Microblogging without Borders: Differences and Similarities

Ruth Garcia
Yahoo! Research, Barcelona, Spain
UPF, Barcelona, Spain
ruthgavi@yahoo-inc.com

Barbara Poblete
DCC, University of Chile
Yahoo! Research Latin America
bpoblete@yahoo-inc.com

Marcelo Mendoza
Yahoo! Research Latin America
mendozam@yahoo-inc.com

Alejandro Jaimes
Yahoo! Research, Barcelona, Spain
ajaimes@yahoo-inc.com

ABSTRACT

The fast increase in the ease of access to computing, coupled with the rapid growth of social media has provided the space and motivated people all over the world to publicly share many kinds of information, from general interest topics such as elections and fashion to private topics such as the user's mood. The widespread use of microblogging services such as Twitter, in particular, have led to vast amounts of data generated by users in many different countries. In spite of this, very little is known about the differences and similarities in the way that people in different countries use such microblogging services. In this paper, we describe the analysis of a large-scale collection of Twitter data. First, we collected more than 550 million tweets from over 76 million users during August 20-29, 2010. Then, we identified the 10 countries with the highest volume of tweets during that period, and finally, selected the users from that period for those 10 countries, and collected all of their tweets for an entire year. Our analysis is based on over 5 billion tweets for 4.7 million anonymous users. We highlighting differences and similarities among these 10 countries with respect to language use, sentiment, and content of tweets.

Categories and Subject Descriptors

H.1.2 [**UserMachine Systems**]: Human factors

Keywords

Twitter, sentiment, cultural differences.

1. INTRODUCTION

Social media platforms such as Twitter have gained popularity in many countries, to the point in which the number of users and volume of activity are both high enough that the data posted by users reflects what is happening in the "real world." Twitter has been used to organize protests in Iran, Egypt, and Tunisia, and massively in reacting to events such as the Chilean and Japanese earthquakes of 2010 and 2011, respectively. But Twitter is also used on a daily basis for mundane purposes by people all over the world, and gaining insights into its use could have important implications in our understanding of different cultures, and have potential implications for development and in the creation of culture-specific services, among others. Eagle et al. [5], for example, investigated the relation between the structure of social networks and access to socioeconomic opportunity and suggest that the diversity of individuals' relationships is strongly correlated with the economic development of communities.

Twitter is an interesting platform, because in contrast to other social networks (e.g., Facebook) the majority of users have public profiles and connections are not necessarily reciprocal: users can follow other users without knowing them personally. Kwak et al. [7] concluded that only 22% of all connections in Twitter are reciprocal. One possible implication of this is that information quickly spreads beyond each user's social circle.

In this paper we report the first part of our ongoing research on cultural differences in social media. Aiming to analyze cultural differences and similarities among countries, we present the results of analyzing over 5 billion tweets for 4.7 million anonymous users in 10 countries. In the following sections, first we briefly explain the data set used and then we address the following questions: a) What are the most popular languages and what is the distribution per country? b) How does "happiness" vary across countries and languages? c) Are there differences in the content of tweets (use of hashtags, links, retweets) between these countries?

1.1 Related Work

Many researchers have analyzed Twitter data, but the analysis is typically performed without considering differences between countries. In terms of culture, one of the most well known studies was done by Hofsted [6], who did extensive surveys in 70 countries and created a framework of dimensions of culture. Based on his analysis he rated cultures based on different dimensions (e.g., individuality, etc.). Work on sentiment analysis on Twitter has been carried out by Doods et al. [3, 4] and by by Bollen et al. [1], among others. Doods et al. [3, 4] measure happiness on blogs and Twitter, and present a detailed study of happiness levels over time using the ANEW lists [2]. Bollen et al.[1] study sentiment in Twitter and showed happiness assortativeness beyond demographic features such as age, sex and race, and conclude that even psychological states such as "loneliness" can be assortative in a social network.

2. DATASET AND ANALYSIS

2.1 Dataset

First we collected all of the tweets that had been publicly posted between August 20 and 29, 2010. This resulted in 550,920,518 tweets from 76,255,339 users. Then we discarded tweets that contained GPS coordinates (which are automatically included in tweets when they are posted from mobile devices) because often those coordinates do not map to real locations. The accuracy of the GPS coordinates varies depending on many factors including whether the mobile device is used indoors or outdoors. The remaining tweets contained a textual field describing location in natural language, which is inherited from the user's profile. In order to validate those locations, we processed them to identify the country [1] and discarded any tweets that did not map to real locations. After such filtering we obtained 229,955,800 tweets generated by 6,263,457 users from 246 countries. Using this information we computed the number of users per country and selected all users from the ten day period for each of the 10 most active countries, yielding 4,736,629 selected users (76% of the total users for the ten day period). For the rest of the study we collected all of the tweets (5,270,609,213) for all 4,736,629 selected users over a period of one year (2010). All processing was anonymous and user ID information was discarded.

Figure 1 gives an overview of the distribution of the number of users over the ten day period, as well as of the level of activity per country over the entire year. Although the U.S. has the most users and most activity, we show later that it is not the country with the highest average $\frac{Tweets}{User}$ ratio.

2.2 Languages used in Countries

We removed URLs and non-alphanumeric characters and used proprietary software to classify the language of each of the 5,270,609,213 tweets. As a result, 69 languages were identified in 99.05% of the tweets. The 10 most popular languages are shown in Figure 2. As expected English is the most used language and corresponds to nearly 53% of all tweets.

[1]Location was found using proprietary Yahoo! software which maps textual descriptions (e.g., New York, NYC) to longitude and latitude coordinates.

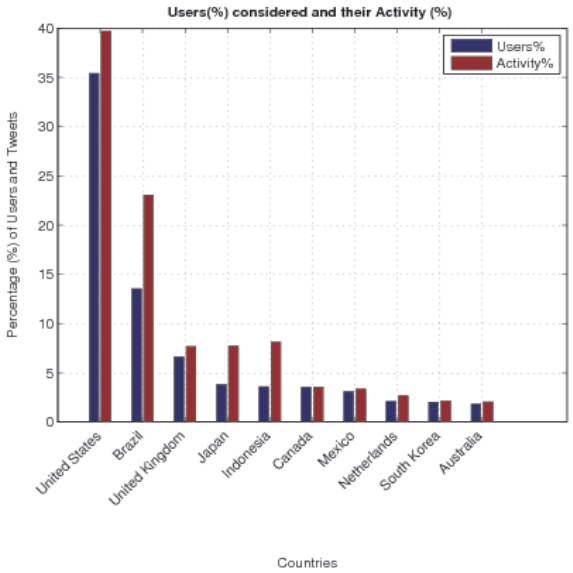

Figure 1: Ten countries with the most number of tweets during August 20-29, 2010 and their activity in 2010. The percentage of users per country is computed with respect to the number of users in the 10 day period, but the percentage of activity is computed with respect to the final one year data set (2010).

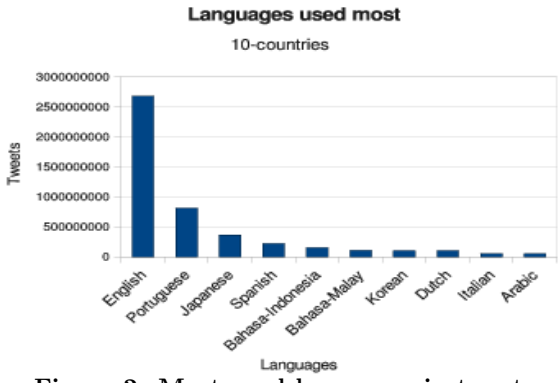

Figure 2: Most used languages in tweets

Figure 3 shows the three most common languages per country as well as the percentage of tweets in each of those languages for each country. It is worth noting that English is one of the top three most used languages in all ten countries, and in The Netherlands, Indonesia, and Mexico more than 10% of the tweets are in English, while in Brazil it is 9%.

Italian appears in Figure 2 although Italy was not considered, and Catalan shows up in Figure 3 even though it is estimated that the number of people that speak it worldwide is significantly smaller than the other languages considered. The explanation for this is that some of the tweets classified as Catalan or Italian were in Portuguese or Spanish. For example, the tweet "*Mexico no hay que llegar primero... si no que ahique saber llegar*" was labeled to be in Catalan although it is in Spanish and "*Um pequeno e valente guerreiro na luta contra o sono*" was classified as Italian although it is in Portuguese. The similarity of these languages, in addi-

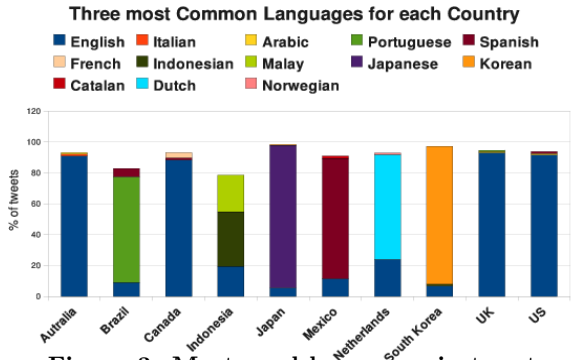

Figure 3: Most used languages in tweets

tion to the frequent presence of slang and misspellings makes automatic language identification particularly challenging.

2.3 Sentiment Analysis

We analyzed the differences in happiness [2] for the 10 selected countries, considering only tweets in *English* and *Spanish*. We used the 1999 Affective Norms for English Words (ANEW) list by Bradley and Lang [2] for English tweets and the corresponding Spanish adaptation of ANEW by Redondo et al. [8] for tweets in Spanish. The ANEW list contains 1,034 words and each word has a score in the range 1-9 which indicate its level of happiness. The scores for the individual words were obtained by asking participants in a study to rate them in that range (e.g., 9 for "completely happy", and 1 for "completely unhappy, annoyed, etc."). For example, the word *loved* has an average happiness value of 8.64 and its equivalent in Spanish (*amado*) has a value of 7.99.

We computed the "weighted average happiness level" per country based on the algorithms of Dodds et al. [4, 3], as follows:

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}(w_i) p_i \qquad (1)$$

where T represents all of the tweets per country during a particular time period for a specific language (Spanish or English), and f_i is the frequency of the ith of N distinct words for which there is an estimate of average happiness (i.e., those words that appear in the ANEW list). The normalized frequency is computed as follows:

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_j} \qquad (2)$$

The results of our sentiment analysis in English are shown in Figure 4, and coincide with those reported by Dodds et al. [4]: the values are between 5 and 7 for both languages and there is also a general increase in happiness towards the end of the year. It's interesting to note that Brazil has the highest values almost every month even though we're not specifically considering Portuguese, but after August the

[2]As in [4], we use the term happiness, but a more standard term is valence, a value that represents the psychological reaction to a specific word within a "happy-unhappy" scale.

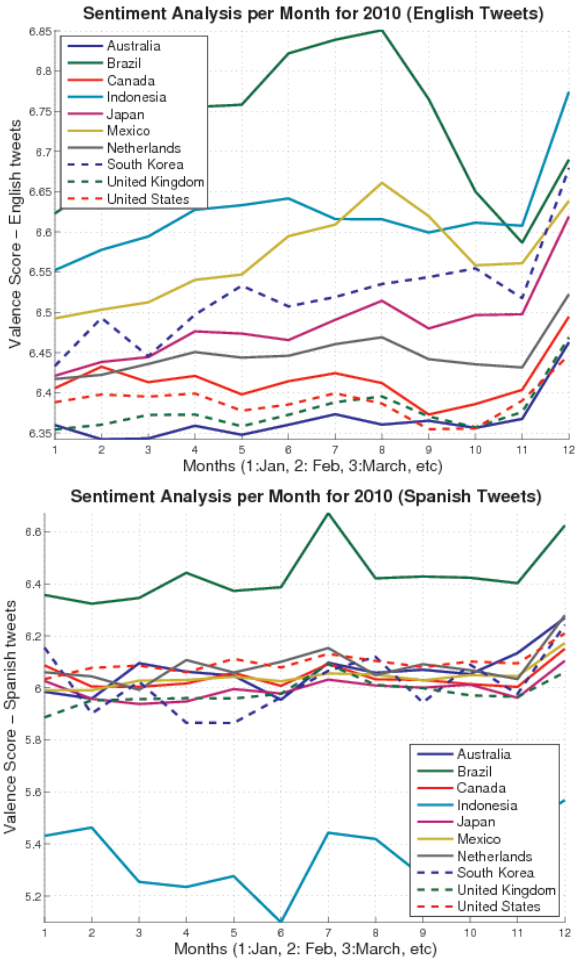

Figure 4: Average Happiness Level per Month for each country - English and Spanish tweets

happiness level in Brazil decreases until November. In December all countries present an increase in their happiness level. Indonesia has a higher increase in this month with scores that are even higher than Brazil. South Korea also presents a strong increase this month, almost scoring the same as Brazil.

Some differences can be seen in the results for tweets in Spanish, also in Figure 4. The number of tweets in Spanish is disproportional as 7 countries account for less than 1% of the tweets, while Mexico, USA and Brazil together account for around 98% of the total. Nevertheless, USA and Mexico have happiness patterns that are similar to the majority of the countries. Only Brazil and Indonesia present interesting results that differ from the other countries: there is a strong increase in happiness from June to July for Brazil and Indonesia. Interesting drops happen in Indonesia during the months of May and August. Brazil has clearly the highest values for all months, but it also presents higher ups and downs in some months.

Country	$\frac{Tweets}{Users}$	$(URL)\%$	$(\#)\%$	$(@)\%$	$(RT)\%$
Indonesia	1813.53	14.95	7.63	58.24	9.71
Japan	1617.35	16.30	6.81	39.14	5.65
Brazil	1370.27	19.23	13.41	45.57	12.80
Netherlands	1026.44	24.40	18.24	42.33	9.12
UK	930.58	27.11	13.03	45.61	11.65
US	900.79	32.64	14.32	40.03	11.78
Australia	897.41	31.37	14.89	43.27	11.73
Mexico	865.7	17.49	12.38	49.79	12.61
S. Korea	853.92	19.67	5.83	58.02	9.02
Canada	806	31.09	14.68	42.50	12.50

Table 1: Average usage percentage per user of symbols in tweets

2.4 Tweet Contents

We analyzed the contents of the tweets for each each country, in particular:

- #: symbols that are used to give the tweet a topic.
- RT : RT is a keyword to forward another tweet and it is generally followed by the name of a user.
- @ : symbol used preceding the user name to mention another user.
- URL : A link to a website to share information.

We computed the averages per user in each country as follows:

$$AVG(symbol) = \frac{\sum_{i=1}^{N} \frac{T(symbol)_{u_i}}{T_{U_i}}}{\sum_{i=1}^{N} U_i} \quad (3)$$

Where AVG(*symbol*) means the average of tweets per user of a particular country containing the *symbol* studied (hashtag, url, mentions, etc). N is the total number of users for a particular country and $T(symbol)_{U_i}$ is the total number of Tweets containing that symbol for user U_i

Table 1 presents these averages per country as well as the ratio $\frac{tweets}{user}$. In our analysis, the appearance of a *symbol* per tweet was counted only once, that is, if a user used two hashtags in one tweet we counted it as one. The countries are ordered according to the ratio $\frac{Tweets}{User}$. The results show that Indonesia ranks first in tweets per user, followed by Japan and Brazil. It is interesting also to see that Indonesia and South Korea have the highest percentage of mentions in contrast to Japan that has the lowest, and it seems also to be the country with the fewest re-tweets in our data set. The Netherlands is the country with the most hashtags per user, while the US seems to be the country with the most links per user.

3. CONCLUSIONS

In this study we have analyzed several aspects of a large-scale Twitter data set. We considered the 10 most active countries during a period of one week and then collected all of the tweets of 2010 from the users belonging to these countries. We studied language use, happiness levels of tweets and content. We studied the most common languages used in each county, from which English is the most common language followed by Japanese and Spanish. We found that the happiness levels for each country differed for English and Spanish (e.g., Indonesia ranked among the first in *happiness* for English tweets but it ranked lowest for Spanish tweets).

With respect to the content of tweets, we found that Indonesia has the highest percentage of tweets with mentions per user while Japan has the lowest percentage of mentions and retweets for 2010. The Netherlands has the highest percentage of tweets with hashtags. In the future, we will analyze correlations between the context and content of the tweets by considering the structure of the network formed by friends and followers, analyze topics of tweets, content of the urls, and identify *cultural clusters*.

3.1 Acknowledgements

This research is partially supported by European Community's Seventh Framework Programme FP7/2007-2013 under the ARCOMEM project and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 (www.cenitsocialmedia.es), "Social Media."

4. REFERENCES

[1] J. Bollen, B. Goncalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Computing Research Repository*, abs/1103.0784, 2011.

[2] M. M. Bradley and P. J. Lang. Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings. In *In Technical Report C-1, The Center for Research in Psychophysiology*, Gainesville, Florida, 1999.

[3] P. Dodds and C. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, August 2010.

[4] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *Computing Research Repository abs/1101.5120v3[physics.soc-ph]*, Feb. 2011.

[5] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[6] G. H. Hofstede. *Culture's consequences : comparing values, behaviors, institutions, and organizations across nations*. Sage Publications, Thousand Oaks, California, 2nd ed edition, 2001.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on WWW, Raleigh NC USA*, pages 591–600. ACM, April 2010.

[8] J. Redondo, I. Fraga, I. Padrón, and M. Comesaña. The spanish adaptation of anew (affective norms for english words). In *Volumne 39*, number 3, pages 600–605. Psychonomic Society Publications, 2007.