Interactions of Language and Multilingual Communities in Twitter

Nabeel Albishry

Department of Computer Science

University of Bristol

Bristol, UK

Email: n.albishry@bristol.ac.uk

Theo Tryfonas

Department of Civil Engineering

University of Bristol

Bristol, UK

Email: Theo.Tryfonas@bristol.ac.uk

Tom Crick

Department of Computing

Cardiff Metropolitan University

Cardiff, UK

Email: tcrick@cardiffmet.ac.uk

Abstract - Whilst emerging research is providing insight into the factors that promote the propagation of information in online social networks following significant events, such as big social events. This paper evaluates the extent to which different language communities engage and interact. We present our analysis of online interactions in various languages that took place on the social networking site Twitter during the international Eurovision Song Contest in May 2016.

By utilising language information from user profiles (N= 1,226,959) and status updates (N=7,926,746) relating to the Contest to identify and categorise communities, we are able to provide insight into the pattern of their interactions, as well as shed some light on multilingual community. The results show that the nature of the event is reflected on the engagement degree and wider interaction of communities. It also shows the participation pattern of multilingual users.

Keywords – predictive, social media marketing, twitter communities, languages

# Introduction

# CONTEXT AND EVENTS TIMELINE

From the contest official website [14], it is a yearly television show that started back in 1956. This year’s contest took place in Stockholm, Sweden, in May 2016. There were 32 countries, each entred with one team. There were two semi-finals, on 12th and 14th May, 18 teams each. Then, 26 qualify for the final on 16th May. This year’s contest was very tough and tense, especially when the Ukrainians won the final. The result was not far from being politicised as we witnessed afterward [16]. Although some articles do not see the contest as influenced by political conflicts or friendship[17][18], some recent news articles articulate on the possibly biased results[19]. The activity on Twitter was also very high throughout the event using #Eurovision hashtag as the main tag. The size of participation collected exceeded 7,900,000 statuses. The activity was produced from 1,226,959 users. Figure 1 below, shows the overall activity on Twitter.

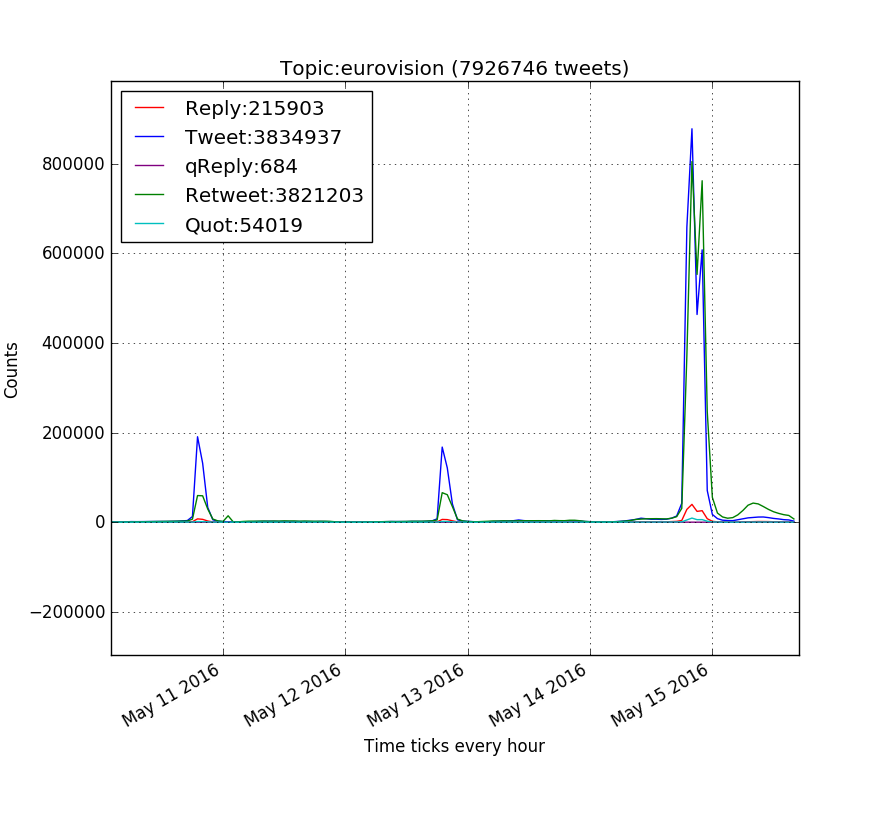


Figure . Overall activity for #Eurovision

# Language Communities

The preliminary analysis shows that tweets and retweets together account for ~97% from the total activity, as we can see in Figure 1. Therefore, we think these two subsets can be representative by their own, without the need to include other sets, such as replays and quotes. It is important to note that tweets and retweets are used to measure actions, and reactions, respectively. However, our analysis will be focusing on original tweets only and usage of different languages in this set. Analysis of language communities begins with two basic techniques. The first is to classify statuses based on their languages. The status language is extracted from the ‘lang’ entity inside status objects. Language used in posting defines which community the status was meant for; a tweet written in Turkish, for example, is meant for the Turkish-speaking community. Output from this will be referred to as ‘posting communities’. The second analysis is to classify users into different communities based on their profile languages, regardless of the posting language they used. Output from this technique will be referred to as ‘profile communities’. As we will see in the following sections, a posting community does not necessarily indicate the profile community for a user.

## Profile and Posting Communities

In the #Eurovision case, there were 49 posting languages. Table 1 shows the top posting languages accounted for 90% of original posts (tweets), out of 3,834,937. As might be expected, the English was the most used posting language. Interestingly, results also show that language of more than 142,721 (3.72%) statuses could not be identified. When investigated, about 40% of those statuses do not contain much text other than hashtags, mentions or URLs. Although this category shows an interesting case in which qualitative content analysis could be involved, it is beyond this study and will not be covered here.

|  |  |
| --- | --- |
| Language | % |
| en | 45.90 |
| es | 17.24 |
| ru | 8.99 |
| fr | 6.20 |
| und | 3.72 |
| nl | 3.71 |
| de | 3.19 |
| it | 2.85 |

Table : Posting languages accounted for 90% of original activities (tweets)

In total, we there were 1,226,959 users participated in the #Eurovision topic. In terms of their profile languages, they form 50 communities. Table 2 shows the profile communities that accumulate 90% of all users. Unlike status language, profile language relies on the user to pick a language for their Twitter profile settings. In general, the default value of this option is the initial placeholder text “Select Language...” or a translated version that might provide hints regarding the user language community. In our dataset, we found that all users chose a language and no users left it to the default value.

|  |  |
| --- | --- |
| Community | % |
| en | 47.06 |
| es | 20.37 |
| fr | 8.00 |
| ru | 7.07 |
| de | 3.53 |
| nl | 3.31 |
| it | 2.25 |

Table : Profile communities (top 90% of the total users)

## Profile-Posting Analysis

From the two tables above, we can see some consistency in the posting and profile communities. Taking an exceptional case as an example, we can see that although French profile community had more presence, Russian posting community is higher by 2.79%. Simple explanation would be that Russian profile community was relatively more active than French. Another reason could be the participation of non-Russian profiles using Russian language for posting. To investigate this, we took a closer look at contribution of profile communities to the Russian posting community. The result in Table 3 shows profile communities that resulted in more than 95% of activity in this posting community.

|  |  |
| --- | --- |
| Community | % |
| ru | 91.25 |
| en | 7.26 |

Table : Active profile communities within Russian posting community

As we saw in this example, posts in Russain were not merely from Russian profile community. This simple illustrative example show one way of exploring relationships between profile and posting communities, especially if we are interested in particular communities.

Another perspective is to explore the posting behavior of one particular community. When considering some profile communities, there is a tendency to assume that communities post in languages that are same as their profile language. To examine this assumption, we investigated participation of ‘en’ profiles, as it forms nearly 50% of users. In total, there were 1,841,205 posts from this community, 81% of which were posted in ‘en’, 15.4% in other languages, and 3.62% were not identified. Table 4 below lists the top 95% posting languages used by this profile community.

|  |  |
| --- | --- |
| Posting Lang | % |
| en | 80.99 |
| und | 3.62 |
| es | 2.69 |
| nl | 2.39 |
| fr | 1.39 |
| ru | 1.36 |
| de | 0.97 |
| it | 0.87 |
| el | 0.86 |

Table : Top 95% participation languages from 'en' profiles

## Language Diversity

By observing language diversity of profile communities we aim to measure language diversity of the topic in general, as well as investigating which community plays more in bridging different profile communities. Diversity here means how many posting languages were used from each profile community, and to what extent they used their own language, as well as other languages.

The general language diversity of the topic is about 17%, while 3.72% were not identified. All of the 50 profile communities used different language in posting. Interestingly, 16 out of those communities did not use their own language, they were low in participation though. Moreover, in terms of using different languages, we found that 32 communities scored at least 50% out of their tweets. We noticed that posting from small profile communities may affect the overall language diversity of the topic. Referring to the top profile communities mentioned in section III.A, Table 5 below shows their diversity in percentage. Russain profile community shows very interesting case as it scored the least diverse profile amongst all the 50 communities although it comes fourth in number of users.

|  |  |
| --- | --- |
| Community | Diversity (%) |
| de | 34.27 |
| nl | 32.78 |
| it | 18.49 |
| fr | 16.65 |
| en | 15.39 |
| es | 10.13 |
| ru | 7.93 |

Table : Diversity of top profile communities

## Multilingual communities

In this section, we group users based on their relationship with posting communities, regardless of their profile language. For example, a user posted in ‘en’ and ‘fr’ will be classified as bilingual, and so on. Based on this grouping technique, with the ‘und’ lang eliminated, we identified 20 sets. The smallest two groups consist of one user each, who posted in 22 and 25 different languages.

As we can see in Figure 2, monolingual users scored about 85% of all users causing 47% of the total original posts. The figure also shows that users and their activity decrease as number of languages used increase.

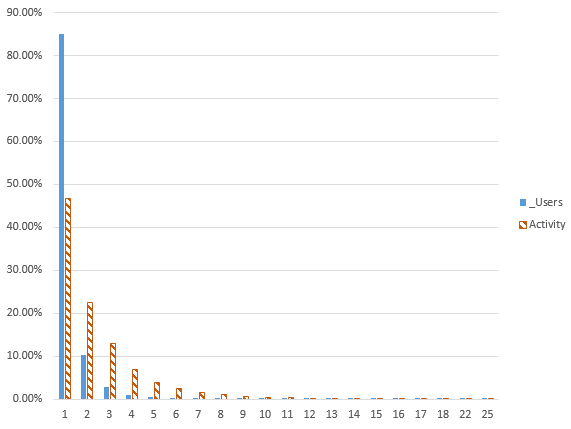


Figure : Multilingual communities and their activity

A closer look at the behavior of these communities shows that, in general, activity per user increases as number of used languages increase, as shown in Figure 3 below.

Although we cannot conclude that there is a correlation between high multilingualism and illegitimacy of accounts, this would be an interesting further research to consider.

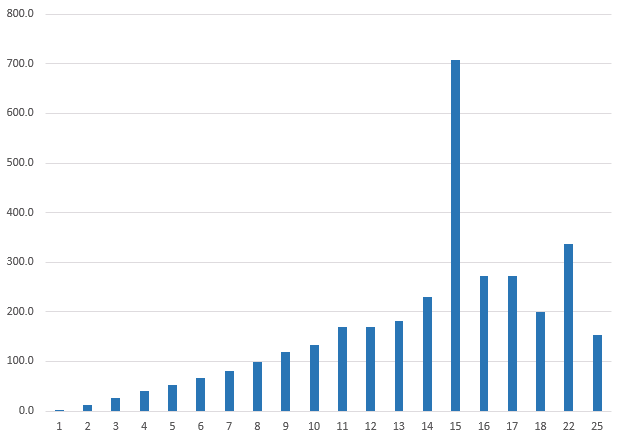


Figure : Average of posts per user for multilingual communities

# Conclusion

This paper presented a study in identifying languages used, language and multilingual communities, and their engagement and interactions on the Twitter platform with respect to real world events, in this instance using the Eurovision Song Contest happened in early May 2016.

As we discussed in section III, there is a positive relationship between size of profile and posting communities. We also showed that a big number in participating profile community does not necessarily imply high language diversity, and that diversity may results from small profile community.

We also presented structure of multilingual communities and their activity. Although most users may use their own profile language in posting, most of the activity came from multilingual users. In few cases, users may use a questionable number of languages, up to 25 different languages. These extreme cases may be very interesting to investigate for possible spammers detection or for sociolinguistics in moderate cases.

The method we presented here can be used in identifying how communities interact with one another, which ones are most active, which languages are mostly used, and at what time. Moreover, within certain contexts, the order of applying these two classifications (posting and profile) will generate results in different perspectives. For example, taking one profile community and dividing it into different posting communities shows the number of languages this community may use, and hence degree of openness and reachability. A possible scenario for governments, politicians or campaigners would be to use this method to measure to what extent other languages are used within a profile community. It may also show how users associate themselves with one community in their profile while using other languages. Monitoring unusual activity for secondary languages, in multilingual communities, may help to uncover important messages or opinions that could not be openly expressed, for a variety of reasons, to the rest of the profile community.

For future work, we plan to have a deeper look at how multilingual communities participate and their reactions network. We believe that differentiation between endorsements (e.g. retweets) and other reactions may provide further insight into the networks and communities. Furthermore, we will apply the methods presented in this paper on other high-profile event/discussion datasets in different domains or contexts, such as for sports and humanitarian actions.

##### Acknowledgment

##### References

1. [1] 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,” *Int. J. Commun.*, vol. 5, p. 31, 2011.
2. [2] P. N. Howard, A. Duffy, D. Freelon, M. Hussain, W. Mari, and M. Mazaid, “Opening Closed Regimes. What Was the Role of Social Media During the Arab Spring?,” 2011.
3. [3] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting Elections with Twitter,” in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 178–185.
4. [4] M. S. Gerber, “Predicting crime using Twitter and kernel density estimation,” *Decis. Support Syst.*, vol. 61, pp. 115–125, 2014.
5. [5] X. Zhang, H. Fuehres, and P. A. Gloor, “Predicting Stock Market Indicators Through Twitter ‘I hope it is not as bad as I fear,’” *Procedia - Social and Behavioral Sciences*, vol. 26. pp. 55–62, 2011.
6. [6] A. Willis, A. Fisher, and I. Lvov, “Mapping networks of influence : Tracking Twitter conversations through time and space,” *J. Audience Recept. Stud.*, vol. 12, no. 1, pp. 494–530, 2015.
7. [7] G. Oatley and T. Crick, “Measuring UK crime gangs,” *ASONAM 2014 - Proc. 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, no. Asonam, pp. 253–256, 2014.
8. [8] R. Kang, S. Brown, and S. Kiesler, “Why do people seek anonymity on the internet?,” *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI ’13*, p. 2657, 2013.
9. [9] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, “Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter's firehose,” *Proc. ICWSM*, pp. 400–408, 2013.
10. [10] S. Kumar, F. Morstatter, and H. Liu, “Twitter Data Analytics,” p. 89, 2013.
11. [11] L. Tan, S. Ponnam, P. Gillham, B. Edwards, and E. Johnson, “Analyzing the impact of social media on social movements: a computational study on Twitter and the Occupy Wall Street movement,” in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, 2013, pp. 1259–1266.
12. [12] Twitter, “Twitter API Documentation.” .
13. [13] M. Graham, S. Hale, and Gaffney，Devin, “Where in the world are you ? Geolocation and language identification in Twitter,” *Prof. Geogr.*, vol. 00, no. 2013, p. 00, 2014.
14. [14] “Eurovision Song Contest website.” [Online]. Available: www.eurovision.tv. [Accessed: 25-May-2016].
15. [15] J. Berlinger, “Baltimore riots: A timeline,” *CNN*, 2015. [Online]. Available: http://edition.cnn.com/2015/04/27/us/baltimore-riots-timeline/index.html. [Accessed: 20-Sep-2015].
16. [16] A. Withnall, “Russian officials threaten to boycott next Eurovision after victory for ‘political’ Ukraine entry,” *The Independent*, 15-May-2016.
17. [17] V. Ginsburgh and A. G. Noury, “The Eurovision Song Contest. Is voting political or cultural?,” *Eur. J. Polit. Econ.*, vol. 24, no. 1, pp. 41–52, 2008.
18. [18] M. Blangiardo and G. Baio, “Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models,” *J. Appl. Stat.*, no. January 2015, pp. 1–11, 2014.
19. [19] A. Kirk, J. Kempster, and S. Franco, “Eurovision 2016: How does country bias affect the result?,” *The Telegraph*, 23-May-2016.
20. [20] M. E. J. Newman, “The structure and function of complex networks.,” *Dialogues Clin. Neurosci.*, vol. 45, pp. 167–256, 2003.