tweets related to US Election 2016 on Twitter [7]. We show that our TBNB approach can effectively identify the 'who', i.e. classify Twitter users into communities. To investigate 'what' these communities have discussed, we apply our time-sensitive topic modelling approach to extract coherent topics. We finally analyse the community-related topics evaluated and selected using our proposed topic coherence metrics.

Overall, we contribute to provide effective approaches to assist social scientists towards analysing political events on Twitter. These approaches include topic coherence metrics, a time-sensitive topic modelling approach and approaches for classifying the community affiliations of Twitter users. Together they make progress to study and understand the connections and dynamics among communities on Twitter.

Supervisors: Iadh Ounis, Craig Macdonald, Philip Habel

The thesis is available at `http://theses.gla.ac.uk/41135/`

References

[1] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. Topics in tweets: A user study of topic coherence metrics for Twitter data. In *Proc. of ECIR*, 2016.

[2] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. Using word embedding to evaluate the coherence of topics from Twitter data. In *Proc. of SIGIR*, 2016.

[3] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. Examining the coherence of the top ranked tweet topics. In *Proc. of SIGIR*, 2016.

[4] Anjie Fang, Iadh Ounis, Craig MacDonald, Philip Habel, Xiaoyu Xiong, and Hai-Tao Yu. An effective approach for modelling time features for classifying bursty topics on Twitter. In *Proc. of CIKM*, 2018.

[5] Anjie Fang, Craig Macdonald, Iadh Ounis, Philip Habel, and Xiao Yang. Exploring time-sensitive variational bayesian inference LDA for social media data. In *Proc. of ECIR*, 2017.

[6] Anjie Fang, Iadh Ounis, Philip Habel, Craig Macdonald, and Nut Limsopatham. Topic-centric classification of Twitter user's political orientation. In *Proc. of SIGIR*, 2015.

[7] Anjie Fang, Philip Habel, Iadh Ounis, and Craig MacDonald. Votes on twitter: Assessing candidate preferences and topics of discussion during the 2016 us presidential election. *SAGE Open*, 9(1):1–17, 2019.