Measuring citizen participation in South African public debates using Twitter: An exploratory study

Selvas Mwanza
ICT4D Research Centre
University of Cape Town
Cape Town, South Africa
smwanza@cs.uct.ac.za

Hussein Suleman
Department of Computer Science
University of Cape Town
Cape Town, South Africa
hussein@cs.uct.ac.za

Abstract

This paper addresses the task of measuring Twitter social attributes that can be used for detecting patterns that show user participation in public debates in South African. We propose a method that leverages observable information on Twitter such as use of language, retweeting user behaviour, and the relationship between topics and the user social network graph. Our experimental results suggest high degrees of citizen participation: people in an otherwise multilingual country tweet in a dominant language; there is more original commentary and interactive discussion; and topics often span natural online communities.

1 Introduction

With the large user base and the ease of publishing content, Twitter has become an ideal platform for many people to communicate and also serves as a platform for expressing opinions on different topics like politics, sports and socio-economic issues. Users on Twitter can converse and interact in different ways. A user can follower another user. A user who follow another user subscribes to receive Twitter messages posted by the followed. Users can reference each other in messages using the @ symbol followed by the username (e.g., I miss @cindy my best friend). Users can also forward a message to others. Twitter adds the key word RT @username at the beginning of all forwarded tweets. The username after the @ symbol is the name of the user who originally posted the message. In addition, Twitter users can use a # symbol to indicate what the message is about.

In 2015, University students in South Africa protested against the increase in school fees (Griffin, 2015). This was mirrored on Twitter when the #FeesMustFall hash tag created for the protests trended on Twitter worldwide. This provides evidence of the adoption of Twitter by citizens in South Africa as a platform to participate in socio-economic issues.

Social media mining is the process of representing, analysing, and extracting actionable patterns from social media data (Zafarani et al., 2014). Twitter data has been mined by different researchers around the world. Examples of Twitter mining includes: financial prediction (Mao et al., 2012), extracting market and business insights (Park and Chung, 2012), political analysis (Monti et al., 2013), mass movement analysis (Borge-Holthoefer et al., 2015) and monitoring of natural disastere and crises (Takeshi et al., 2010). Although a lot of research has been done, little attention has been given to Twitter data produced in Africa.

In this paper, we address the task of measuring citizen participation in public debates on Twitter. We use standard methods like language detection in text, graph partitioning and graph centrality measures to detect patterns of use of language, retweeting user behaviour, and the relationship between topics and user communities to measure user participation in public debates in South African.

The paper is organized as follows. Section 2 introduces the literature review on social media analysis. Section 3 describes in detail our methodology for measuring citizen participation in South Africa using Twitter data, while Section 4 reports on the experiment design and the results. Finally, in Section 5 we discuss the conclusions and outline future work.

2 Literature Review

This section looks at previous work that is related to our work.

2.1 Graph partitioning

Graph partitioning or community detection aims to identify groups in a graph by only using the information encoded in the graph topology (Lancichinetti and Fortunato, 2009). Lancichinetti and Fortunato (2009) reviewed various disjoint community detection algorithms. Disjoint community detection algorithms partition a graph into disjoint groups and has a wide application. Recently, with the introduction of social media mining, attention has been given to overlapping community detection algorithms. Overlapping community detection algorithms identify a set of partitions that are not necessarily disjoint (Xie et al., 2013). A node in the graph can be found in more than one partition. People in social media usually have connections to several social groups like family, friends, and colleagues. Java (2007) used an overlapping detection algorithm called clique percolation method (CPM) to detect overlapping communities in a Twitter network. CMP was used to find how communities connect to each other by overlapped components. Overlapping community detection has also been used to explain how information cascades through Twitter communities (Barbieri et al., 2013). The authors used a community detection algorithm to find the level of authority and passive interest of a node in each community it belongs to.

2.2 Graph centrality measures

Node centrality measures node involvement in the walk structure of a network (Freeman, 1978). Freeman defined three centrality measures, namely, degree, closeness and betweenness. Degree centrality is a count of the number of edges incident upon a given node. Closeness defines the total geodesic (the length of a walk is defined as the number of edges it contains, and the shortest path between two nodes is known as a geodesic) distance from a given node to all other nodes. Betweenness measures the geodesics that pass through a given vertex. Centrality measures have been used in ranking and understanding nodes in social networks. Ediger (2010) used betweenness centrality to rank nodes in clusters of conversations on Twitter data. Betweenness centrality score has also been used to detect spammers in Twitter (Yang et al., 2011). The authors used the betweenness centrality to rank users in a graph then use the ranking score to identify spammers.

2.3 Use of language detection in Twitter

Language detection is the task of detecting the natural language in which a document is written (Lui et al., 2004). Hong and Convertino (Hong and Convertino, 2011) used language detection in Twitter data to discover cross-language differences in adoption of features such as URLs, hashtags, mentions, replies, and retweets. The authors used a combination of LingPipe text classifier and Google language API to to classify 62,556,331 tweets into languages. The data was downloaded for a period of four weeks. The authors then analyzed how each cluster uses URLs, hashtags, mentions, replies, and retweets. Use of language has also been used as a primary tool for detecting spam in tweets (Martinez-Romo and Araujo, 2013). The authors examine the use of language in the topic, a tweet, and the page linked from the tweet. They make an assumption that the language model for a spam tweet will be substantially different: the spammer is usually trying to divert traffic to sites that have no semantic relation. They exploit this divergence between the language models to effectively classify tweets as spam or non-spam.

3 Methodology

In this section we describe in details three types of social attributes that can help in measuring citizen participation: **use of language, retweeting user behaviour,** and **relationship between topics and the user network graph.** We use these three metrics to detect patterns that measure citizen participation in public debates in South Africa.

3.1 Use of Language

South Africa is a multilingual country with nine official languages, namely: English, Afrikaans, Zulu, Xhosa, Ndebele, Northern Sotho, Tsonga, Tswana and Venda. English and Afrikaans are high resource languages while the other languages, which are Bantu languages, are low resource languages. In our work, we are interested in detecting English and Afrikaans in tweets. Tweets that cannot be detected as English or Afrikaans are categorized as other.

Tweets are informal. They contain special tokens such as @ for usernames, # for trending topics and they have http links for related content. They also contain slang, misspellings and grammatical errors. We implemented a program called SATwitterCleaner that cleans the dataset before language detection. Cleaning involved doing the following:

1. Removing usernames: The program removes all usernames in the dataset by searching for words that starts with the @ symbol. This follows the convention that all usernames in Twitter messages are prefixed with the @ symbol.

2. Removing hash tag (#) symbol in the messages: The program removes all hash tags by searching for the # symbol.

3. Removing URLs in the messages: Twitter users reference external sources by inserting URLs in their messages. SATwitterCleaner implements a string pattern that identifies URLs in Twitter messages and removes them.

4. Remove emoticons from the message: An emoticon is a representation of a facial expression used in electronic communication to convey the writer's feelings. The online community uses different types of emoticons for different expressions. We compiled 15 emoticons used for happy expressions and 11 emoticons used for sad expressions. The program used this list to indentify and remove emoticons from messages.

5. Expand slang words into their actual meaning: Slang is the use of informal words and expressions that are not considered standard in the speaker's language or dialect but are considered acceptable in certain social settings. Example: 2b means to be. We created a slang dictionary of 5,364 slang words. Each slang word in the dictionary was mapped to its actual meaning. The slang dictionary was used by SATwitterCleaner to expand all slang words found in the dataset.

6. Correcting spelling and grammatical errors in English tweets: The program employed a LanguageTool library to correct the grammer in tweets. LanguageTool (LT) is based on surface text processing, without deep parsing, yet, it manages to get significantly better results for some languages than commercially available products (Mikowski, 2010). Spelling check and correction was done by using the jazzy spell checker (Idzelis, 2005). Jazzy spell checker integrates the DoubleMetaphone phonetic matching algorithm and the Levenshtein distance using the near-miss strategy. The jazzy spell checker was chosen because it gives suggestions if the word is not properly spelled. SATwitterCleaner employs the spell checker to pick the first option in the suggestion list as a replacement for the mispelled word. The method used in our work for grammar and spell checking is limited to English text as we could not find equivalent libray tools for Afrikaans. Hence, only English text was corrected on grammar and spelling.

7. Replacing repeated characters in words with the correct number of characters: We developed a method for English text that can remove repeated characters in words. English seldom uses words with more than two character repetition. However, there are words with three character repetition. We compiled a list of 21 English words with three character repetition. The program ignores all the words with repeated characters that are found in the compiled list. Otherwise, if a word has repeated characters, the program first reduces the repeated characters to two. Then, using the jazzy spell checker (Idzelis, 2005), the program checks if the word is a correct English word. If not, the spell checker is used to get the suggested close word. The program then computes the cosine similarity distance between the suggested word and the original word. If the distance is below a threshold, the suggested word is taken as a replacement, otherwise the program skips the replacement. We chose the similarity distance threshold of 1.

After data cleaning, we used a combination of the Naive Bayesian method and simple word statistics to detect the English and Afrikaans tweets. We used LangDetect, which implements a Naive Bayes classifier, using a character n-gram based representation without feature selection, with a set of normalization heuristics to improve accuracy (Nakatani, 2010). Lui and Baldwin (2014) compared the performance of eight off-the-shelf language detection systems to determine which would be the most suitable for Twitter data. They compared langid.py (Baldwin et al., 2013), CLD2 (McCandless, 2010), LangDetect (Lui and Baldwin, 2014), LDIG (Nakatani, 2012), whatlang (Brown, 2013), YALI (Majlis, 2012), TextCat (Scheelen, 2003) and MSR-LID (Goldszmidt et al., 2013). They compared the systems on four different Twitter datasets. They found that LDIG outperforms all the algorithms though it supports a limited number of languages and Afrikaans is not one of them. Overall, they concluded that, in their off-the-shelf configuration, only three systems (LangDetect, langid.py, CLD2) perform consistently well on language detection of Twitter messages. Our Twitter messages cleaner, SATwitterCleaner and the language detection program was developed in Java hence we chose LangDetect because it has Java support. Simple word statistics classify tweets by counting the number of words in a tweet that are English or Afrikaans. If the number is higher than or equal to 50%, a tweet is classified as English or Afrikaans respectively. All the tweets that were not detected as English by LangDetect were classified by the simple word statistics. This allowed us to compensate for the inacuracy of the LangDetect system. Only tweets with more than three words were considered for language detection.

3.2 Retweeting user behaviour

Twitter adds the key word RT @username to all forwarded tweets. RT mean retweet and @username refers to the name of the user who originally made the tweet.

In our work, we want to measure how many tweets and retweets are present in the dataset. To find the number of original tweets, we counted all tweets that do not start with RT @. To find the number of retweets in the dataset, we counted all the tweets that starts with RT @ keyword.

3.3 Relationship between topics and user network graph

We created a social graph using retweets. Galuba (2010) showed that retweets is the most powerful mechanism to diffuse information and a strong indication of the direction of information flow in Twitter. We created a graph using retweets because we wanted to see and measure how a graph form around the tweets. Users form vertices in the graph. We add an edge from user @A to user @B whenever @A retweets a tweet from @B. We treat the graph as undirected, so an edge from @A to @B also connects @B back to @A. All loops are discarded from the graph. Loops are formed when a user retweets his/her own tweet. We also ignore duplicate user interactions so that only unique user interactions are represented in the graph. Our graph had 30,114 vertices and 55,578 edges. In this paper we analyse the graph at two different levels, network level and group level. Network level is the view of the entire graph. Group level is the view of sub-graphs/communities in the graph.

Network level

At a network level we calculated the betweenness centrality of all the nodes in the graph. Freeman (1978) defines betweenness centrality as: let g$_{ij}$ denote the number of geodesic paths from node i to node j, and let g$_{ikj}$ denote the number of geodesic paths from i to j that pass through intermediary k. Then the betweenness centrality is defined as follows:

$$C_k^{\mathrm{BET}} = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$$

Betweenness centrality measures the influence/centrality of a node in a graph. According to the definition, a node with high betweenness centrality sits at a connection point of subgraphs. A node plays a major role in the movement of the data from one subgraph to the other. Freeman applied the betweenness to connected and undirected graphs. Social networks often share common characteristics. Natural clusters form, but the clusters do not partition the graph (Mislove et al., 2007). We use this characteristic to make an assumption that our graph will be largely connected and hence the betweenness can be applied.

We also performed another measurement at the network level we called resourceful measure. We calculated how many tweets from each node in the graph have been retweeted at least once. A node with a high resourceful measure has a high number of tweets retweeted at least once by other users. Resourceful measure measures how many tweets each node has contributed in the graph. In our work, we compared the resourceful measure with the betweenness centrality measure of nodes to find the relationship between the top producers of tweets in the graph and the top users who propagate tweets to subgraphs. The Jaccard similarity coefficient (Jaccard, 1902) is a common index for binary variables. It is defined as the quotient between the intersection and the union of the pairwise compared variables among two objects. Jaccard is calculated as follows: given two groups b and c, the percent similarity = [a/(a + b + c)] where a = number of elements present in both groups, b = number of elements present only in group b, and c = number of elements present only in the group c. Jaccard coefficient is a number from 0 and 1. If the coefficient is 0, it means the two groups are completely unidentical. If the coefficient is 1, then the two groups are completely identical. We used the Jaccard coefficient to measure the similarity between the nodes with high betweenness centrality and the nodes with high resourceful measure.

Group level

We partitioned the graph into communities. Xie, Kelley and Szymanski (2013) did a review of the state of the art in overlapping community detection algorithms. They reviewed a total of fourteen algorithms and concluded that, for low overlapping density networks, SLPA (Xie et al., 2011), OSLOM (Lancichinetti et al., 2011), Game (Chena et al., 2010) and COPRA (Gregory, 2010) offer better performance than the other tested algorithms. For networks with high overlapping density and high overlapping diversity, both SLPA and Game provide relatively stable performance. We evaluated two algorithms, namely, COPRA and SLPA. We observed that SLPA performed better than COPRA on our graph both in computer time and modularity. The modularity of a partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities (Girvan and Newman, 2002). After evaluation, we used the SLPA overlapping algorithm for community detection in the graph.

Topic Categories	**Topics**
Controversial topics	#OscarPistorius Kim Martin #FeesMustFall #Sarafina BCCSA Esethu #TaxiStrike Durban protests Mbuyisa
Developmental topics	#ProjectKhanya Wastestopswithme Cleanerjoburg TeamUpToCleanUp #JobSeekersWednesday Durban protests
Entertainment topics	Pearl thusi #ExpressoShow #BangOut #FreshAT5 #KentPhonikFridays #GenNext2016 #iGazi #DateMyFamily #FridayStandIn Jessica Nkosi Ertugral #AskAMan
Political topics	#NandosDMgathering #SpyTapes #ANCGPManifestoLaunch #ANCFriday #FillUpFNBStadium #FillUpFNB FNB Stadium Luthando Mbinda Mavuso
Road accident topics	Bellville N1 North #PTATraffic
National Event topics	#YouthDay Soweto
Other	#WomenMustKnowThat Shoprite #TNABizBrief

Table 1: Keywords used for downloading tweets. Keywords were determined by following trending topics in South Africa from 4th June 2016 to 19th June 2016

4 Results and Discussion

This section discusses our experimental set up and results.

4.1 Experimental Settings

To do this experiment, trending topics in South Africa shown in Table 1 were used to download tweets from 4th June 2016 to 19th June 2016. Twitter implements a proprietary algorithm that shows the trending topics in Twitter data. Trending topics can either be hash tagged words or non hash tagged words. We manually observed trending topics in South Africa from the Twitter website for 16 days and used the Web API to download 131,790 tweets from 37,876 Twitter accounts. The topics were categorized into seven (7) groups, namely: controversial topics, developmental topics, entertainment topics, political topics, road accident topics, national events topics and other.

4.2 Experimental Results

We first start with the results of the language detection. Our experiments show 94.64% of tweets were in English, 2.61% of tweets were in Afrikaans and 2.75% was detected as other. Other means the tweet was neither English nor Afrikaans. During the experiment, we noticed that tweets were repeating in the dataset. This is because users can retweet the same tweet, causing repetition. So, before detecting the language, we filtered out all the repeating tweets. After filtering, the number of tweets in the dataset was reduced to 66,378. The result show that despite having many languages, South Africa tweets in a common language. This pattern suggests that people tweet so that their message can be read across a larger spectrum of the population.

The next result describes the tweet-retweet behaviour. The downloaded dataset had 58.88 % tweets and 41.12 % retweets. This pattern suggests that there is more original contribution in public debates.

The last set of results show the analysis of the social graph. Our results shows that 79.5% of users in our dataset participate in conversation. To measure participation in conversation, we counted all the users in our dataset who retweeted other user's tweets or their tweets were retweeted by others. We used the Jaccard coefficient to measure the similarity of users with high betweenness centrality and users with high resourceful measure. Users with high betweenness centrality play a major role in the movement of tweets in the graph. Users with high resourceful measure have a high number of tweets retweeted at least once by other users. We took the top 50 users with the highest resourceful measure and top 50 users with the highest betweenness centrality and computed the Jaccard coefficient. The coefficient is the number between 0 and 1. A coefficient of 0 means the two groups are completely unidentical. If the coefficient is 1, then the two groups are completely identical. Our calculation yielded a coefficient of 0.23. This result concludes that, top users who provide information in the graph are not the top users who propagate the tweets through communities. Finally, we compared topics in the communities to find overlaps. SLPA (Xie et al., 2011) was used to partition the graph into communities. SLPA is a non-determistic algorithm, so we ran the algorithm 11 times and recorded the average performance. The algorithm produced 2,200 communities with an overlap of 7.3%. This shows that our graph had a low overlapping density. Table 2 shows that all communities tweeted about Oscar Pistorius and there is not a clear cut division among communities with regards to topics. Though communities focus on certain topics - group 5 and 10 talk more about political topics, group 3 entertainment topics, all communities talk about other issues too. These graph patterns suggests that citizens participate in public debates on a variety of topics.

5 Conclusions and Future Work

We presented social attributes that help identify patterns that measure citizen participation in public debates in South Africa. Africa is highly multilingual, hence we chose the use of language as an attribute that can indicate participation in online public discussions. We also considered user retweeting behavior and how topics relate to online communities. This exploratory study provides the first step in Twitter analysis on South African online data. This paper considers only a snapshot of the South African Twitter data. In future, we aim to consider the temporal aspects of the graph.

Group No.	No. Members	Top 4 topics mentioned in each community	% of topics mentioned from all the topics in Table 1
1	10564	#OscarPistorius (4043) #NandosDMgathering (175) #Sarafina (89) #SpyTapes (71)	72.09%
2	1704	#BangOut (419) #OscarPistorius (18) #Sarafina (8) #AskAMan (8)	37.21%
3	1330	#AskAMan (567) #OscarPistorius (88) Shoprite (21) #BangOut (20)	65.12%
4	333	#BangOut (87) #Sarafina (14) #AskAMan (6) #OscarPistorius (9)	30.23%
5	300	#NandosDMgathering (94) #ANCGPManifestoLaunch (35) #OscarPistorius (16) #YouthDay (5)	30.23%
6	288	#FreshAT5 (96) #KentPhonikFridays (10) #BangOut (4) #OscarPistorius (4)	13.95%
7	273	#BangOut (81) #OscarPistorius (14) #GenNext2016 (11) #AskAMan (4)	23.26%
8	147	#NandosDMgathering (25) #JobSeekersWednesday (22) #ANCGPManifestoLaunch (13) #OscarPistorius (3)	30.23%
9	126	#BangOut (46) Shoprite (3) #OscarPistorius (2) Soweto (2)	20.93%
10	124	#OscarPistorius (130) #NandosDMgathering (21) #ANCGPManifestoLaunch (3) #FeesMustFall (2)	13.95%
11	114	#OscarPistorius (26) #ANCGPManifestoLaunch (23) #NandosDMgathering (17) #AskAMan (15)	23.26%
12	102	#OscarPistorius (46) Ertugral (6) Soweto (4) #YouthDay (1)	13.95%

Table 2: Relationship between topics and communities. The table shows communities with more than 100 members. Column three shows the top mentioned topics in each community. The number of mentions is indicated in brackets. Column four shows the percentage of topics mentioned in each community out of all the topics used for downloading the data shown in Table 1.

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text,how different social media sources? *In Proceedings of the 6th International Joint Conference on Natural Language Processing*.

Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2013. Cascade-based community detection. *Sixth ACM international conference on Web search and data mining*, pages 33–42.

Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind egyptian political polarization on twitter. *ACM Conference on Computer Supported Cooperative Work and Social Computing*, (18):700–711.

Ralf Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. *16th international conference on text, speech and dialogue*.

Duanbing Chena, Mingsheng Shanga, Zehua Lvb, and Yan Fua. 2010. Detecting overlapping communities of weighted networks via a local algorithm. *Physica A: Statistical Mechanics and its Applications*, pages 4177–4187.

David Ediger, Karl Jiang, Jason Riedy, David A. Bader, and Courtney Corley. 2010. Massive social network analysis: Mining twitter for social good. *39th International Conference on Parallel Processing*, pages 583–593.

Linton C. Freeman. 1978. Centrality in social networks conceptual clarification. *Social Networks*, pages 215–239.

Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Outtweeting the twitterers - predicting information cascades in microblogs. *WOSN'10 Proceedings of the 3rd Wonference on Online social networks*, pages 3–3.

Michelle Girvan and Mark E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*, pages 7821–7826.

Moises Goldszmidt, Marc Najork, and Stelios Paparizos. 2013. Boot-strapping language identifiers for short colloquial postings. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Steve Gregory. 2010. Finding overlapping communities in networks by label propagation. *arXiv:0910.5516 [physics.soc-ph]*.

Nosakhere Griffin. 2015. Fees must fall and the possibility of a new african university. *www.Face2FaceAfrica.com*.

Lichan Hong and Gregorio Convertino. 2011. Language matters in twitter: A large scale study. *Fifth International AAAI Conference on Weblogs and Social Media*.

Mindaugas Idzelis. 2005. Jazzy: The java open source spell checker. *http://jazzy.sourceforge.net/*.

Paul Jaccard. 1902. Lois de distribution florale. *Bulletin de la Socet Vaudoise des Sciences Naturelles*, pages 67–130.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. *9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65.

Andrea Lancichinetti and Santo Fortunato. 2009. Community detection algorithms: a comparative analysis. *arXiv:0908.1062*.

Andrea Lancichinetti, Filippo Radicchi, Jose, Javier Ramasco, and Santo Fortunato. 2011. Finding statistically significant communities in networks. *PLoS One*, 6(4).

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. *NICTA VRL*.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2004. Automatic detection and language identification of multilingual documents. *Proceedings of the 2004 ACM Symposium on Applied. Computing*.

Martin Majlis. 2012. Yet another language identifie. *Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 46–54.

Yuexin Mao, Wei Wei, and Bing Wang. 2012. Correlating s&p 500 stocks with twitter data. *ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, (1):69–72.

Juan Martinez-Romo and Lourdes Araujo. 2013. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications: An International Journal*, 40:2992–3000.

Michael McCandless. 2010. ccuracy and performance of googles compact language detector. *http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html*.

Marcin Mikowski. 2010. Developing an open source, rule based proofreading tool. *Software: Practice and Experience*, 40:543–566.

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42.

Corrado Monti, Alessandro Rozza, Giovanni Zappella, Matteo Zignani, Adam Arvidsson, and Elanor Colleoni. 2013. Modelling political disaffection from twitter data. *International Workshop on Issues of Sentiment Discovery and Opinion Mining*, (2).

Shuyo Nakatani. 2010. Language detection library (slides). *http://www.slideshare.net/shuyo/language-detection-library-for-java*.

Shuyo Nakatani. 2012. Short text language detection with infinity-gram. *http://shuyo.wordpress.com/2012/05/17/short-text-language-detection-with-infinity-gram/*.

Jaimie Y. Park and Chin-Wan Chung. 2012. When daily deal services meet twitter: understanding twitter as a daily deal marketing platform. *Annual ACM Web Science Conference*, (4):233–242.

Frank Scheelen. 2003. libtextcat. *http://software.wise-guys.nl/libtextcat/*.

Sakaki Takeshi, Okazaki Makoto, and Matsuo Yutaka. 2010. Earthquakeshakes twitter users: Real-time event detection by social sensors. *International conference on World wide web*, (19):851–860.

Jierui Xie, Boleslaw K. Szymanski, and Xiaoming Liu. 2011. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *IEEE ICDM workshop on DMCCI*.

Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(43).

Chao Yang, Robert Chandler Harkreader, and Guofei Gu. 2011. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. *14th international conference on Recent Advances in Intrusion Detection*, 6961:318–337.

Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social Media Mining*. Cambridge University Press, NY, USA.