“*Come Together!*”: Interactions of Language Networks and Multilingual Communities on Twitter

Nabeel Albishry
Department of Computer Science
University of Bristol
Bristol, UK
n.albishry@bristol.ac.uk

Theo Tryfonas
Department of Computer Science
University of Bristol
Bristol, UK
theo.tryfonas@bristol.ac.uk

Tom Crick
Department of Computing
Cardiff Metropolitan University
Cardiff, UK
tcrick@cardiffmet.ac.uk

ABSTRACT

Emerging research is providing insight into the factors that promote the propagation of information in online social networks following significant activities, such as high-profile international social or societal events; this paper provides insight into how people are linked, by how different language communities engage and interact. We present our analysis of two significant online interactions in various languages that took place on the social networking site Twitter: during the Baltimore protests in April 2015 in the USA and the Eurovision Song Contest in May 2016.

By utilising language information from user profiles (Baltimore: N=716,494; Eurovision: N=1,226,959) and status updates (Baltimore: N=1,257,065; Eurovision: N=7,926,746) to identify and categorise communities, we are able to provide insight into the pattern of their interactions, as well as constructing their network graphs to shed light on these multilingual community. The results show that the nature of the event is reflected on the engagement degree and wider interaction of communities, as well as showing the participation pattern of multilingual users. This analysis of language communities may also help in deciding which group of users to engage with, and hence increase the chance of influential action when participating on Twitter conversations.

KEYWORDS

TBC

ACM Reference format:
Nabeel Albishry, Theo Tryfonas, and Tom Crick. 2017. “*Come Together!*”: Interactions of Language Networks and Multilingual Communities on Twitter. In *Proceedings of ACM Conference on Hypertext and Social Media, Prague, Czech Republic, July 2017 (HYPERTEXT 2017),* 10 pages.
DOI:

N.B. The first part of the title of this paper comes from the motto of the 2016 Eurovision Song Contest, which along with the theme artwork was “inspired by the dandelion, symbolising the power of resistance and resilience but also of regeneration”.

Author’s addresses: N. Albishry & T. Tryfonas, Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK; T. Crick, Department of Computing, Cardiff Metropolitan University, Cardiff CF5 2YB, UK.

HYPERTEXT 2017, Prague, Czech Republic

DOI:

1 INTRODUCTION

1.1 Online Social Networks

In recent years, online social networks (OSNs) have been utilised as means to express ideas and opinions, spread information about events, or even stimulate and propagate calls for civic engagement and societal action. Social networking sites such as Twitter, Facebook, LinkedIn and YouTube have also empowered individuals to promote their viewpoints and interests – professional or otherwise – to a broad and diverse global audience. The engagement of certain demographics with social networks offers the opportunity for researchers interested in observing and interpreting society to apply established theory and methods to an emerging digital culture.

To satisfy the demand for various types of communities, interactions and engagement, there are now vast numbers of social media sites and platforms[1], along with a number of attempted categorisations. By 2018, there will be an estimated 2.5 billion active social network users (up from 1.9 billion in 2014); they are producing massive amounts of data (volume) on a real-time basis (velocity) with implicit sociological attributes such as beliefs, opinions, sentiments, behaviours, structures and influences (variety) [8]. These data exhibit the key traits of what is now referred to as big data: volume, velocity and variety [44]. In this age of big data and an increasingly interconnected digital society, there is a new challenge – the application of robust and scalable methods and tools that can be applied to digitised social behaviour generated via social networks so as to be able to efficiently analyse big social data to provide insight into real-world events and actions [8, 28].

Recent work [1, 2, 33, 34, 38] has analysed what people say on social media to identify distinctive words, phrases, and topics as functions of known attributes of people such as gender, age, location, or psychological characteristics. This can thus be collated and aggregated, inferring gender, age, location and sentiments, from social media data. Potential negative implications of these approaches include the fact that they can be easily applied to large numbers of people or groups in society without obtaining their explicit consent or even being aware it is being done. Data-driven commercial companies, governmental entities, or even one’s followers or friends are able to use software to infer personality and other attributes – such as sexual orientation or political affiliations – that an individual may have decided not to share [27, 44].

There are various projects that have used Twitter corpora and related datasets to make predictions about elections [43], stock markets [50], and crimes and policing [16, 35]. Twitter played an

[1]This list is by no means exhaustive: http://en.wikipedia.org/wiki/Listofsocialnetworkingwebsites

important role during what was then known as the “Arab Spring”, which has been extensively examined in the social network analysis domain [6, 14, 22, 31, 47]. While the use of Twitter data has been demonstrated to provide insight – and sociologically relevant demographics [39] – into major social and physical events such as riots [36] and terror attacks [9], often all is not what it may seem; for instance, many tweets may not a crowd make [29].

1.2 Languages and Communities

Despite the widespread engagement with Twitter globally, little research has investigated the differences amongst users of various languages; there is a tendency to assume that the behaviours of English users generalise to other language users [21]. Language has featured as a facet of research on the geographies of Twitter networks [40], especially whether offline geography still matter in online social networks [25]. Linguistic-inspired studies have been done on hashtags [15], as well as the volume and proportional of tweets in English and Arabic, as part of an analysis of the Arab Spring [6]. Nevertheless, language is clearly a vital component of affiliation and discourse on the web [48], with the creation and curation of emerging multilingual networks and communities, representing well-established creative and cultural norms, including for minority languages such as Welsh [20], as well as investigations into the economics of linguistic diversity [18].

1.3 Social Network Analysis

In the social network analysis (SNA) domain, centrality measures provide the ability to assess network graphs that are constructed from collected data (for example, tweets). Selection of these centrality measures is dependent on the goal of the analysis; for example, the degree of node helps to identify nodes with high number of connections within the network [4, 30, 37]. In a representation of a real world network, this metric may help to identify highly connected persons, such as political leaders, sports stars or celebrities, who are potential “information spreaders” [5, 11, 49]. Centrality measures such as degrees, betweenness, clustering coefficient, modularity and cliques have been used in many projects to measure influence or detect the emergence of new communities [35, 46].

Clustering users in communities has been an important analytic factor in social networking analysis; numerous work has focused on clustering users based on their locations. However, for the sake of anonymity, many users tend not to disclose information about their identity, such as locations [23]. It has also been reported in the literature that geotagged tweets are generally low in number [26, 32, 41], the exponential growth in social media over the past decade has been joined by the rise of location as a central organising theme [29] of how users engage with online information services and, more importantly, with each other [10, 13].

1.4 Users and Location

It is important to understand how geotagging works in Twitter. The ‘place’ entity included in a Twitter status does not necessarily indicate precisely where the actual posting was made, as stated in the Twitter API documentation[2]:

> “*Tweets associated with places are not necessarily issued from that location but could also potentially be about that location.*

For the sake of anonymity many users tend not to disclose information about their identity, particularly locations; this has also been supported by the literature that geotagged tweets are generally low in number [23]. An alternative location-based option to consider is based on profile location, but this may not serve the need for location clustering for a multitude of reasons, especially with a significant proportion of Twitter users not setting their profile location [19].

1.5 Language Communities

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 two case studies, a posting community does not necessarily indicate the profile community for a user. Therefore, the second step is to examine the relationship between profile and posting communities. We will also explore relationships amongst profile communities, in term of action-reaction. We will investigate the intra- and inter-profile communities interactions by constructing network graphs and generate a visual representation.

1.6 Overview of Paper

The techniques we introduce in this paper through two real-world case studies are based on language settings in users’ profiles and those for statuses[3]. The remainder of this paper is organised as follows: Sections 2 and 3 presents the 2015 Baltimore protests and 2016 Eurovision Song Context case studies, along with an analysis of the key data and results. Section 4 concludes the paper with a wider discussion and a summary of the potential application of our approach.

2 CASE STUDY: 2015 BALTIMORE PROTESTS

Following the peaceful funeral of Freddie Gray that took place on the morning of Monday 27 April 2015 in Baltimore, Maryland, USA, a protest hit the city. According to the timeline published on the CNN website “*The city exploded on Monday after the funeral of Freddie Gray, a 25-year-old black man who mysteriously died on April 19, a week after Baltimore Police arrested him.*” [45]. The nature of the Baltimore protests is a good representation of a planned event in which a sudden escalation of violence hits a geographical area. The event manifested itself on Twitter as `#BaltimoreRiots`, and resulted in more than 1,250,000 status updates.

Figure 1 and Figure 2 present how the event manifested itself on Twitter once a “purge” was scheduled. We can see that what

[2] https://dev.twitter.com/overview/api/places

[3] The term ‘status’ is a generic term used to refer to any Twitter post (tweet, retweet, reply, or quote).

was happening on the ground was quickly reflected on the activity in Twitter, as Figure 2 indicates. More detailed analysis reveals that within one hour the topic started to go "viral"; more precisely, at approximately 15:00 at which the "purge" was scheduled. The topic jumped from roughly 1,200 to 8,000 tweets per hour. Then, it peaked with 98,000 between 22:00 and 23:00.

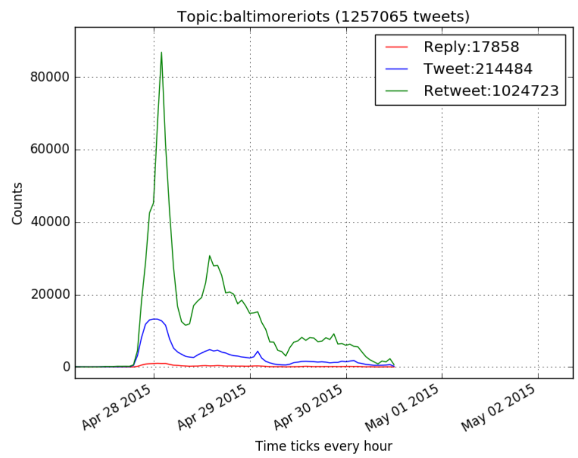

Figure 1: Overall activity for `#BaltimoreRiots`.

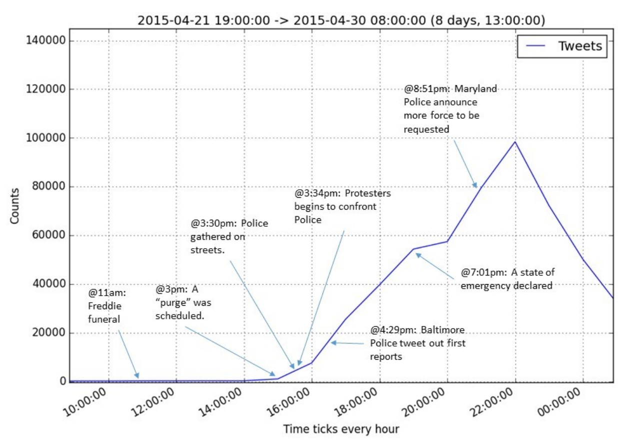

Figure 2: Main events during the lifetime of `#BaltimoreRiots`.

2.1 Locations

As mentioned in Section 1, for the sake of anonymity many users tend not to disclose information about their identity, particularly locations; this has also been supported by the literature that geotagged tweets are generally low in number [23]. We took the step to verify this claim in our datasets; in the best cases, the ratio of geotagged tweets did not exceed 2%. In the case of the `#BaltimoreRiots` dataset, only 1% of collected statuses were associated with places. Moreover, out of this geotagged subset, only 4% were associated with the city where the event took place (Baltimore).

An alternative location-based option to consider is based on profile location, but it still does not serve the need for location clustering for a multitude of reasons. Firstly, we found that less than 45% of users have set their profile location, which is in line with other studies [19]. Secondly, although Twitter suggests certain presets for setting profile location, users are given the option to enter any text they wish; this results in a considerable amount of noise.

2.2 Posting Communities

In the `#BaltimoreRiots` case, for original posts () there were 39 posting languages, including und. As we can see in Figure 3, English was the dominating language by far. Interestingly, results also show that language of more than 7% statuses could not be identified. When investigated, those statuses mostly do not contain text other than hashtags, pictures or URLs. Although, this is not a big portion, it came second after English. Although this category shows an interesting case in which qualitative content analysis would be involved, it is beyond this study and will not be covered here.

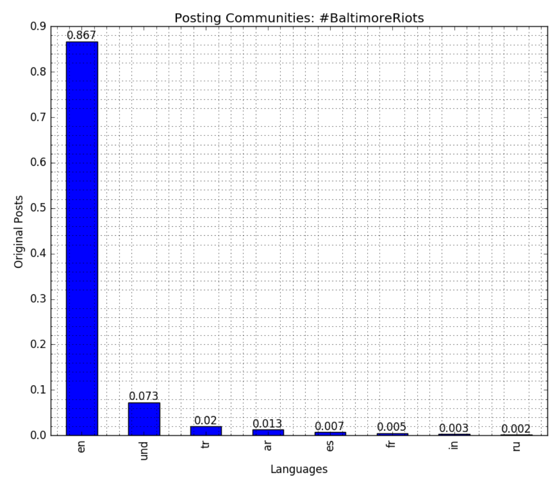

Figure 3: Most frequently used languages in `#BaltimoreRiots`: (*en:* English; *es:* Spanish; *tr:* Turkish; *fr:* French; *en-gb:* British English; *ar:* Arabic; *de:* German; *ru:* Russian; *it:* Italian; *pt:* Portuguese)

2.3 Users' Language Communities

In the majority of cases, users choose to pick a language for their Twitter profile settings. In our dataset we found that out of 716,494 users, only 45 had not chosen any language. However, the language entity returned by the API for those cases is the initial placeholder text "*Select Language...*" or a translated version that might provide hints regarding the user language community. Figure 4 shows that about 94% of the users came from '*en*' profile community.

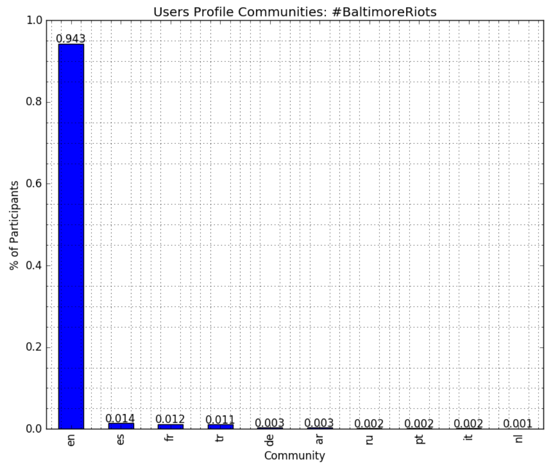

Figure 4: Top 10 profile language communities in `#BaltimoreRiots`

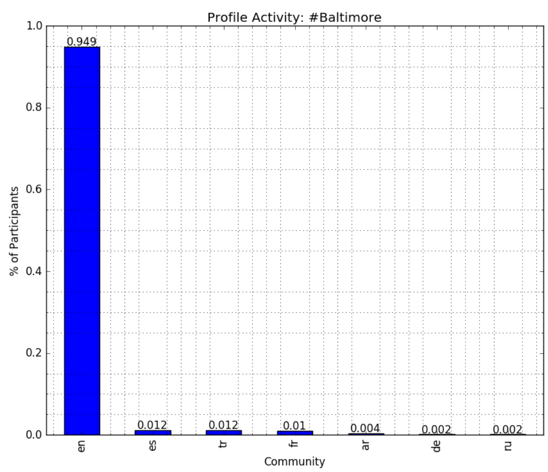

Figure 5: Profile language communities casuig %99 of activity in `#BaltimoreRiots`

As we can see in figure 5, activity from profile communities is not far from their sizes. Also, from these two outputs, we can see that nearly all of the topic activity came from one particular community using one particular language. This extreme pattern may accompany extreme and geographically constrained real world events such as riots and terror attacks.

2.4 Profile-Posting Graph

To investigate whether the '*en*' posting community is linked to particular profile communities, we constructed a bipartite graph as presented in figure 6, representing the profile-posting language network. In this graph, nodes that are prefixed by "*p_*" represent profile language community, and nodes that are prefixed by "*s_*" represent posting language community. Size of node represents the weighted indegree, whereas colour represents the outdegree; the darker the colour, the higher outdegree, hence, totally white nodes have no out degree and help to easly distinguish posting communities from profile ones. The graph confirms the domination pattern we highlighted earlier; furthermore, it shows the relationships between the profile and posting communities.

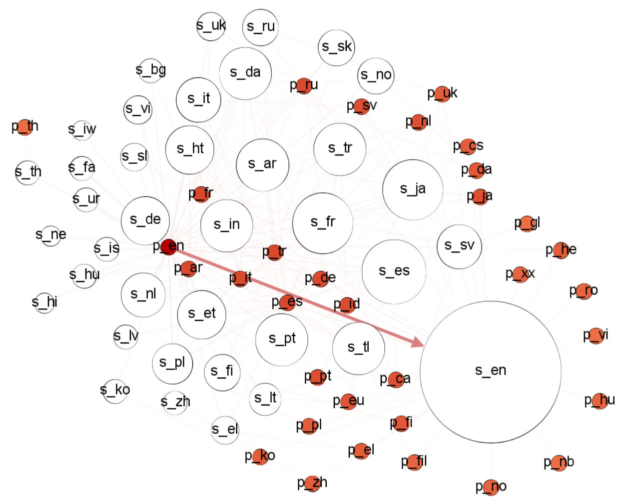

Figure 6: Profile-Posting network graph, selfloop included

Furthermore, to examine users behaviour in using languages other than their own (profile language), while same-language communities are filtered out for clarity, the result is shown in figure 7. In this figure, relationships can be easly observed by looking at weight of their edges. As we can see in Table 1, in this context, English profile users have mostly been posting in Arabic, followed by Espanish profiles posting in English. This observation shed the light on those relatioships, and could be of use for further analysis, such as highly dissiminated message that fall into these relationships and thier contents.

Profile-Posting Edge	Weight
en-ar	1791
es-en	727
ar-en	697
fr-en	571
tr-en	558
en-tr	550

Table 1: Users behaviour in using languages different to their profile

For an extreme case of one dominating posting language, we wanted to investigate participation of different communities. We

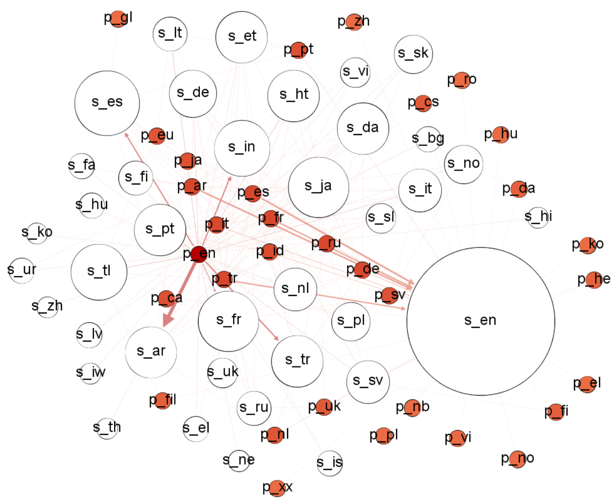

Figure 7: Profile-Posting network graph, no selfloop

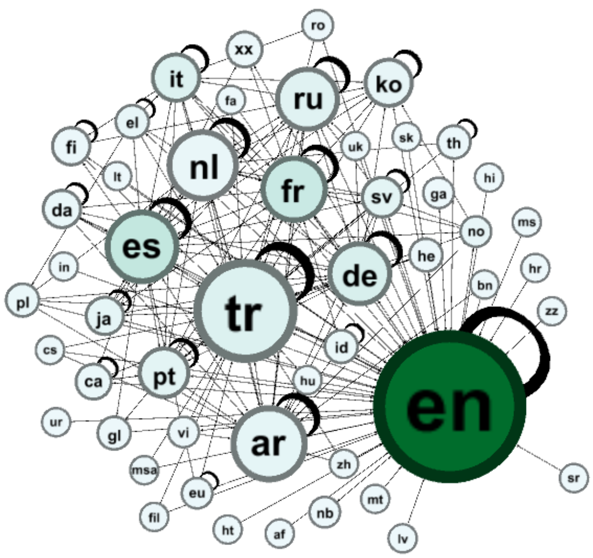

Figure 8: Profile-profile network graph

thus filtered out all non-'*en*' statuses, and then identified different profile communities with the resultant set. For each community, we classified statuses into two sets: *actions* and *reactions*; this result is shown in Table 2. This shows the highest scoring communities, where the first column represents the category of status (action or reaction), community column represents profile language community, and last one shows percentage of '*en*' posts by that community.

Category	Community	%
Reaction	*en*	81.08
Action	*en*	15.43
Reaction	*es*	0.68
Reaction	*fr*	0.59
Reaction	*en-gb*	0.47
Reaction	*tr*	0.19
Reaction	*de*	0.18
Reaction	*pt*	0.13

Table 2: Activity and categories of most active profile language communities

From the results above we can infer that there is a dominating player in both domains: posting languages and profile communities. Therefore, for the case of `#BaltimoreRiots`, we can conclude that the case was substantially localised.

2.5 Reaction Networks

Another important perspective to identify and capture is how different profile communities relate to each other in terms of reactions. Therefore, we constructed the graph in Figure 8 to show interactions amongst language communities. Edges are directed and are drawn from reacting nodes (retweeter, or replier) to the original acting node (tweeter). Node size represents weighted indegree; darkness of colour shows oudegree. The figure clearly shows that communities are most reactive to their own than to others.

3 CASE STUDY: 2016 EUROVISION SONG CONTEST

The Eurovision Song Contest (*Concours Eurovision de la chanson*) – sometimes popularly called Eurovision – is the longest-running annual international TV song competition, held, primarily, among the member countries of the European Broadcasting Union since 1956. Each participating country submits an original song to be performed on live television and radio and then casts votes for the other countries' songs to determine the most popular song in the competition. The contest has been broadcast every year for sixty years, and is one of the longest-running television programmes in the world. It is also one of the most watched non-sporting events in the world, with audience figures varying in recent years from 100 million to 600 million globally[4]. The emergence of social networking in recent years has dramatically changed the range and scope of audience interaction and engagement, particularly for different language communities.

The 2016 Eurovision Song Contest[5] took place in May in Stockholm, Sweden. There were 32 countries taking part, with two semi-finals taking place on 12 and 14 May. 26 countries qualified for the final on 16 May. This yearfis contest was perceived by many commentators to be tense and politically motivated, especially with Ukraine eventually winning the final [42]. Varying analyses see the contest as being influenced by political conflicts, friendships or

[4]https://www.eurovision.tv
[5]https://www.eurovision.tv/page/stockholm-2016/all-participants

cultural bias [3, 7, 12, 17], with a range of news articles explicitly discussing the possibly biased results [24]. Twitter activity was very high throughout the event on the main #Eurovision hashtag. The participation exceeded 7,900,000 statuses, produced by 1,226,959 users; Figure 9 shows the overall Twitter activity.

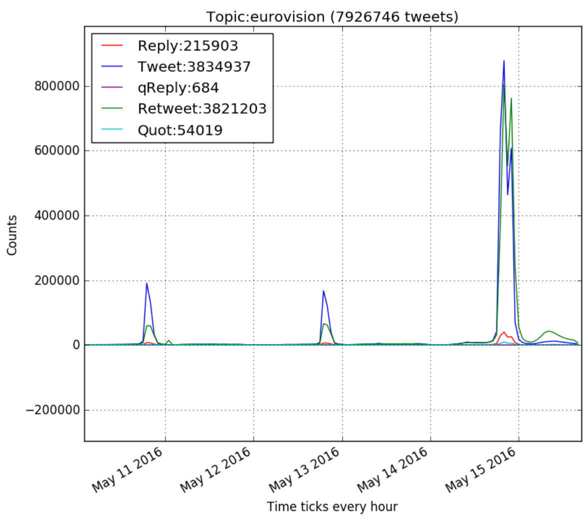

Figure 9: Overall activity for #Eurovision.

Preliminary analysis shows that tweets and retweets together account for c.97% from the total activity, as shown in Figure 9. These two subsets can be representative on their own, without the need to include other interaction sets, such as replies and quoted tweets. 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 the usage of different languages in this set.

3.1 Locations

As mentioned in Section 1, for the sake of anonymity many users tend not to disclose information about their identity, particularly locations; this has also been supported by the literature that geotagged tweets are generally low in number [23]. We took the step to verify this claim in our datasets; in the best cases, the ratio of geotagged tweets did not exceed 2%. An alternative location-based option to consider is based on profile location, but it still does not serve the need for location clustering for a multitude of reasons. Firstly, we found that less than 45% of users have set their profile location, which is in line with other studies [19]. Secondly, although Twitter suggests certain presets for setting profile location, users are given the option to enter any text they wish; this results in a considerable amount of noise.

3.2 Profile and Posting Communities

In the #Eurovision case, there were 49 posting languages. Table 3 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, the results show that the language of 142,721 (3.72%) statuses could not be identified. When investigated further, c.40% of these statuses did not contain much text other than hashtags, user mentions or URLs. Although this category shows an interesting case in which qualitative content analysis could be involved, it is beyond the scope of this study and will not be addressed 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 3: Most active profile language communities, accounting for 90% of original tweets

In total, 1,226,959 users interacted with the #Eurovision hashtag. In terms of their profile languages, they formed 50 communities. Table 4 shows the profile communities from the top 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 had selected a language and no users with the default value.

Community	%
en	47.06
es	20.37
fr	8.00
ru	7.07
de	3.539
nl	3.31
it	2.25

Table 4: Profile communities, for top 90% of users

3.3 Profile-Posting Analysis

From the previous two tables, we can see some similarities between the posting and profile communities. Taking an exceptional case as an example, we can see that although the French profile community had more presence, the Russian posting community is larger by 2.79%. A simple explanation would be that the Russian profile community was relatively more active than French due to the focus on related countries; another reason could be the participation of non-Russian profiles using the Russian language for posting. To investigate this, we investigated the contribution of profile communities to the Russian posting community. The result in Table 5 shows profile communities that resulted in more than 95% of activity in this posting community.

Community	%
ru	91.25
en	7.26

Table 5: Active profile communities within the Russian posting community

As we can see in this example, posts in Russian were not merely appearing from the Russian profile community. This show one way of exploring relationships between profile and posting communities, especially if we are interested in particular communities.

Another approach is to explore the posting behaviour of one particular community. When considering certain profile communities, there is a tendency to assume that communities only post in languages that are the same as their profile language. To examine this assumption, we investigated participation of '*en*' profiles, as they form 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 6 lists the top 95% posting languages used by this profile community.

Language	%
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 6: Top 95% of participation languages from '*en*' profiles

3.4 Language Diversity

By observing the language diversity of profile communities, we aim to measure language diversity of the topic in general, as well as investigating which community plays a key role in bridging different profile communities. Diversity in this context 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 c.17%, while 3.72% were not identified. All of the 50 profile communities used different languages 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 discussed in Section 3.2, Table 7 shows their diversity by percentage. The Russian profile community is again an interesting case, as it scored the least diverse profile amongst all the 50 communities although it comes fourth in number of users.

Language	%
de	34.27
nl	32.78
it	18.49
fr	16.65
en	15.39
es	10.13
ru	7.93

Table 7: Diversity of the top profile communities

3.5 Multilingual Communities

In this section, we group users based on their relationship with posting communities, regardless of their profile language. For example, a user posting in both '*en*' and '*fr*' will be classified as bilingual, and so on. Based on this grouping technique, with the '*und*' lang category 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 10, monolingual users scored about 85% of all users, creating 47% of the total original posts. The also shows that users and their activity decrease as number of languages used increase.

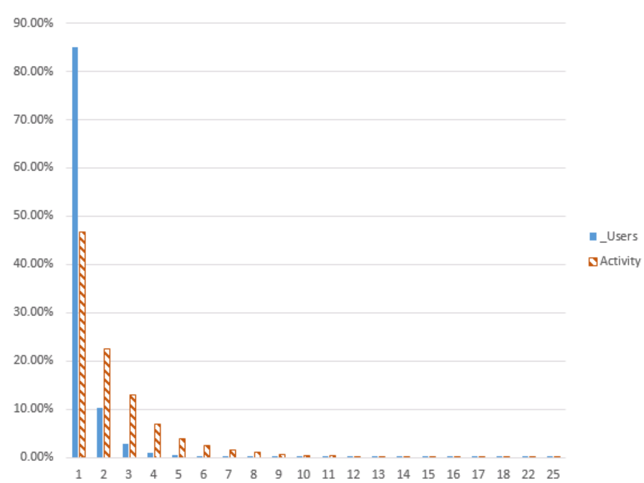

Figure 10: Multilingual communities and their associated activities.

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

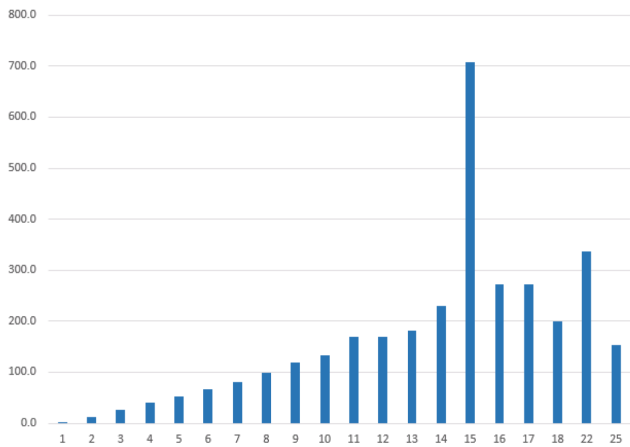

Figure 11: Average number of posts per user for multilingual communities.

4 CONCLUSIONS

This paper presents two real-world case studies – the 2015 Baltimore protests in the USA and the 2016 Eurovision Song Contest – in identifying languages used, language and multilingual communities, and their engagement and interactions on the Twitter platform. As we discussed in Section 1.5, the nature of the event (e.g. being a local or global) may be reflected on community conversations on Twitter. We found that most of posting activity comes from the main community (the language community in which the incident has happened or tightly related to). This is especially the case when the online conversations are triggered by a real world incident. The same for posting languages, users mostly to use the language of the main community. Although most of topic activity comes from that main community, we noticed that, with local events, other communities work, mostly, as information spreaders. Furthermore, there is a positive relationship between size of profile and posting communities; we have also showed that a large number in participating profile community does not necessarily imply high language diversity, and that diversity may results from small profile community. We also presented the 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 a few cases, users may use a significant number of languages, up to 25 different languages. These extreme cases may be interesting to investigate for possible spammer/false account detection or for sociolinguistics in more moderate cases.

We also presented a network graph (using Gephi[6] and the Networkx Python package[7]) showing how language communities relate to each other in the form of action-reaction (action: tweet, reaction: retweet, reply, quote). Another interesting graph that we produced to show relations between profile and posting communities. We find that this graph is important to facilitate comparing users defined profile language with their posting language. Some event might be termed as 'partially scheduled' as their end was different to how they were planned in the first place. In such tense situation, we noticed that diversity of languages and communities are very low, and there always be a dominating community and language.

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. Applying these techniques on data pouring from the Twitter Stream API[8] would be applicable to a wide number of domains. For example, these methods can be used in social network marketing and publicity to increase the probability of influential posts. In practice, for a given `#<Brand>`, by monitoring the activity of different language community, one can decide the time to post well-tailored tweets targeting certain communities. This can be fine-tuned further by mentioning key players in that community, e.g. users with high closeness scores.

Moreover, within certain contexts, the order of applying these two classifications (posting and profile) will generate different results. 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 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 the social network analysis domain, this method provides a different perspective for influence analysis. Endorsement from different profile communities cannot be measured similar to those coming from the same community. For example, in a controversial Arabic topic, we noticed that high support came from other profile communities.

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 reaction networks. We believe that differentiation between endorsements (e.g. retweets) and other reactions may provide further insight into the networks and communities. Also, we plan to add further classifications to

[6] https://gephi.org/
[7] https://networkx.github.io/
[8] https://dev.twitter.com/streaming/overview

the reaction network (as presented in Section 2.5). 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, music contests and civil rights/humanitarian actions.

ACKNOWLEDGMENTS

This work has been supported by a doctoral research scholarship for Nabeel Albishry from King Abdulaziz University, Kingdom of Saudi Arabia.

REFERENCES

[1] Benjamin Blamey, Tom Crick, and Giles Oatley. 2012. R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora. In *Research and Development in Intelligent Systems XXIX*. Springer, 207–212.

[2] Benjamin Blamey, Tom Crick, and Giles Oatley. 2013. 'The First Day of Summer': Parsing Temporal Expressions with Distributed Semantics. In *Research and Development in Intelligent Systems XXX*. Springer, 389–402.

[3] Marta Blangiardo and Gianluca Baio. 2014. Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models. *Journal of Applied Statistics* 41, 10 (2014), 2312–2322.

[4] Stephen P. Borgatti and Martin G. Everett. 2000. Models of core/periphery structures. *Social Networks* 21, 4 (2000), 375–395.

[5] Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2012. Locating privileged spreaders on an online social network. *Physical Review E* 85, 066123 (2012).

[6] Axel Bruns, Tim Highfield, and Jean Burgess. 2013. The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks. *American Behavioral Scientist* 57, 7 (2013), 871–898.

[7] Oliver Budzinski and Julia Pannicke. 2016. Culturally biased voting in the Eurovision Song Contest: Do national contests differ? *Journal of Cultural Economics* (2016), 1–36.

[8] Peter Burnap, Omer Rana, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, Luke Sloan, and Javier Conejero. 2015. COSMOS: Towards an integrated and scalable service for analysing social media on demand. *International Journal of Parallel, Emergent and Distributed Systems* 30, 2 (2015), 80–100.

[9] Pete Burnap, Matthew L. Williams, Luke Sloan, Omer F. Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4, 1 (2014).

[10] James Caverlee, Zhiyuan Cheng, Daniel Z. Sui, and Krishna Yeswanth Kamath. 2013. Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media. *IEEE Data Engineering Bulletin* 36, 3 (2013), 33–41.

[11] Meeyoung Cha, Fabrício Benevenuto, Hamed Haddadi, and Krishna Gummadi. 2012. The World of Connections and Information Flow in Twitter. *IEEE Transactions on Systems, Man, and Cybernetics* 42, 4 (2012), 991–998.

[12] Nicholas Charron. 2013. Impartiality, friendship-networks and voting behavior: Evidence from voting patterns in the Eurovision Song Contest. *Social Networks* 35, 3 (2013), 484–497.

[13] Z. Cheng, J. Caverlee, and K. Lee. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM'10)*. ACM Press, 759–768.

[14] Francesca Comunello and Giuseppe Anzera. 2012. Will the revolution be tweeted? A conceptual framework for understanding the social media and the Arab Spring. *Islam and Christian-Muslim Relations* 23, 4 (2012), 453–470.

[15] Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. 2011. Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In *Proceedings of the Workshop on Languages in Social Media (LM'11)*. 58–65.

[16] Matthew S. Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.

[17] Victor Ginsburgh and Abdul G. Noury. 2008. The Eurovision Song Contest. Is voting political or cultural? *European Journal of Political Economy* 24, 1 (2008), 41–52.

[18] Victor Gisnburgh and Shlomo Weber. 2011. *How Many Languages Do We Need? The Economics of Linguistic Diversity*. Princeton University Press.

[19] Mark Graham, Scott A. Hale, and Devin Gaffney. 2014. Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer* 66, 4 (2014), 568–578.

[20] Elin Haf Gruffydd Jones and Enrique Uribe-Jongbloed (Eds.). 2013. *Social Media and Minority Languages: Convergence and the Creative Industries*. Multilingual Matters Ltd.

[21] Lichan Hong, Gregorio Convertino, and Ed H. Chi. 2011. Language Matters In Twitter: A Large Scale Study. In *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*.

[22] Philip N. Howard, Aiden Duffy, Deen Freelon, Muzammil M. Hussain, Will Mari, and Marwa Maziad. 2011. Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? (2011). available at: http://dx.doi.org/10.2139/ssrn.2595096.

[23] Ruogu Kang, Stephanie Brown, and Sara Kiesler. 2013. Why do people seek anonymity on the internet?: informing policy and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2657–2666.

[24] Ashley Kirk, Jack Kempster, and Stefano Franco. 2016. Eurovision 2016: How does country bias affect the result? http://www.telegraph.co.uk/music/news/eurovision-2016-how-country-bias-affects-the-result. (May 2016). (accessed 2016-10-28).

[25] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. 2012. Geographic Dissection of the Twitter Network. In *Proceedings of the 6th International AAAI Conference on Web and Social Media (ICWSM)*.

[26] Shamanth Kumar, Fred Morstatter, and Huan Liu. 2014. *Twitter Data Analytics*. Springer.

[27] Renaud Lambiotte and Michal Kosinski. 2014. Tracking the Digital Footprints of Personality. *Proceedings of the IEEE* 102, 12 (2014), 1934–1939.

[28] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323, 5915 (2009), 721–723.

[29] Yuan Liang, James Caverlee, Zhiyuan Cheng, and Krishna Y. Kamath. 2013. How big is the crowd?: event and location based population modeling in social media. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT'13)*. 99–108.

[30] Wei Liu, Matteo Pellegrini, and Xiaofan Wang. 2014. Detecting Communities Based on Network Topology. *Scientific Reports* 4, 5739 (2014).

[31] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and Danah Boyd. 2011. The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions. *International Journal of Communication* 5 (2011), 1375–1405.

[32] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitterfis Streaming API with Twitterfis Firehose. In *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM)*. 400–408.

[33] Mohammed Mostafa, Tom Crick, Ana C. Calderon, and Giles Oatley. 2016. Incorporating Emotion and Personality-Based Analysis in User-Centered Modelling. In *Research and Development in Intelligent Systems XXXIII*. Springer.

[34] Giles Oatley and Tom Crick. 2014. Changing Faces: Identifying Complex Behavioural Profiles. In *Human Aspects of Information Security, Privacy and Trust*. Lecture Notes in Computer Science, Vol. 8533. Springer, 282–293.

[35] Giles Oatley and Tom Crick. 2015. Measuring UK Crime Gangs: A Social Network Problem. *Social Network Analysis and Mining* 5, 1 (2015).

[36] Rob Procter, Jeremy Crump, Susanne Karstedt, Alex Voss, and Marta Cantijoch. 2013. Reading the riots: what were the police doing on Twitter? *Policing and Society* 23, 4 (2013), 413–436.

[37] M. Puck Rombach, Mason A. Porter, James H. Fowler, and Peter J. Mucha. 2014. Core-Periphery Structure in Networks. *SIAM Journal on Applied Mathematics* 74, 1 (2014), 167–190.

[38] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8, 9 (2013).

[39] Luke Sloan, Jeffrey Morgan, William Housley, Matthew L. Williams, Adam Edwards, Pete Burnap, and Omer F. Rana. 2013. Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online* 18, 3 (2013).

[40] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. 2012. Geography of Twitter networks. *Social Networks* 34, 1 (2012), 73–81.

[41] Li Tan, Suma Ponnam, Patrick Gillham, Bob Edwards, and Erik Johnson. 2013. Analyzing the impact of social media on social movements: A computational study on Twitter and the Occupy Wall Street movement. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[42] The Telegraph. 2016. Eurovision 2016: Furious Russia demands boycott of Ukraine over Jamala's 'anti-Kremlin' song. http://www.telegraph.co.uk/news/2016/05/15/eurovision-2016-furious-russia-demands-boycott-of-ukraine-over-j. (May 2016). (accessed 2016-10-28).

[43] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the 4th International AAAI Conference on Web and Social Media (ICWSM)*.

[44] UK Parliamentary Office of Science and Technology. 2014. *Social Media and Big Data*. Technical Report Report 460.

[45] Wikipedia. 2015. 2015 Baltimore protests. https://en.wikipedia.org/wiki/2015_Baltimore_protests. (2015). (accessed 2016-08-14).

[46] Alistair Willis, Ali Fisher, and Ilia Lvov. 2015. Mapping networks of influence: tracking Twitter conversations through time and space. *Participations: Journal of Audience & Reception Studies* 12, 1 (2015), 494–530.

[47] Gadi Wolfsfeld, Elad Segev, and Tamir Sheafer. 2013. Social Media and the Arab Spring: Politics Comes First. *The International Journal of Press/Politics* 18, 2 (2013), 115–137.

[48] Michele Zappavigna and J. R. Martin. 2012. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Continnuum.

[49] Jian-Xiong Zhang, Duan-Bing Chen, Qiang Dong, and Zhi-Dan Zhao. 2016. Identifying a set of influential spreaders in complex networks. *Scientific Reports* 6, 27823 (2016).

[50] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. 2011. Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”. *Procedia – Social and Behavioral Sciences* 26 (2011), 55–62.