Predicting Ideological Friends and Foes in Twitter Conflicts

Zhe Liu*
College of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania 16802
zul112@ist.psu.edu

Ingmar Weber
Qatar Computing Research Institute
PO Box 5825, Doha, Qatar
iweber@qf.org.qa

ABSTRACT

The rise in popularity of Twitter in recent years has in parallel led to an increase in online controversies. To monitor and control such conflicts early on, we design and evaluate a language-agnostic classifier to tell pairs of ideological friends from foes. We build the classifier using features from four different aspects: user-based, interaction-based, relationship-based and conflict-based. By experimenting with three large data sets containing diverse conflicts, we demonstrate the effectiveness of language-agnostic classification of ideological relation, achieving satisfactory results across all three data sets. Such a classifier potentially enables studies of diverse conflicts on Twitter on a large scale.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

political ideology,conflict,classification,Twitter,social network

1. INTRODUCTION

As more and more people start engaging regularly in discussions about controversial issues online, "digital hate" on Twitter grows at an alarming speed. According to Simon Wiesenthal Center's report [4], Twitter has encountered an incredible 30 percent surge in this kind of traffic in 2012. Tension on Twitter can also result in irreconcilable online or offline chaos, such as the Arab Spring uprisings and the Occupy Wall Street movement. In fact, some research [3, 5] hints at social media facilitating the actual mass aggregations and outbreak of violence in several contexts. Given the rising tide of Internet violence, we propose a classifier to predict the ideological relation, friend or foe, between any two Twitter users. We build our classifier in a language-agnostic manner by using only contextual clues such as following information or the frequency of interactions. We chose three conflicts of very different natures to build and evaluate the classifier: the Palestine-Israel conflict, the Democrat-Republication political polarization, and the FC Barcelona-Real Madrid football rivalry. In all three cases, we identify likely supporters of either camp by using retweet signals.We validate our classification model on all three data sets and show that using only language-agnostic features achieves satisfactory classification results on ideological relations.

*Most of this work was done while the first author was at Qatar Computing Research Institute.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576964.

2. DATA SETS

To obtain the data, we started with three pairs of key official accounts with opposing orientations as our seed users, including "@AlqassamBrigade" and"@IDFSpokesperson" from the conflict of Palestine vs. Israel (PA vs. IL), "@TheDemocrats" and "@GOP" from Democrat vs. Republican (DEM vs. REP), and "@FCBarcelona_es" and "@realmadrid" from Barcelona vs. Real Madrid (FCB vs. RMCF). We intentionally chose these seed nodes because, first, they are all high-profile figures which are actively involved in the conflicts. Second, these seed users covered conflicts that happen in different parts of the world across different topics. This allows us to test the generalizability of our proposed classifier.

For each of our seed users we obtained their latest 3,200 publicly available tweets and for each tweet we identified up to 100 retweeters. As retweet often indicates endorsement and preference of a message, we labeled those retweeters as the ideological-friends of the corresponding seed account, and obtained their public tweets. We removed neutral nodes, such as journalists and news publishers, based on their lower relative difference in total retweeting frequencies of the top most popular accounts from each side. In addition, we also removed non-active members in each conflict according to their lower retweeting frequencies.

We validated our labeling results via CrowdFlower[1]. For each conflict, we created a job, including 100 users randomly selected from each side. For each sampled user, we randomly selected 10 of their related tweets based on chosen keywords (such as "Palestine", "Israel", "Hamas", and their translations in 5 other most frequently used languages in a conflict). All non-English tweets were detected with Google's Compact Language Detector[2] and translated using Google Translate[3]. With both a user's profile and related tweets displayed in a HIT, we asked the workers to identify the ideology of the listed user. Each HIT was distributed to three workers and the final label for each HIT was decided via majority voting. To control

[1] `http://crowdflower.com/`
[2] `http://code.google.com/p/chromium-compact-language-detector/`
[3] `http://translate.google.com/`

the trustworthiness of the workers, we created 20% of pre-labeled HITs as "gold standard" data. By comparing user's pre-assigned ideology to the majority-voted label obtained from CrowdFlower, we found that our retweet-based labeling method worked well over all three data sets, yielding an average accuracy of 96.2%.

Table 1 presents the overall statistics for our collected data sets. Since we have much less inter-ideological communications that intra-ideological ones, to avoid classification bias and get better results, we balanced our data set by taking all cross-party interactions plus the same number of randomly selected within-party records.

Conflict	#Users	#Inter-Interactions	#Intra-Interactions	#Interaction Classification
PA - IL	9,937	4,829	178,255	9,658
DEM - REP	17,869	20,257	576,848	40,514
FCB - RMCF	28,218	21,089	152,799	42,178

Table 1: Data Statistics

3. AUTOMATIC RELATION DETECTION

In total, we proposed 74 features which could be used in the context of ideological friend-foe identification. We grouped these features into four categories based on their formation characteristics. Many of the features follow previous works [1]. All features are extracted from both sides of Twitter interactions: the initiators (U_a) and the receivers (U_b).

User-based features depict both the explicit characteristics of a user, as well as their style of tweeting, including: the number of followers, followees, tweets, percentage of tweets containing mentions, formal retweets (retweet by clicking on the "retweet" button), and informal retweets.

Interaction-based features characterized the communication patterns between two user, including: one user mentions the other, formally retweets the other, the position of the mention, and the average length of the mention tweet.

Relationship-based features covered the bidirectional Twitter relations between two users. These include: if one user is following the other, as well as if one user is followed by the other.

Conflict-based features specified a user's involvement in a conflict. We simply defined the engagement level as the percentage of times that the seed user with the same ideology has been mentioned. Similarly, we defined the level of aggressiveness as the percentage of times that the seed user with the opposing ideology has been mentioned. Note that these features require the preexisting knowledge that a conflict exists, along with the identity of Twitter users central to this conflict. We decided to include this feature set to gauge the value that such information could have.

We did not include any language-dependent features, such as insulting or cursing words, repeated words, or even capital letters in our classification method. We deemed this necessary because we are looking for an approach to be used across different languages in an ever increasingly globalized world.

4. EXPERIMENTS

Based on all four sets of features proposed, we next built a binary classifier to automatically label cross and within-ideological relationship in an interaction. We tested classification algorithms implemented in Weka, such as naive Bayes, Bayes network, SVM and J48 decision tree etc, all with default values for all parameters using 10-fold cross-validation. For all three data sets, the J48 decision tree algorithm achieved the best classification performance. A summary of the results is shown in Table 2. We found that our proposed classification method outperformed the simple baseline model, which always simply predicts the majority class (50% precision, recall and F_1 given our balanced data sets). Given that only language-agnostic features were included in our classifier, our classifier on all three data sets achieved satisfactory performance in identifying friends from foes, especially in the Palestine-Israel conflict. We think this might be related to the level of conflict intensity, although this needs to be proved in future studies.

Apart from the overall performance of our proposed model, we further identified the most informative features in classification of ideological relations. Table3 shows the top 10 features in each of our data set. We noticed that the relationship-based features have the strongest discrimination power in two political data sets, whereas, to our surprise, we found that the conflict-based features are not that useful (except in PA - IL), which pointed out that relying only on the mentioning patterns of the super-active users cannot effectively differentiate ideological friends from foes. In addition, interaction-based features, such as whether or not two users retweet each other also revealed significant potential for accurate friend-foe identification, and this is in line with findings in [2].

Conflict	TP Rate	FP Rate	Prec.	Recall	F_1	ROC
PA-IL	0.97	0.03	0.97	0.97	0.97	0.98
DEM-REP	0.86	0.15	0.86	0.86	0.86	0.90
FCB-RMCF	0.90	0.11	0.90	0.90	0.90	0.90

Table 2: Results for friendship classification

PA-IL	DEM-REP	FCB-RMCF
U_b is followed by U_a	U_b is followed by U_a	U_a's formal RT of U_b
U_a's engagement	U_b is following U_a	U_a's informal RT of U_b
U_a's informal RT of U_b	U_a's informal RT of U_b	U_b's # of followees
U_b's # of followees	U_b's # of followees	U_b's informal RT of U_a
U_b's total # of tweets	U_a's formal RT of U_b	U_b's % of total url
U_b's engagement	U_b's % of mention	U_b's total # of tweets
U_b's % of informal RT	U_b's total # of tweets	U_b's % of non-RT url
U_b's aggressiveness	U_b's % of non-RT non-url	U_b's # of followers
U_b's % of mention	U_b's # of followers	U_b's % of RT url
U_b's # of followers	U_b's % of total url	U_b's % of non-RT non-url

Table 3: Top 10 features selected

5. CONCLUSION

We proposed a language-agnostic model that based on features from four different aspects, enables classification of ideological relations between a pair of users. Our work differs from previous work on prediction of dyadic relations, as our method is language and domain agnostic and can be generalized to help identify any conflicts anywhere in the world. We believe our study is a first step in implementing tools to monitor and control the potential threats that occur due to large online conflicts.

6. REFERENCES

[1] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, pages 675–684, 2011.

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *ICWSM*, 2011.

[3] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and D. Boyd. The revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *IJOC*, 5:1375–1405, 2011.

[4] D. MacMillan. Twitter aids rise of web-based hate forums. **http://www.bloomberg.com/news/2013-05-07/twitter-aids-rise-of-web-based-hate-forums-report-finds.html**, May 2013. Accessed: 2013.

[5] I. Weber, V. R. K. Garimella, and A. Batayneh. Secular vs. islamist polarization in egypt on twitter. In *ASONAM*, pages 290–297, 2013.