“*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 tools and methodologies are 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 indicating 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 actions – when participating in large-scale Twitter conversations.

KEYWORDS

Language networks, Multilingual communities, Community discovery, Online participation, Social networks

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:

1 INTRODUCTION

1.1 Online Social Networks

In recent years, online social networks have been utilised as means to express ideas and opinions, share information about events, or even stimulate and propagate calls for civic engagement and societal action. Social networking sites such as Twitter, Facebook, YouTube and Instagram 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 specialist and generalist 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 referred to as big data: volume, velocity and variety [46]. In this age of big social 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 these big social data to provide insight into real-world events and actions [8, 29].

Recent work [1, 2, 34, 35, 37, 40] has analysed what people say and do 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. Data can thus be collated and aggregated, inferring gender, age, location and sentiments, from large-scale 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 [28, 46].

There are various projects that have used Twitter corpora and related datasets to make predictions about elections [45], stock

N.B. The first part of the title of this paper was taken from the motto of the 2016 Eurovision Song Contest, which along with the theme artwork was said to be “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
© 2017 Copyright held by the owner/author(s). . . .$15.00
DOI:

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

markets [52], crimes and policing [17, 36], even allowing us to quantify controversy for topics that spark the most heated debates on social media [16]. Twitter played an important role during what was then known as the "Arab Spring", which has been extensively examined in the social network analysis domain [6, 14, 23, 32, 49]. While the use of Twitter data has been demonstrated to provide insight – and sociologically relevant demographics [41] – into major social and physical events such as riots [38] and terror attacks [9], often all is not what it may seem; for instance, many tweets may not a crowd make [30].

1.2 Languages

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 [22]. Language has featured as a facet of research on the geographies of Twitter networks [42], especially whether offline geography still matter in online social networks [26]. 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 [50], 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 [21], as well as investigations into the economics of linguistic diversity [19].

1.3 Social Network Analysis

In the social network analysis 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, 31, 39]. 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, 51]. 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 [36, 48].

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 [24]. It has also been reported in the literature that geotagged tweets are generally low in number [27, 33, 43], the exponential growth in social media over the past decade has been joined by the rise of location as a central organising theme [30] of how users engage with online information services and, more importantly, with each other [10, 13].

1.4 Overview of Paper

The remainder of this paper is organised as follows: in Section 2 we introduce the methodology and key language themes. Sections 3 and 4 present the 2015 Baltimore protests and 2016 Eurovision Song Context case studies, along with an analysis of the key data and results. Section 5 concludes the paper with a wider discussion and a summary of the potential application of our approach.

2 METHODOLOGY

The primary purpose of this study is to examine if the nature of event is reflected in language uses, communities, and diversity on Twitter. 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[2]. The first step is to identify relationships between language settings of users and languages used in original posts (tweets). Then, we will discuss language diversity and how it is affected by the nature of topic.

2.1 Users and Locations

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[3]: "*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 [24]. 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 [20]. 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 Language Communities

Analysis of language communities begins with two basic techniques. The first is to classify statuses based on their languages, with the status language 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. Output from this technique will be referred to as '*profile communities*'. We can then create network graphs to explore relationships between profile and posting communities. As we will see in the following two case studies, a posting community does not necessarily indicate the profile community for a user.

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

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

2.3 Language Diversity

In this section, we present two language diversity measurements. The first, which we call 'diversity', is to measure uses of languages different to the profile's i.e. we are referring to 'non-selfloop' edges in the generated network graphs. The second one is to measure the magnitude of this diversity, which can be calculated as the total weight of 'non-selfloop' over total edge weights in the profile-posting graph. We will make observation on these two measurements for both cases.

By observing the language diversity of profile communities, we aim to measure language diversity of the topic in general. The same technique can be applied on individual communities within the topic. By doing so, it can help identify communities that act as bridges between different profile communities. Moreover, the technique can be applied to measure diversity at individual users level.

3 CASE STUDY: 2015 BALTIMORE PROTESTS

Following a peaceful funeral 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.*" [47]. The nature of the Baltimore protests is a good representation of a partially planned event in which a sudden escalation of civil unrest hits a geographical area. The event manifested itself on Twitter as `#BaltimoreRiots`, and resulted in more than 1,250,000 status updates.

Figure 1 shows how the event manifested itself on Twitter once a "purge" was scheduled. We can see that what was happening on the ground was quickly reflected in the online activity on Twitter. More detailed analysis reveals that within one hour the topic started to go "viral"; more precisely, at approximately 15:00 on 27 April 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.

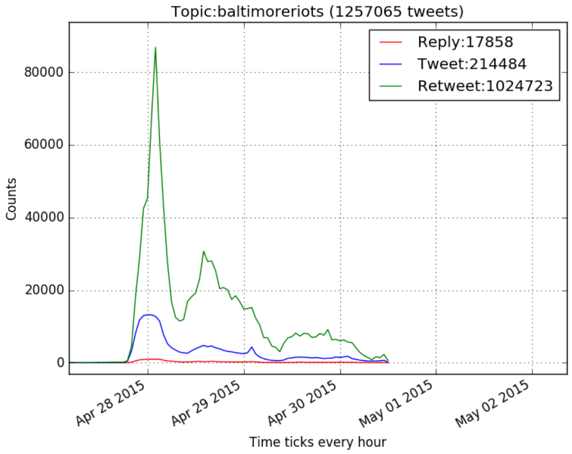

Figure 1: Overall activity for `#BaltimoreRiots`.

3.1 Posting Communities

In the `#BaltimoreRiots` case, 38 posting languages were used for original posts. As we can see in Figure 2, English was the mostly frequently-used language by far. Interestingly, results also show that language of more than c.7% statuses could not be identified. When investigated, those statuses mostly did not contain text other than hashtags, pictures or URLs. Although, this is not a big proportion of the overall sample, it came second after English. This category shows an interesting case in which qualitative content analysis could potentially be used, but it is beyond the scope of this study and will not be covered here.

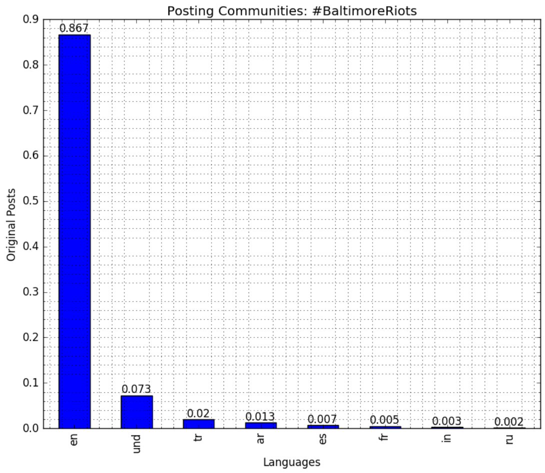

Figure 2: Most frequently used languages in `#BaltimoreRiots`: (*en:* English; *es:* Spanish; *tr:* Turkish; *fr:* French; *en-gb:* British English; *ar:* Arabic; *de:* German; *ru:* Russian; *it:* Italian; *pt:* Portuguese)

3.2 Profile 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 opted to select a language. However, the language entity returned by the API for those cases is the initial placeholder text "*Select Language...*" or a translated version that may provide hints to the user language community. Figure 3 shows that about 94% of the users came from '*en*' profile community.

As we can see in Figure 4, activity from profile communities is not far from their relative 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 civil unrest and terrorist attacks.

3.3 Profile-Posting Network

This section explores the network graph we are able to build from profile-posting relationships; constructing this graph is an essential step for our core analysis of language diversity. For example, to

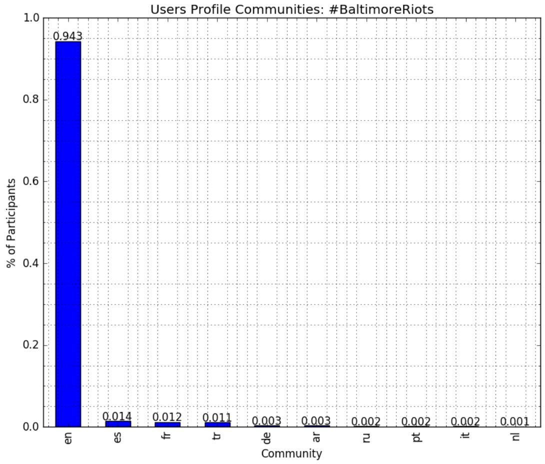

Figure 3: Top 10 profile language communities in `#BaltimoreRiots`

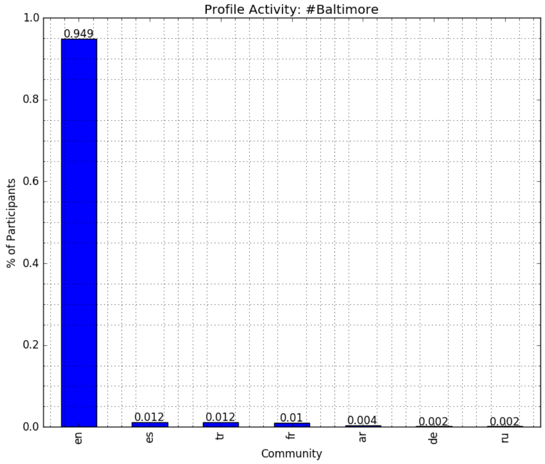

Figure 4: Profile language communities causing %99 of activity in `#BaltimoreRiots`

investigate whether the '*en*' posting community is linked to particular profile communities, we used the bipartite graph as presented in figure 5. In this graph, nodes that are prefixed by "*p_*" represent profile language community, and nodes that are prefixed by "*s_*" represent posting language community. The size of node represents the weighted indegree, whereas colour represents the outdegree: the darker the colour, the higher outdegree; hence, completely white nodes have no outdegree and help to easily distinguish posting communities from profile ones. From the graph 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.

3.4 Diversity and Multilingual Communities

To examine users' behaviour in using languages other than their own (profile language), same-language communities are filtered out. We found that English profile users frequently posted in Arabic, as a secondary posting language. This sort of activity could be obtained from edge weights in the profile-posting graph. Table 1 shows the top portion of these communities. This observation sheds light on those relationships, and could be of use for further analysis, such as identifying highly disseminated message that fall into these relationships and their contents.

Profile-Posting Edge	Weight
en-ar	1.00
es-en	0.49
ar-en	0.47
fr-en	0.38
tr-en	0.37
en-tr	0.36

Table 1: Users' behaviour in using languages different to their profile (normalised)

The topic diversity in `#BaltimoreRiots` is 0.91, which is not surprising when we recall that there were 38 languages used in original tweets. However, when the magnitude of diversity was measured, it gave a score of 0.04, a generally low score. This will be of use when compared to the data from our second case study in Section 4.

Additionally, measuring multilingual communities is another use of profile-posting network graph at individual users' level. We grouped 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 nine sets. As we can see in Figure 6, monolingual group contain the overwhelming majority of users (nearly 99%), contributing approximately 94% of the topic activity too.

4 CASE STUDY: 2016 EUROVISION SONG CONTEST

The Eurovision Song Contest (*Concours Eurovision de la chanson*) – generally known as 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

[4] https://www.eurovision.tv

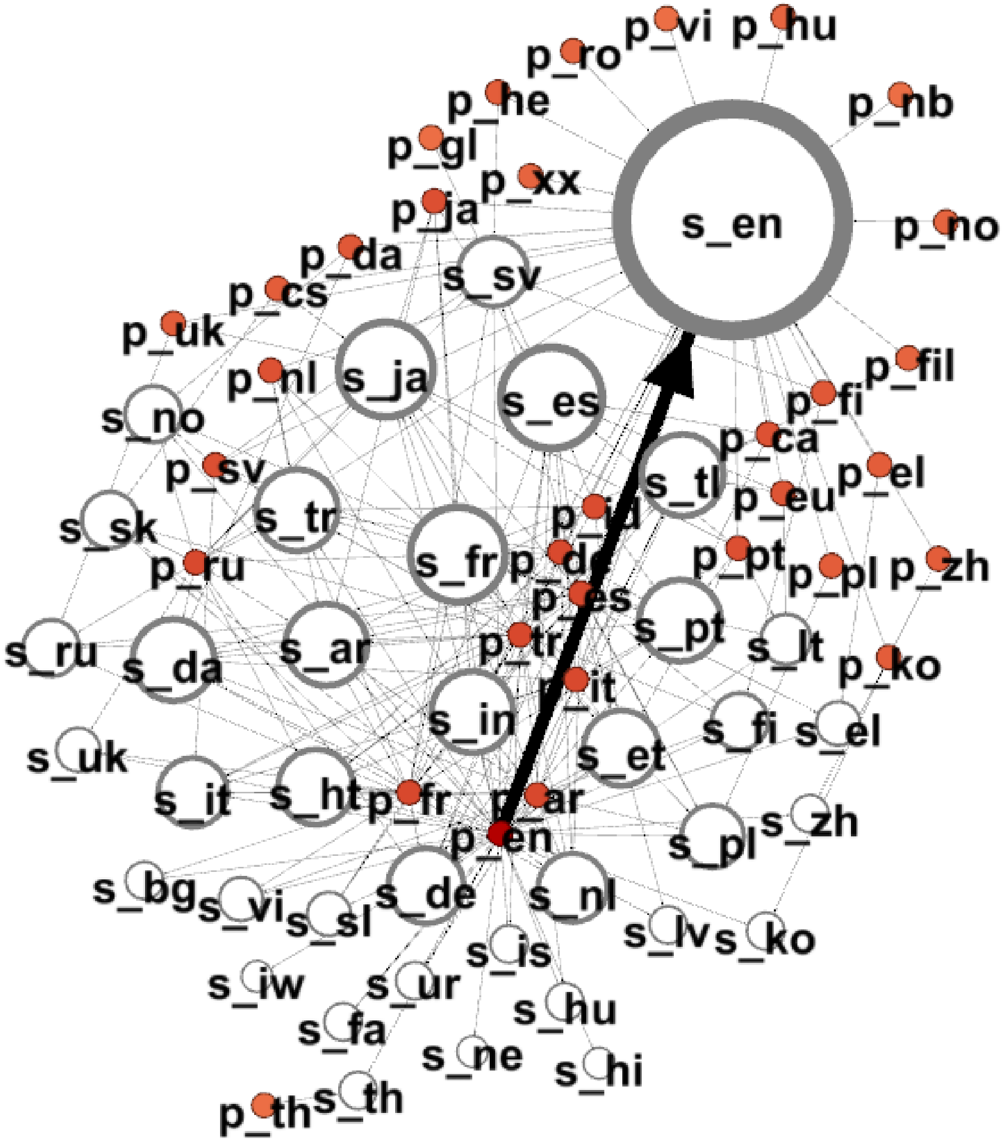

Figure 5: Profile-posting network graph

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, with 26 countries qualifying 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 [44]. Varying analyses see the contest as being influenced by political conflicts, friendships or cultural bias [3, 7, 12, 18], with a range of news articles explicitly discussing the possibly biased results [25]. Twitter activity was very high throughout the event on the primary #Eurovision hashtag. The participation exceeded 7,900,000 statuses, produced by 1,226,959 users; Figure 7 shows the overall Twitter activity.

Preliminary analysis shows that tweets and retweets together account for 97% from the total activity; 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.

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

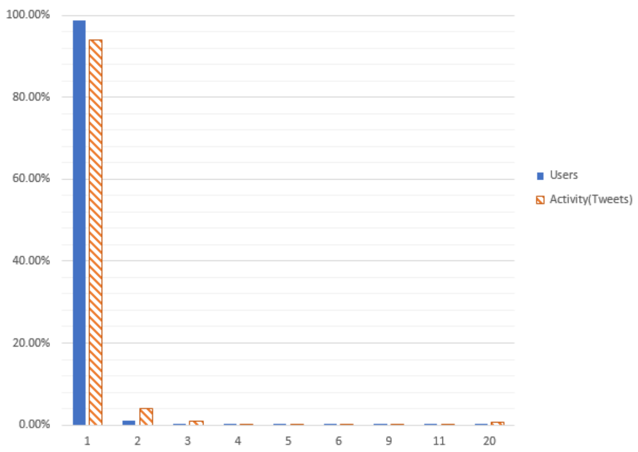

Figure 6: Multilingual communities and their activity in #BaltimoreRiots

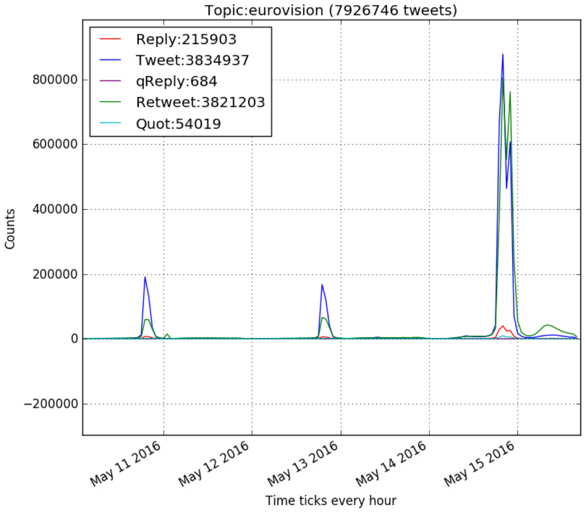

Figure 7: Overall activity for #Eurovision.

4.1 Posting Communities

In the #Eurovision case, there were 49 posting languages. Fig 8 shows the top posting languages (tweets), out of 3,834,937. As might be expected, English was the most frequently-used posting language, while c.4% statuses could not be identified. However, as we can see, it is not as dominating as in the #Baltimoreriots case. Interestingly, it took participation from eight posting communities to match the activity ratio of the '*en*' posting community in #Baltimoreriots (0.87).

4.2 Profile Communities

In total, 1,226,959 users interacted with the #Eurovision hashtag. In terms of their profile languages, they formed 50 communities.

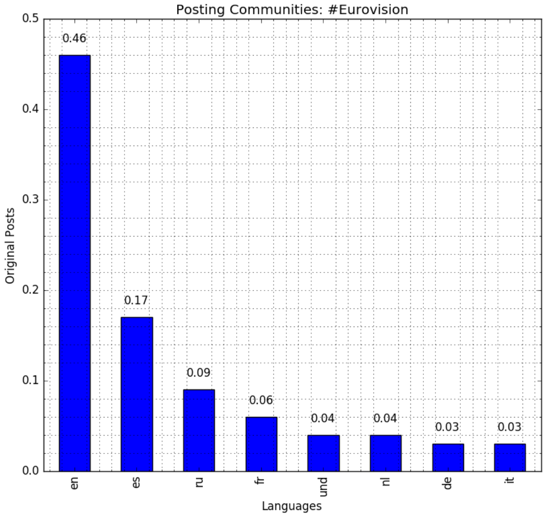

Figure 8: Most frequently used languages in #Eurovision.

Figure 9 shows some of the top profile communities from of all users, ordered by size. 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 #Eurovision dataset, we found that all users had selected a language and no users with the default value. For completeness, we also show profile communities grouped by their activity in 10.

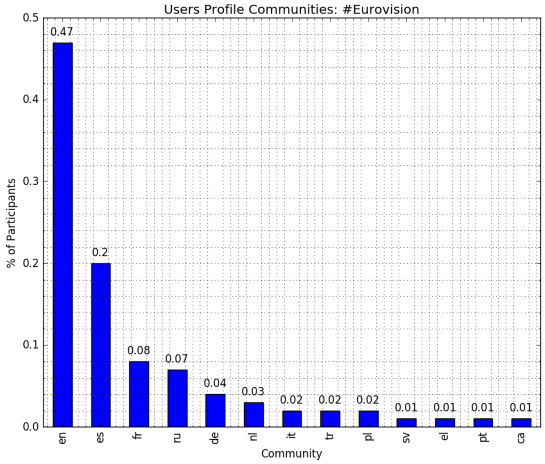

Figure 9: Profile communities by size in #Eurovision.

From the previous three figures, we can see clear similarities between the posting and profile communities. Statistically, we found that these three measures (i.e. profile community size, profile

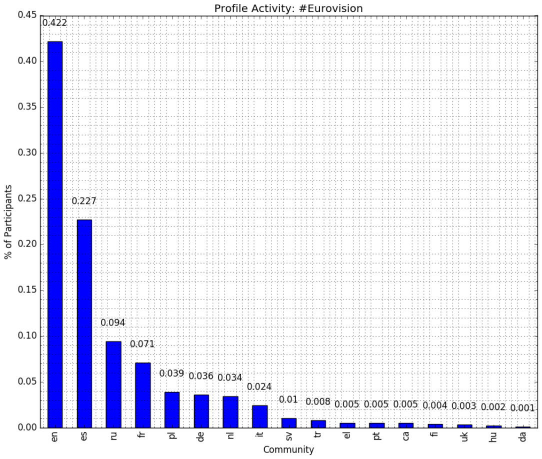

Figure 10: Profile communities by activity in #Eurovision.

activity, and frequency of language uses) are highly correlated, for both case studies. An interesting example to explore here is the comparison between two profile communities, '*fr*' and '*ru*' . We can see that although the French community had more presence, the Russian posting community is greater. 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; further discussion of this example follows in the next section.

4.3 Profile-Posting Analysis

To explore the posting behaviour from profile communities, we constructed the graph shown in Figure 11, allowing us to continue analysing the '*fr*' and '*ru*' example. The graph helps us to evaluate contributions of profile communities to the Russian posting community. The result in Table 2 shows profile communities that resulted in more than 95% of activity in this posting community. The values in this table represent the weights from those profile communities to the Russian posting community.

Community	%
ru	91.25
en	7.26

Table 2: Active profile communities within the Russian posting community

Posts in Russian were not just 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 3 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 3: Top 95% of participation languages from '*en*' profiles

4.4 Diversity and Multilingual Communities

All of the 49 profile communities have used different languages in posting. 16 out of those communities did not use their own language, although they were low in numbers of tweets and engagement. Moreover, in terms of using different languages, we found that 32 communities scored at least 50% out of their original tweets.

We also wanted to measure the diversity and its magnitude for the topic. We found that the language diversity of the topic as a whole is 0.96. Although the diversity here is not far from #BaltimoreRiots case, when we compare magnitude of diversities, it scored 0.22, which is relatively high.

In terms of multilingual communities, we group users based on their relationship with posting communities, as we did in the #BaltimoreRiots case. 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 12, monolingual users scored about 85% of all users, creating 47% of the total original posts. 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.

5 CONCLUSIONS

This paper presents work in identifying languages used, language and multilingual communities using two real-world case studies: the 2015 Baltimore protests in the USA and the 2016 Eurovision Song Contest, and their associated engagement and interactions on Twitter. As we discussed in Section 2.2, the nature of the event (e.g. being a local or global) may be reflected on community conversations on Twitter. We found that the majority of posting activity comes from the main community (the language community in which the incident has happened or closely related). This is

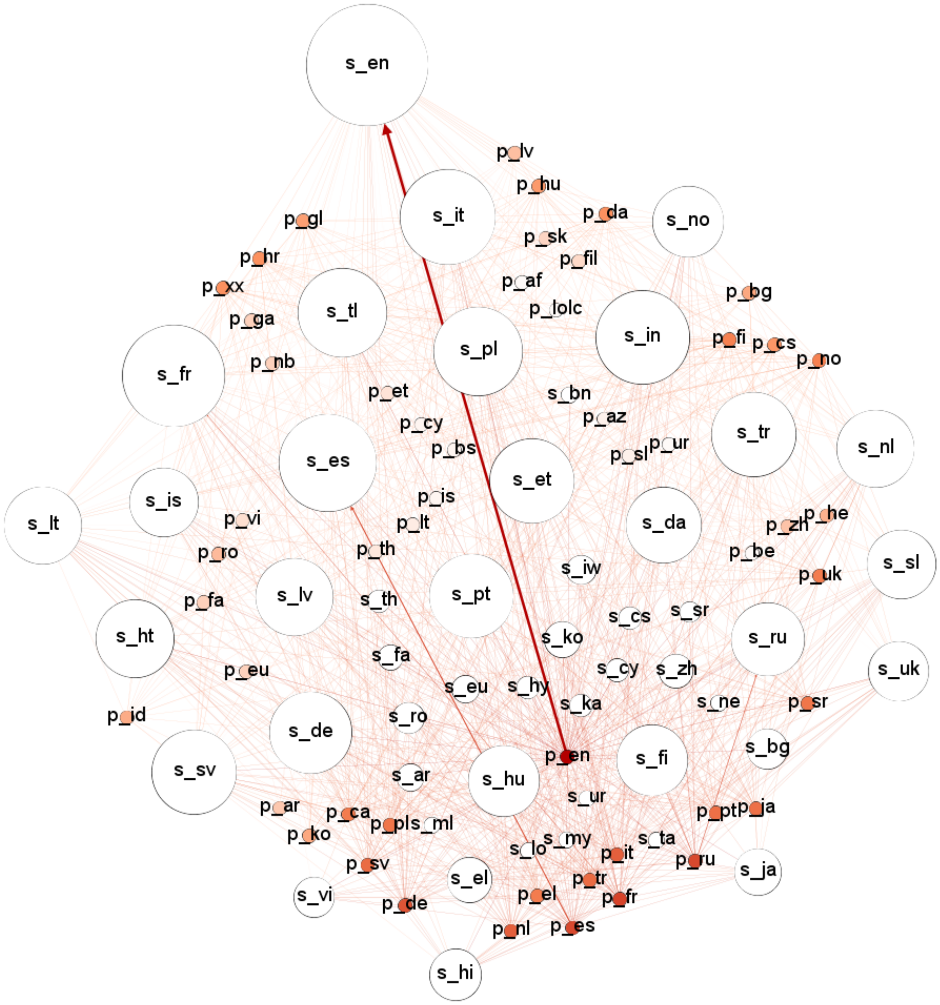

Figure 11: Profile-posting network graph for #Eurovision

especially true when the online conversations are triggered by a real-world incident. The same is the case for posting languages – users mostly use the language of the main community. Furthermore, there is a positive relationship between size of profile and posting communities; we have also demonstrated that a large number of people in participating profile communities does not necessarily imply high language diversity, as single posts from many profile communities are enough to dramatically affect the diversity of the topic. We also presented the diversity magnitude measurement, and showed that it is highly effective in eliminating, or at least minimising, the noise caused by those odd posts. We also discussed the structure of multilingual communities and their activity. 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 graphs showing how language communities relate to each other, as well as relations between profile and posting communities. We find that this second graph is important to facilitate comparing users' defined profile language with their posting language. Some events might be termed as 'partially scheduled' as their end was different to how they were planned in the first place; in such situations, we noticed that there always be a dominating community and language.

The methods we have 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[6] 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

[6] https://dev.twitter.com/streaming/overview

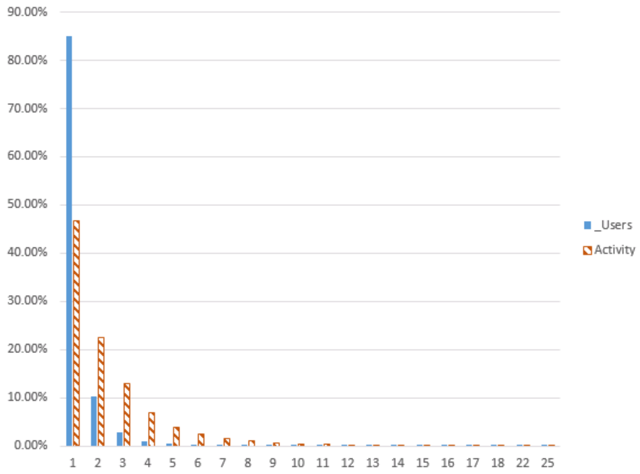

Figure 12: Multilingual communities and their associated activities.

given #<Brand>, by monitoring the activity of different language community, one can decide the time to post well-tailored tweets targeting certain communities.

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.

As we saw in the comparison between the two case studies, we can conclude that the nature of the event is mostly reflected on the "language diversity magnitude" than any other measurement. For future work, we will explore in more detail how multilingual communities interact and participate, as well as their reaction networks. We believe that differentiating 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, to provide further insight into the methodology and overall approach.

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] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16)*. 33–42.

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

[18] 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.

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

[20] 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.

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

[22] 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)*.

[23] 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.

[24] 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.

[25] 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 2017-02-17).

[26] 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)*.

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

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] Giles Oatley, Tom Crick, and Mohamed Mostafa. 2015. Digital Footprints: Envisaging and Analysing Online Behaviour. In *Proceedings of 2015 Symposium on Social Aspects of Cognition and Computing Symposium (SSAISB)*.

[38] 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.

[39] 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.

[40] 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).

[41] 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).

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

[43] 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)*.

[44] 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 2017-02-17).

[45] 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)*.

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

[47] Wikipedia. 2015. 2015 Baltimore protests. https://en.wikipedia.org/wiki/2015_Baltimore_protests. (2015). (accessed 2017-02-17).

[48] 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.

[49] 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.

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

[51] 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).

[52] 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.