Exploiting Geo-tagged Tweets to Understand Localized Language Diversity

Amr Magdy[1], Thanaa M. Ghanem[2], Mashaal Musleh[3], Mohamed F. Mokbel[4]

[1,2,3,4]KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, KSA
[1,4]Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, MN
[2]Dept. of Information and Computer Sciences, Metropolitan State University, Saint Paul, MN
{[1]amr,[4]mokbel}@cs.umn.edu, [2]thanaa.ghanem@metrostate.edu,
[3]mmusleh@gistic.org

ABSTRACT

Social media services are the top-growing online communities in the last few years. Among those, Twitter becomes the *de facto* of microblogging services with millions of tweets posted everyday. In this paper, we present an analytical study for localized language usage and diversity in Twitter data using a half billion geotagged tweets. We first identify local Twitter communities on a country-level. For the identified communities, we examine (1) the language diversity, (2) the language dominance within the community and how this differs from local to global views, (3) demographics representativeness of tweets for real population demographics, and (4) the spatial distribution of different cultural groups within the countries. To this end, we group the tweets on two levels. First, we group tweets per country to identify the local communities. Second, we group tweets within each local community based on the tweet language. Our study shows useful insights about language usage on Twitter which provide important information for language-based applications on top of Twitter data, e.g., lingual analysis and disaster management. In addition, we present an interactive exploration tool for the spatial distribution of cultural groups, which provides a low-effort and high-precision localization of different cultural groups inside a certain country.

1. INTRODUCTION

Twitter is one of the most popular social networks where people used to tweet about their opinions, feelings, desires, on-going activities,...etc. Currently, Twitter receives 500+ million tweets that are posted by 255+ millions active users everyday [31]. With such plethora of incoming tweets, and other types of micro-messages, e.g., Facebook comments, Foursquare check-in's, social media data has entered the era where it can be used for new and different analysis tasks. For example, several techniques have been proposed to discover, track, and analyze local events based on Twitter data [1, 15, 17, 20, 23, 27, 34]. In addition, Twitter data analysis spans several areas including recency ranking [7], real-time recommendations [19], and modeling social interests and relations [12]. Social media analytics, then, became a big industry [26, 28] to the extent that major IT companies spend millions of dollars to incorporate such services [2, 30]. Combined with the widespread of smart mobile devices, social media providers are currently able to enrich their data with location information. On February 2014, 184 million users, around 76% of Twitter monthly active users, accessed Twitter from mobile devices [31]. As almost all smart mobile devices are GPS-equipped, Twitter has the ability to attach location information to most of its data, per user agreement. The same observation is used in releasing popular services like Facebook places and Foursquare check-in's. This generated a plethora of spatial information in social media which enables rich spatial analytics tasks on social media data.

With the popularity of Twitter service, a plethora of techniques in the literature have adapted language-based analysis approaches on tweets. This includes semantic and sentiment analysis [4, 5, 13, 18, 25], news extraction [21], recommendation [19], disaster management [33], entity linking [10, 16], and word-based analysis [24]. In most of these tasks, an implicit assumption has been made that English language is dominating other languages on Twitter to the extent that it could work as a language proxy for other languages [29]. However, some crucial applications, like disaster management, are highly dependent on the local language. For example, during China floods in 2012, propagating information about victims' locations on the Chiense Twitter (Sina Weibo) saved more than two hundred souls [6]. Thus, language-based applications on Twitter data need to be carefully aware of the language usage on the popular social network.

In this paper, we conduct a study to analyze and understand different aspects of spatial-language interaction in Twitter data. We analyze language data of recent tweets posted in the period of October 12, 2013 to March 6, 2014 worldwide. We closely focus on relating tweets' language to their spatial distribution to examine two aspects of language-spatial interactions in Twitter data: (1) the language diversity and usage in Twitter communities local-

§This work is supported by KACST GIS Technology Innovation Center at Umm Al-Qura University, under project GISTIC-13-06, and was done while the first, second, and fourth authors were visiting the center.

GeoRich '14 June 22-27 2014, Snowbird, UT, USA

http://dx.doi.org/10.1145/2619112.2619114 ...$15.00.

ized by country, and (2) the spatial distribution of cultural groups inside the country. This gives useful insights on the language usage on Twitter distinguished based on the spatial extent. First, this would give a pretty good idea on the difference of language usage in the global Twitter community and each of the local communities. In other words, we would identify the dominating languages as well as the distribution of other languages in each local community. Second, we study the language dominance in both global Twitter community and local communities deducing fruitful insights on language usage in Twitter data. Third, comparing different language diversity measures with the data collected by international organizations, e.g., UNESCO, would show if the virtual Twitter local community can be a representative sample for the actual population demographics. Fourth, analyzing the spatial distribution of languages inside countries enables a low-effort and high-precision localization of different cultural groups inside the country which is of interest for several users, e.g., administrative authorities in the country to deal with certain situation for a specific cultural group like Syrian refugees, new comers to multi-cultural countries who prefer to approach a similar culture community, or ethnicity-specific organizations that are interested to keep track of the spatial distribution changes of their ethnicity people.

To this end, we localize tweets based on country to identify what is called the local Twitter community in the country. For each local community, we study different statistical measures that show language diversity within the community. We show that statistical measures that take into account the language distribution within the community are more robust and consistent to show the countries with the highest language diversity. In addition, we develop new insights on language dominance within the different countries. We show that local Twitter communities are dominated by local language with probability of 65%. Also, we show that English language cannot be assumed a general language proxy ignoring the spatial distribution of the tweets. We, then, compare our collected statistics and measures with those that are published by UNESCO to assess the representativeness of Twitter data for population demographics.

To analyze cultural groups within the country, we group tweets within the country local community based on the tweet language at different levels of spatial granularity. For this, we build an adaptive pyramid structure that is able to power efficient querying for language data distributions at different spatial zoom levels. According to our analysis, tweets currently are posted in 55+ different languages, as well as in different dialects for the same language, from 206 countries worldwide. This reflects a rich cultural information that are embedded in Twitter data as well as a widespread cross cultures and space. To the best of our knowledge, there is no previous study that focuses on language usage and diversity aspects in Twitter data. Only one study [29] generally spotted different aspects of Twitter data based on the attached geographical information. However, this study hardly analyzed the interaction between language and space for only language dominance, for five weeks of Twitter data in year 2012, and reported a global domination of English language so that it can be considered as a global proxy language in tasks like geotagging. However, we show in our study over a longer period of time during 2013-2014 that the spatial extent matters in language usage on Twitter.

The rest of this paper is organized as follows. Section 2 introduces the dataset used in our study along with some background definitions. Section 3 shows our analysis for Twitter local communities in different countries. Section 4 presents our tool to localize and explore the different cultural groups within the country. Section 5 concludes the paper.

2. BACKGROUND AND DEFINITIONS

In this section we set a background for our study. First, we introduce basic definitions we are going to use throughout the rest of the paper. Then, we describe our Twitter dataset along with some statistics that show the relevance of studying language diversity in Twitter data. We also highlight some other data sources that we used in our study.

2.1 Definitions

In our study, we mainly work on two concepts: Twitter local community and cultural group. Twitter local community of a certain country is defined by the set of all tweets posted within the spatial extent of this country. Thus, when we analyze certain aspects of the local community, we basically analyze this set of tweets. The cultural group is simply defined as the group of tweets that are posted in the same language. Intuitively, language is a proxy for the cultural information. Thus, people who consistency tweet in the same language are considered to share some cultural background.

Throughout the paper, we will use Greenberg's language diversity index (LDI) [32] as one of the measures to assess cultural diversity. LDI gives the probability of randomly selecting two persons with different native languages from a certain group of people. Assume we have n language groups, LDI is given with the following equation:

$$LDI = 1 - \sum_{i=1}^{n} (\frac{c_i}{C_{total}})^2,\ C_{total} = \sum_{i=1}^{n} c_i$$

Where c_i is the number of people whose native language is i. Thus, a value of one represents total diversity where every individual has a different native language. The higher LDI value, the higher cultural diversity. LDI is used in UNESCO World Report on Cultural Diversity [32].

2.2 Data

In our study, we use 445+ millions geotagged tweets that are collected through Twitter public streaming APIs during the period of October 12, 2013 to March 6, 2014. We use Twitter filtering APIs where we set a spatial filter to the whole world region to get geotagged tweets from anywhere. The collected Tweets are geotagged on two levels: either (1) spatial region, e.g., landmark, city, or country, or (2) exact latitude/longitude coordinates. For the former category, extracting tweet country is just a parsing task. For the later category, which forms 15% of the dataset, we have built a spatial grid index, based on Simplified World Polygons data [11], that facilitates country name extraction based on a given latitude/longitude coordinates. For language data, we use the language attribute, that is attached to tweets, as it comes from Twitter.

To have rich and practical insights from our measured statistics from Twitter data, we compare those with statistics collected by official organizations and major geographical database providers. Specifically, we use datasets from ISO [1], UNESCO [2], and GeoNames [3] as baselines for our

[1]http://www.iso.org
[2]http://www.unesco.org
[3]http://www.geonames.org

Country	# of Languages to cover all tweets
USA	44
Japan, India	39
Germany, UK, Turkey, Indonesia	38
Spain, France, Brazil, Malaysia	37
Italy, Saudi Arabia	36

Table 1: Diversity by total # of languages

Country	# of Languages to cover 80% of tweets
Macedonia	9
Austria	8
AAT	7
NA, Armenia, Bulgaria, Burma, Germany, Switzerland, Cambodia	6
Morocco, Luxembourg, Georgia, Bangladesh, Hungary	5

Table 2: Diversity by # of languages to cover 80% of tweets

comparison. We use ISO 3166 [14] and GeoNames country information [9] datasets for getting country names and statistics on spoken languages. We also use UNESCO World Report on Cultural Diversity [32] for getting UNESCO values of Greenberg's language diversity index (LDI) for different countries. The details on the usage of the aforementioned datasets and the comparison with our measured statistics are presented in the following sections of the paper.

3. LOCALIZED LANGUAGE DIVERSITY

In this section, we present our study results on Twitter local communities in different countries. First, we present our results about language diversity and distribution within the countries. Then, we discuss language domination in Twitter local communities showing that English cannot be a global language proxy for Twitter data. Finally, we discuss the representativeness of Twitter data for the actual demographics of different countries.

3.1 Language Diversity

In our study, we identified 206 Twitter local communities, each is associated with exactly one country. Each community is divided into cultural groups. Our analysis shows that the whole dataset contains 55+ different languages with average of 18 languages used within individual communities and standard deviation of 12. Tables 1 - 3 show the language diversity in local Twitter communities, of the indicated countries, based on different measures. Table 1 shows the top-5 countries based on total number of languages that are used within the community. As shown, USA encounters tweets with 44 different languages, which is a relatively high number that indicates a high diversity. However, 85% of USA tweets are posted in English. Thus, the total number of languages is not the most indicative measure as it does not take into consideration the distribution of the languages. Table 2 [4] shows the top-5 countries based on number of languages that cover at least 80% of the tweets. As the reader can notice, a completely different set of countries appear in this list, except Germany, which means that all countries in the first column cover most of their tweets with less than 9 languages. Actually, all of them cover the 80% of the tweets with only 1-3 languages. Table 3 shows the top-5 countries based on LDI value. We can find strong correlation between countries in Tables 2 and 3, which shows that distribution-based measures are more consistent than the total number of languages.

Country	LDI
Macedonia	0.884
AAT	0.865
NA	0.857
Austria, Armenia	0.832
Morocco	0.821

Table 3: Diversity by LDI

3.2 Language Domination

Our analysis shows that tweets of 133 countries (∼65% of the countries) are dominated by the first spoken language in the country. This clearly shows that language domination in local Twitter communities is mostly for local language rather than international languages like English. In fact, most of the countries that are dominated by English although it is not the first language, which are 41 countries, encounter low Twitter activity that represents only 13% of the tweet activity in the remaining 73 countries. This shows that English cannot work as a language proxy for other languages when the application is concerned with the spatial extent.

In general, it is widely known that English language dominates Twitter data [31]. This observation is confirmed in our analysis where 33.9% of the tweets are posted in English from different countries. However, considering the spatial distribution of languages shows that English tweets are dominating due to the high Twitter activity from USA and UK. In fact, 28.6% of the tweets, in different languages, are posted from USA and UK. Obeying the 65% probability of being dominated by the first spoken language, 87.5% of USA and UK are posted in English, which represents 25% of the whole dataset. Consequently, 75% of the English tweets worldwide are posted from USA and UK.

Table 4 shows the most frequent language in the whole dataset and Table 5 shows the countries with the highest tweet activity and not dominated by their first language. It is worth noting that tweets of the seven languages in the first column of Table 4 form 81.6% of the whole dataset while the rest of 48 languages form only 18.4%. This confirms the observation that global language dominance exist in Twitter global community which does not contradict with the dominance of local languages in local communities.

3.3 Demographics Representativeness

Twitter is so popular that around 15% of the whole human population are registered users on Twitter and a quarter of

[4] AAT: Australian Antarctic Territory, NA: Netherlands Antilles

Language	%
English	33.9
Indonesian	17.2
Spanish	10.2
Portuguese	8.8
Turkish	4.6
Japanese	3.6
French	3.3

Table 4: Dominant Languages

Country	Dominant Language	% of Dominant Language
Malaysia	Indonesian	57.6
South Africa	English	77.8
Ukraine	Russian	77.4
Belarus	Russian	90.2
Switzerland	French	48.5
Lebanon	English	52.5
Pakistan	English	66.1

Table 5: Countries dominated by foreign language

those are active users. Thus, tweets are posted from so many people that it can be a representative sample for the human population. In 2012, a previous study [29] using five weeks of tweets has shown a Pearson correlation of 0.79 between the locations of tweets and the location where the electricity is available worldwide. They deduced that such high correlation makes Twitter data a valid baseline for evaluating the accuracy of geographic methods. In this part of our study, we consider one aspect of population demographics, which is the language diversity, and compare its outcomes from both Twitter data and from the real-world at different levels of spatial granularity. By this, we try to assess the validity of using Twitter population as a representative for actual population either worldwide or per country.

For our identified 206 local Twitter communities, we calculated the LDI index which represents the language diversity in certain population. In addition, we extracted LDI values for the same countries from UNESCO World Report on Cultural Diversity [32] as a baseline for comparison. Worldwide, we found a weak Pearson correlation of 0.25 between LDI values that are calculated from Twitter data and those that are reported by the UNESCO. However, we identified 33 countries (∼16% of the countries) that having less than or equal to 10% difference in the LDI value between the local Twitter community value and the UNESCO value. This brings the attention again for focusing on local aspects of Twitter data. Although the global Twitter community does not look representative for the human population, certain local Twitter communities look representative to its actual population. Table 6 shows countries with the least difference in LDI values, which are the most promising candidates for more investigation on the demographics representativeness of their Twitter local communities. Table 7 shows the number of countries that encounters a certain difference in LDI values. For example, there are 16 countries with difference in LDI values less than or equal 5%.

Country	% of LDI Difference
Samoa	0.2
Qatar, Italy	0.4
Bosnia and Herzegovina	1
Yemen	1.7
Ireland	1.8

Table 6: Countries with least LDI difference

% of LDI Difference	# of Countries
1	4
3	10
5	16
7	22
10	33

Table 7: # of Countries with LDI difference

4. LOCAL CULTURAL GROUPS

As the findings in Section 3 show, analyzing the interaction between language and spatial attributes of tweets could give fruitful insights on different aspects that are related to language usage and cultural diversity. In this section, we present a tool that enables visual analysis for language spatial distribution within a certain country. Using this tool, one can visually identify the spread of local cultural groups within the country through a web-based interface. This may be of interest for different types of users, e.g., administrative authorities in the country to deal with certain situations for a specific cultural group like Syrian refugees, new comers to multi-cultural countries who prefer to approach a community with a similar culture, or ethnicity-specific organizations that are interested to keep track of the spatial distribution changes of their people of interest.

Our tool employs an adaptive pyramid structure [3] (similar to a partial quad tree [8]) that stores percentages of different languages in all areas of the country at different levels of spatial granularity. Building the pyramid structure goes through two phases: (1) Structuring phase, and (2) Computation phase. In the structuring phase, we determine the structure of the pyramid. It is first initialized by one root cell that covers the whole country space. Then, the root cell is divided into four quadrant disjoint cells, each covering a quarter of the country space, and the tweets are distributed over the cells based on their spatial locations. Any cell that has number of tweets larger than a parameter *Capacity* is divided further into four children cells. The process is repeated recursively for each cell until the leaf cell has tweets less than or equal to *Capacity*. When the structuring process is completed, the partial pyramid structure is then fed to the computation phase. In the computation phase, the language distribution in each pyramid cell, either leaf or non-leaf cell, is precomputed and stored. The language distribution is computed as percentages of different languages in this cell. For example, if a certain cell has 80 English tweets, 60 Spanish tweets, 40 Italian tweets, and 20 Arabic tweets, then the language distribution for this cell would have four pairs of *<English,40%>*, *<Spanish,30%>*, *<Italian,20%>*, and *<Arabic,10%>*. The pyramid is then

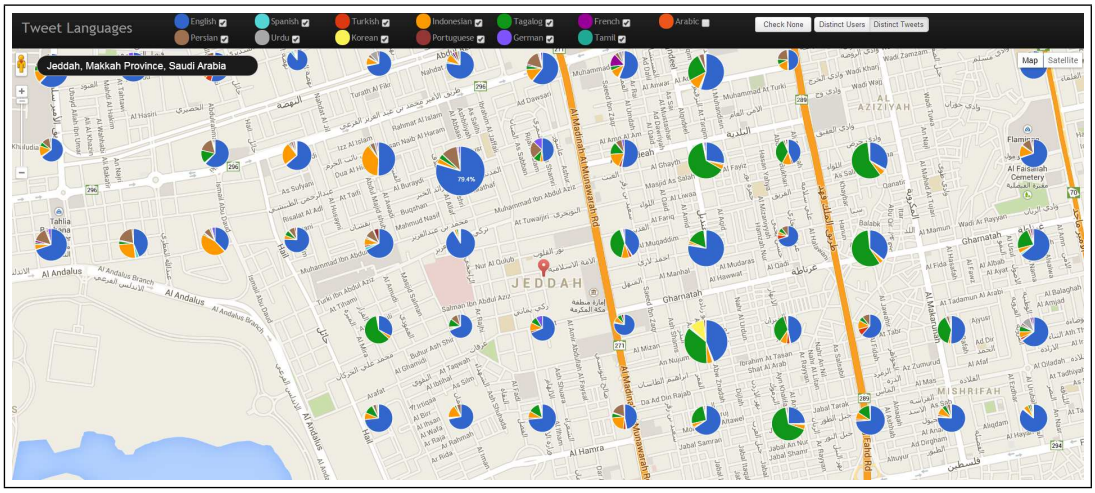

Figure 1: Language Distribution in Jeddah City, Saudi Arabia.

materialized to disk with each cell storing its language distribution. When the web-based interface is launched, the pyramid structure is loaded from the disk with its precomputed language distributions for fast query, navigation, and visualization.

Our tool takes a query input of a spatial region and a zoom level. The pyramid structure is then navigated to the appropriate level (based on the input zoom level) where the language distributions in all sub-regions of the spatial scope are retrieved and visualized using Pie charts on Google Maps.

We apply our tool to Saudi Arabia tweets, which has the highest Twitter penetration all over the world [22], and set *Capacity* parameter to 50 which enables to show the language distribution on the district-level all over the country. Figure 1 shows the language distributions in districts of Jeddah city in Saudi Arabia. In this figure, Arabic tweets, the first language in Saudi Arabia, are optionally excluded as it dominates everything else. As the figure shows, one can visually identify places where certain languages are popular. Other than English, the figure shows that specific districts where Indonesian, Tagalog, and Persian languages are popular. At a coarser level of granularity, one can see popular languages in city-level instead of district-level. The reader can check the interactive visualization tool on http://www.gistic.org/TwitterLanguages.

The presented tool, along with its web-based interface, facilitates a low-effort and high-precision localization for different cultural groups around the country. Such application, specifically, is of special interest to the administrative authorities in certain countries. For example, several countries currently have problems with refugees or groups of people who violates the immigration/work systems. Usually, refugees have a common cultural background which makes them likely to speak a common language. Localizing them through Twitter data is not costly, yet, expected to be highly effective due to the high accuracy of language detection on Twitter.

The presented tool can be used for other types of spatial analysis for categorical data attributes. For example, if one extracted the mobile device brand, e.g., iPhone, Samsung phone, or Nokia phones, from tweets that are posted from mobile devices, the same tool can be used to visually explore the penetration of each brand over the space at different levels of granularity. Then, this could be a useful low-effort analysis for endless number of applications.

5. CONCLUSIONS AND DISCUSSION

In this paper, we exploit geotagged tweets to understand localized patterns in language usage and diversity in different countries. The main component of our study is analyzing the distribution of languages over space. First, we have studied the distribution of languages on country-level where the tweets are posted from 206 countries in 55+ different languages. Using the country-level distribution, we have collected statistics that help to understand three aspects that are related to language: (1) Language diversity, (2) Language dominance, and (3) Demographics representativeness. Language diversity on country-level is an interesting aspect about demographics of different populations. International organizations like UNESCO consider language diversity among its measures to report while discussing the cultural diversity worldwide. In our study, we considered three measures of language diversity: (a) Total number of languages that are posted from the country, (b) Number of languages that covers 80% of the tweets, and (c) LDI index. It has been shown that the total number of languages is not very indicative as it does not consider the distribution of the languages inside the country. Thus, although USA encounters the highest total number of languages (44 languages), 85% of its tweets are posted in English and hence cannot be considered the highest language diversity among the countries. On the contrary, the second and third measures have shown a strong correlation as both of them take the language distribution into account. For language dominance, only seven languages, leaded by English, have shown to cover 81% of the whole tweets. However, considering the spatial extent, per country, shows that 65% of local communities are dominated by their first spoken language. Even countries that are not dominated by their first language encounter low tweet activity in English and usually are dominated by the neighborhood language which usually is related to cultural and historical bonds. Thus, English cannot be treated as global language proxy. Instead, the spatial extent should be considered while dealing with language related stuff in Twitter data. Comparing our LDI measured index with the

values reported by UNESCO, we found a weak statistical correlation of 0.25 for all the countries. However, for certain countries, the difference between the measured value and UNESCO value is negligible so that it could be considered very representative. We found 33 countries that have less than 10% difference, which are the top candidates to have a strong demographics representativeness from Twitter data.

Second, we presented a visual analysis tool that explores the spatial distribution of languages within a certain country. To this end, we depend on an adaptive pyramid structure that materializes the language distribution of all subregions within the country at different levels of spatial granularity. The language distribution is represented as relative percentages of the different languages within the cell. Setting the capacity of each pyramid cell controls the granularity of the navigation from city-level to even district-level. The pyramid structure is constructed once for the country and then stored to the disk. Then, when the application runs, it is loaded with the precomputed language distribution for fast querying, navigation, and interactive visualization.

Our study shows that Twitter data is rich with cultural and demographics information. The plethora of languages that are used indicates a wide interest in social media services from all cultures from all over the world. In addition, the inconsistency of the demographics outcomes from Twitter data and the real-world data indicates a gap between the virtual and real worlds which can be defined as "the gap of easy Internet connectivity". In other words, people who do not have ease access to the Internet are not expected to contribute to social media websites as it would not look like a necessity for them. One example for that is the Bengali people who live in Saudi Arabia. It is widely notice that this cultural group encounters problems in Internet connectivity. Thus, although there is a large number of individuals who can be easily noticed by everyone living in Saudi Arabia, one cannot find a considerable portion of tweets posted in Bengali so that it shows up on the map of Twitter languages in the country. This somehow explains the gap between Twitter demographics and actual population demographics. In fact, this envisions that Twitter data cannot be representative for the actual human population unless there is an easy Internet access to everyone on the planet, which does not look like a short-term goal.

6. REFERENCES

[1] H. Abdelhaq, C. Sengstock, and M. Gertz. EvenTweet: Online Localized Event Detection from Twitter. In *VLDB*, 2013.

[2] Apple buys social media analytics firm Topsy Labs. http://www.bbc.co.uk/news/business-25195534, 2013.

[3] W. G. Aref and H. Samet. Efficient Processing of Window Queries in the Pyramid Data Structure. In *PODS*, 1990.

[4] A. Bermingham and A. F. Smeaton. Classifying Sentiment in Microblogs: Is Brevity an Advantage? In *CIKM*, 2010.

[5] A. Bermingham and A. F. Smeaton. An Evaluation of the Role of Sentiment in Second Screen Microblog Search Tasks. In *ICWSM*, 2012.

[6] Sina Weibo, ChinaŠs Twitter, comes to rescue amid flooding in Beijing. http://thenextweb.com/asia/2012/07/23/sina-weibo-chinas-twitter-comes-to-rescue-amid-flooding-in-beijing/, 2012.

[7] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: Improving recency ranking using twitter data. In *WWW*, 2010.

[8] R. A. Finkel and J. L. Bentley. Quad Trees: A Data Structure for Retrieval on Composite Keys. *ACTA*, 4(1), 1974.

[9] GeoNames Country Information Data. http://download.geonames.org/export/dump/countryInfo.txt, 2014.

[10] W. Guo, H. Li, H. Ji, and M. T. Diab. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In *ACL*, pages 239–249, 2013.

[11] Humanitarian Information Unit Data. https://hiu.state.gov/data/data.aspx, 2014.

[12] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering Geographical Topics In The Twitter Stream. In *WWW*, 2012.

[13] Y. Hu, F. Wang, and S. Kambhampati. Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2640–2646, 2013.

[14] ISO 3166 Data. http://datahub.io/dataset/iso-3166-1-alpha-2-country-codes, 2014.

[15] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *ICDE*, 2012.

[16] X. Liu, Y. Li, H. Wu, M. Zhou, F. Wei, and Y. Lu. Entity Linking for Tweets. In *ACL*, pages 1304–1311, 2013.

[17] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*, 2011.

[18] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM*, 2012.

[19] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *RecSys*, 2009.

[20] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW*, 2010.

[21] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. In *GIS*, 2009.

[22] Saudi Arabia Records Highest Twitter Penetration in the World. http://arabcrunch.com/2013/11/saudi-arabia-records-highest-twitter-penetration-in-the-world.html, 2013.

[23] V. K. Singh, M. Gao, and R. Jain. Situation Detection and Control using Spatio-temporal Analysis of Microblogs. In *WWW*, 2010.

[24] C. Tan, L. Lee, and B. Pang. The Effect of Wording on Message Propagation: Topic- and Author-Controlled Natural Experiments on Twitter. In *ACL*, 2014.

[25] D. Tang, F. Wei, N. Yang, M. Zhou, B. Qin, and T. Liu. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *ACL*, 2014.

[26] Topsy Pro Analytics: Find the insights that matter. http://topsy.com/, 2013.

[27] TweetTracker: track, analyze, and understand activity on Twitter. http://tweettracker.fulton.asu.edu/, 2013.

[28] Twitter Analytics Business. https://business.twitter.com/products/analytics, 2014.

[29] The Geography of Twitter: Mapping the Global Heartbeat. http://irevolution.net/2013/06/09/mapping-global-twitter-heartbeat/, 2012.

[30] New features on Twitter for Windows Phone 3.0. https://blog.twitter.com/2013/new-features-on-twitter-for-windows-phone-30, 2013.

[31] Twitter Statistics. http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/, 2013.

[32] UNESCO World Report on Investing in Cultural Diversity and Intercultural Dialogue. http://www.unesco.org/new/en/culture/resources/report/the-unesco-world-report-on-cultural-diversity/, 2009.

[33] I. Varga, M. Sano, K. Torisawa, C. Hashimoto, K. Ohtake, T. Kawai, J.-H. Oh, and S. D. Saeger. Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster. In *ACL*, pages 1619–1629, 2013.

[34] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *CIKM*, 2011.