


Extracting Graph Topological Information and Users' Opinion

Mirko Lai^{1,2}, Marcella Tambuscio¹, Viviana Patti¹, Giancarlo Ruffo¹, and Paolo Rosso²

¹ Dipartimento di Informatica, Università degli Studi di Torino, C.so Svizzera 185, 10149, Torino, Italy
{lai,patti,tambusci,ruffo}@di.unito.it,
² PRHLT Research Center, Universitat Politècnica de València, Camino de Vera, s/n 46022, Valencia, Spain
proso@dsic.upv.es

Abstract. This paper focuses on the role of social relations within social media in the formation of public opinion. We propose to combine the detection of the users' stance towards BREXIT, carried out by content analysis of Twitter messages, and the exploration of their social relations, by relying on social network analysis. The analysis of a novel Twitter corpus on the BREXIT debate, developed for our purposes, shows that like-minded individuals (sharing the same opinion towards the specific issue) are likely belonging to the same social network community. Moreover, opinion driven homophily is exhibited among neighbours. Interestingly, users' stance shows diachronic evolution.

Keywords: Stance Detection, BREXIT, Community Detection

1 Introduction

The political public debate is radically changed after the increasing usage of social media in last years. Politicians use them in order to conduct their political campaigns, and to engage users. On the other hand, users interact each other sharing their opinions and beliefs about political agenda or public administration. In this domain, techniques to study and analyse social media users' activity have been gaining importance in recent years, and (now more than ever) automatic approaches are needed in order to deal with this enormous amount of users' generated content. For instance, interest is growing in opinion mining, considered an important task to classify and monitor users' sentiment polarity [8], and in Stance Detection (SD), a finer grained task where the focus is on detecting the orientation *pro* or *con* that users assume within debates towards specific target entity, e.g., a controversial issue [7]. SD could be very useful to probe the citizens' perspective towards particular national and international political issues. Many recent works also suggest the exploitation of users' social community to develop features helping to detect their opinions [9, 2]. To learn more about the role of social relations in the formation of public opinion we

address two research questions: first, if individuals that share the same opinion towards a specific issue are likely to belong to the same community [6]; second, if link formation can be better understood in term of homophily (i.e., users with the same opinion are more likely to be connected to each other). We also explore the possibility to have a diachronic evolution in stance, e.g., people changing their stance after some particular events, happening when the debate is still active [3]. Here, we analysed the political discussion in United Kingdom (UK) about the European Union membership referendum, held on June 23rd 2016, commonly known as BREXIT, on Twitter. We showed that our hypotheses are supported by the analysis of real data proposing a new SD annotation scheme that takes into account temporal evolution, and a method for SD based on SVM in order to label the stance of users involved in the discussion.

2 Dataset

Data collection. In order to explore social relations and temporal evolution of users' stance, we collected about 5M of English tweets containing the hashtag #brexit using the Twitter Stream API, during the time span between June 22nd and 30th. First, we grouped tweets according to three time intervals, corresponding to relevant clear-cut events related to the referendum, in a short and highly focused time window:

- “*Referendum Day*” - the 24 hours preceding the polling stations closing (between June 22nd at 10:00 p.m. and June 23rd at 10:00 p.m.);
- “*Outcome Day*” - the 24 hours following the formalisation of referendum outcome (between June 24nd at 8:00 a.m. and June 25nd at 8:00 a.m.);
- “*After Pound Falls*” - the 24 hours after the financial markets' turbulence that followed the referendum (between June 28nd at 12:00 p.m. and June 29nd at 12:00 p.m.).

Then, we selected a random sample of 600 users from 5,148 that wrote at least 3 tweets in each time interval. We defined a *triplet* as a collection of three random tweets written by the same user in a given time interval. Finally, we created the TW-BREXIT corpus that consists of 1,800 triplets.

Manual annotation. We employed CrowdFlower³ to annotate the so-obtained corpus. We asked the human contributors to annotate the user's stance on the target *BREXIT* (i.e. UK exit from EU). In particular, given a triplet posted by an user, they had to infer the user's stance, by choosing between three options:

- *Leave*: if they think that the user is *in favour* of the UK exit from EU;
- *Remain*: if they think that the user supports staying within the EU (i.e. the user is *against* BREXIT);
- *None*: if they could not infer user's stance on BREXIT (e.g., all the messages are unintelligible, or the user do not express any opinion about the target, or the user expresses opinion about the target, but the stance is unclear).

³ <http://www.crowdfunder.com>

The final TW-BREXIT corpus contains 1,760 labelled triplets in agreement (majority voting)⁴.

Social Network. By the *friends/list* Twitter API, we collected the follower list for the 4,548 available⁵ users over 5,148 that wrote at least 3 tweets in each interval in order to explore users’ social network. We obtained a graph where a node represents a user and an edge between two users will exist if one follows the other. The graph consists in 4,114,523 nodes connected by 13,189,524 edges. We then extracted a sub-graph consisting in 198,419 nodes connected by 6,604,298 edges after removing friends having less than 10 relations in order to reduce computational issues.

3 Content and Network Analysis

Diachronic evolution of stance. In order to provide insights on temporal evolution, we analysed the label distribution in TW-BREXIT over the three temporal intervals. Not surprisingly, we observe an unbalanced distribution for stance as shown in Table 1. We used the hashtag #brexit for collecting data: despite it is apparently a neutral hashtag, a recent study [4] shows that most of tweets containing #brexit were posted by people that expressed stance in favour of Brexit, but since we are not interested in predicting the referendum outcome this bias is not crucial for the next analysis. It is more important to notice that label distribution changes over the time, in particular between “Outcome Day” and “After Pound Falls” phases. Then, we considered the point of view of a single user exploring if her/his own stance changes over time. We found that 57,66% of the users was labelled with the same stance in all the three temporal intervals (37,16% Leave, 15,5% None, 5% Remain). Very interestingly, 42,33% of users’ labelled stance changes across different temporal intervals. In particular, 9,5% of users’ stance varies from *Leave* (L) to *None* (N) (7% L → L → N; 2,5% L → N → N). From these results we cannot infer that users effectively changed opinion, but for sure they express their stance in their tweets in a different way depending on the phase of the political discussion. This is an argument in favour of the hypothesis that stance should be analysed not in isolation but also in a diachronic perspective, which will be matter of future deeper investigations.

Table 1. Label distribution over the time

Time span	Leave	Remain	None
Average	961 (51%)	236 (14%)	563 (35%)
Referendum day	55.67%	13.67%	30.67%
Outcome day	55.67%	14%	30.33%
After Pound falls	50%	13.67%	36.33%

Automatic content analysis: stance detection. We aim to automatically estimate the stance of all users of our dataset in order to explore how the

⁴ Inter-Annotator Agreement: 65.48. The corpus is available for research purposes.

⁵ Some users set privacy in order to hide profile information, while others shut down their profile after the referendum.

stance is distributed in the social network. Then, we propose a machine learning supervised approach using SVM to annotate the stance s of the remaining 3,948 users, using the following five features computed over a triplet: bag of words (BoW), structural-based (structural), sentiment-based (sentiment) (described in [5]), community-based (community), and temporal-based (temporal). The community feature returns the community of the user who wrote the triple, while the temporal one, the given time interval of the triplet. The F-Measure $\frac{F_{leave} + F_{remain}}{2}$ obtained by SVM using all the mentioned features is 67% and it overcomes the performance of SVM trained with unigrams (58.25%) and unigrams plus n-grams (60.14%) (baselines proposed by [7]).

Community Detection. Subsequently, we analysed the network topology. Figure 1(a) shows the graph plotted by the software Gephi⁶ coloured by user’s community. The users community’s membership was assigned by the Louvain Modularity method [1]. Figure 1(b) shows the graph where users have been coloured according to the annotated stance computed with SVM. Table 2 highlights that the percentage of users’ stance in community D is evidently biased towards the stance “Remain”; in communities B, E, and F towards the stance “Leave”; in communities A and C towards the stance “None”. The existences of communities so defined in term of stance could allow filter bubble phenomena to occur.

Neighbourhood Overlap. Lastly, we evaluated the stance similarity among couples of connected nodes. Then, we defined users *agreement* as a measure of the likelihood that two users i and j have the same stance (i.e., $s(i) = s(j)$) in the same time interval, and then we explored how the agreement between two users changes depending on the rate of the common neighbours. *Neighbourhood Overlap* (NO) is defined as the number of neighbours that nodes i and j have in common divided by the sum of neighbours of both i and j (not counting i and j themselves):

$$NO(i, j) = \left(\frac{|N_i \cap N_j|}{|N_i \cup N_j| - \{i, j\}} \right) \quad (1)$$

where N_i and N_j are the sets of neighbours of nodes i and j respectively. Table 3 shows how to compute the agreement score $A_{i,j}$ between i and j . Considering $E_l = \{(i, j) \in E \mid NO(i, j) = l\}$ as the subset of edges that are incidents to neighbours of both i and j , when $NO(i, j)$ value is exactly equal to l , we computed the *agreement* A_l related to the NO level l as it follows:

$$A_l(i, j) = \sum_{i, j \mid (i, j) \in E_l} \frac{A_{i,j}}{|E_l|}. \quad (2)$$

Roughly speaking, through this measure we want to explore if the opinion agreement among users changes accordingly the rate of common neighbourhood. We computed users’ agreement in our dataset for each time interval and then we took the averaged value for different values of neighbourhood overlap. Figure 2 shows that the agreement between two users increases depending on the percentage of friends. Results showed the tendency of users to associate with similar others according to opinion driven homophily.

⁶ <http://gephi.org>

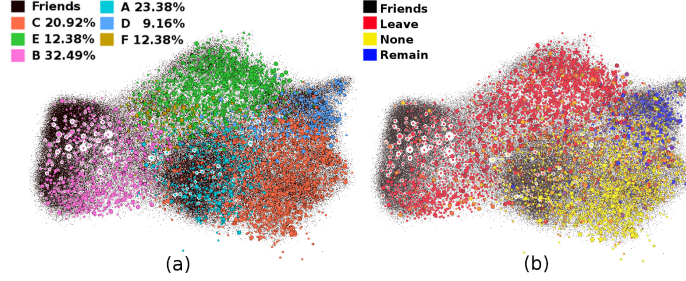


Fig. 1. In (a), each node is coloured depending on assigned community. Otherwise, in (b), they are coloured according to the annotated stance of the user’s triplet by SVM (red for *Leave*, yellow for *None*, blue for *Remain*, mixed colours when stance changes over time). Followers and the remaining users are black-coloured.

Table 2. Users’ stance distribution over communities. The percentage shows the average users’ distribution in communities over the three temporal phases.

Community	A	B	C	D	E	F
Leave	29.63%	84.61%	26.31%	18.96%	85.6%	75%
Remain	11.11%	0.37%	17.02%	57.47%	2.37%	0%
None	56.79%	14.28%	54.38%	18.39%	10.06%	22.92%

Table 3. The table shows the Agreement score for couple of users (i, j) over the temporal phases. The maximum value is 1 in the case i and j agree ($s(i) = s(j)$) in all the three temporal phases, 0 if one or both users have label “None” and -1 otherwise.

	Agreement	One or both None	Disagreement
Referendum day	0.33	0	-0.33
Outcome day	0.33	0	-0.33
After Pound falls	0.33	0	-0.33
	1	0	-1

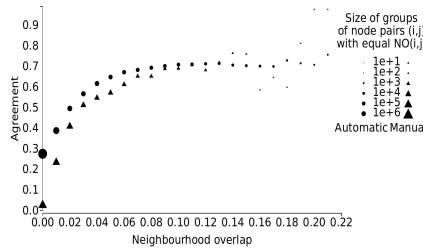


Fig. 2. A shape (circle or triangle) represents a group of node pairs (i, j) with equal $NO(i, j)$ (rounded to two decimal points). Shape size is proportional to the size of such groups. The agreement score $A_{i, j}$ was computed with manual annotation stance (triangle) and with user’s stance computed by SVM (circle). We noted that the *affinity* among two users increases depending on the rate of NO.

4 Discussion

In this paper we have shown that users having the same stance towards a particular issue tend to belong to the same social network community. Moreover, we found evidences that the neighbours are more likely to have similar opinions. The obtained results show that stance verified by human annotators over the same user varies over time, even though we exclusively focused on three 24-hours time slots in a time span of only 8 days. This suggests that stance should be studied considering the diachronic evolution of the debate. We are planning to combine the diachronic evolution of users' stance with the dynamic social network perspective and to explore this methodology on other political corpora. In our future research we would also like to understand the role that influencers could have on the stance change. Moreover, we would like to investigate the use of irony within polarised communities in order to figure out if social network relations influence the use of this figurative language.

Acknowledgments

The work of the last author has been partially funded by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) under the research project SomEMBED TIN2015-71147-C2-1-P and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10 (2008)
2. Deitrick, W., Hu, W.: Mutually enhancing community detection and sentiment analysis on Twitter networks. *Journal of Data Analysis and Information Processing* 1, 19–29 (2013)
3. Gelman, A., King, G.: Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* 23(04), 409–451 (1993)
4. Howard, P.N., Kollanyi, B.: Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. *ArXiv e-prints* (2016)
5. Lai, M., Hernandez Farias, D.I., Patti, V., Rosso, P.: Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets. In: Sidorov, G., Herrera-Alcántara, O. (eds.) Part I. *Lecture Notes in Artificial Intelligence. Advances in Computational Intelligence. 15th Mexican International Conference on Artificial Intelligence, MICAI 2016*. vol. 10061, pp. 152–165 (october 2016)
6. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. *Annual review of sociology* pp. 415–444 (2001)
7. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. pp. 31–41. ACL, San Diego, US, CA (2016)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2(1-2), 1–135 (2008)

9. Xu, K., Li, J., Liao, S.S.: Sentiment community detection in social networks. In: Proceedings of the 2011 iConference. pp. 804-805. iConference '11, ACM, New York, US, NY (2011)