Portuguese Geolocated Tweets: An Overview

Gaspar Brogueira
ISCTE-IUL / INESC-ID
Lisboa, Portugal
gmrba@iscte.pt

Fernando Batista
ISCTE-IUL / INESC-ID
Lisboa, Portugal
fmmb@iscte.pt

Joao P. Carvalho
INESC-ID / IST,
Universidade de Lisboa,
Portugal
joao.carvalho@inesc-id.pt

Helena Moniz
INESC-ID / FLUL - CLUL,
Lisboa, Portugal
helena.moniz@inesc-id.pt

ABSTRACT

This paper describes an existing database of geolocated tweets that were produced in Portuguese regions. The existing database was collected during eight consecutive days and contains about 307K tweets, produced by about 11K different users.

A detailed analysis on the content of the messages suggests a predominance of teenagers and young adult authors that use Twitter as a way to communicate their feelings, ideas and comments to their colleagues. An overview of the dataset suggests that tweets have a very personal content, often describing family bonds and school activities and concerns. This is a suitable source of information for a number of tasks, including sociolinguistic studies, sentiment analysis, among others.

Categories and Subject Descriptors

H.3.1 Content Analysis and Indexing;

Keywords

Twitter, Portuguese tweets, Twitter APIs, Data analysis

1. INTRODUCTION

Twitter is one of the most widely used and well-known social networks worldwide. It provides rapid communication and experience sharing among its users by providing an infrastructure for sending and receiving messages, containing 140 characters at most. According to [1, 2], about 646 million users produce approximately 400 million tweets every day.

Access to the Twitter data is facilitated for the scientific community by a number of APIs (Application Programming Interfaces). The Streaming APIs offer a “low latency access to Twitter's global stream of Tweet data". The REST API also offers a vast number of resources for fetching of tweets, based on timelines, searches, users, and other lists. Most of the APIs impose restrictions to the amount of data that can be retrieved. For example, the statuses/sample API returns a random sample of about only 1% of all public tweets produced everyday [3]. The statuses/filter API allows to collected information based on query filters and permits to specify, for example, geographic criteria.

This paper describes a database of tweets that was collected over eight consecutive days and is restricted to geolocated tweets produced in Portuguese regions and written in Portuguese. The data was retrieved using the statuses/filter API, by specifying geographic coordinates for covering the mainland and also the Portuguese archipelagos Azores and Madeira. From a preliminary inspection to our 8-day sample we can say that the data is fairly spontaneous, obviously coded in a written form, produced by young people with very personal content, often describing family bonds and school activities and concerns. Building on this assumption, tweets may be eventually used for training spontaneous models for Automatic Speech Recognition (ASR). Therefore, future work will tackle the use of tweets to train language models and to evaluate such models in different spontaneous speech domains. The database can be used in several perspectives, including: geolocated analysis of users and content; characterization of the different Portuguese regions; age-identification and characterization; sociolinguistic studies; sentiment analysis; among others.

2. RELATED WORK

Previous studies over twitter data are commonly found in the literature. However, those concerning Portuguese tweets are rather scarce, and sometimes limited to smaller databases containing only few thousand tweets.

A considerable number of research work report the use of Portuguese language Tweets for Sentiment Analysis [4][5]. [4] uses a database of 1700 tweets to evaluate the impact of different preprocessing techniques and negation modeling in the tweet sentiment classification. [5] also focuses on Sentiment Analysis, adapting state of the art approaches to Portuguese language. The author uses a collection of 300 thousand tweets, filtered according to the presence of certain verbs, such as sentir"/feel.

Portuguese twitter data was also used by [6] to predict Flu Incidence. In this recent study, the authors use about 14 million tweets originated in Portugal together with a search engine query logs to estimate the incidence rate of influenza like illness in Portugal. Portuguese tweets are also currently being used for Machine Translation tasks. For example, [7] provides a link to databases of parallel corpora that also include Portuguese language.

3. DATA ANALYSIS

The data here analyzed was collected between March 14th and 21th 2014, totaling about 307K tweets, produced by 11391 distinct users. The set of tweets correspond to a daily average of about 48K tweets. The number of tweets varies considerable per user, ranging from 1 to around 1100 tweets in the 8-day period. Figure 1 shows the user activity by hour, revealing that the most active period starts at 17h and goes until midnight. In fact, about 58% of the data is produced during that period. Figure 2 reveals that activity per hour is also not exactly the same all over the week. The number of tweets during weekend evenings is lower than during the remainder of the evenings, which might indicate that Twitter usage is mainly domestic during the evenings, *i.e.*, users do not usually tweet as much when going out with friends.

Conference'10, Month 1–2, 2010, City, State, Country.

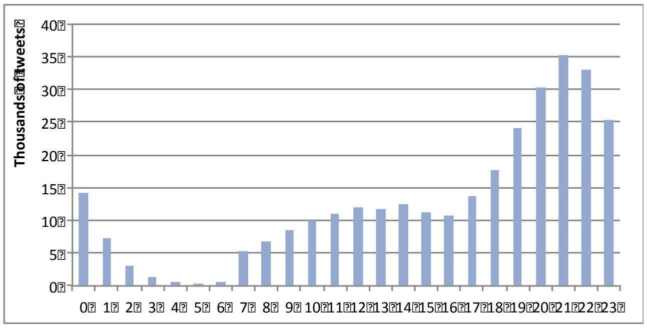

Figure 1. Average number of tweets produced per hour.

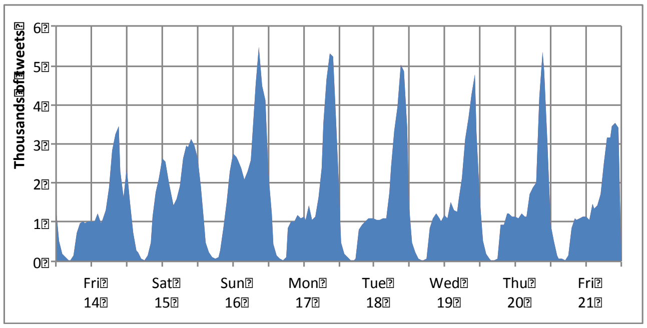

Figure 2. Number of tweets produced by hour during the week.

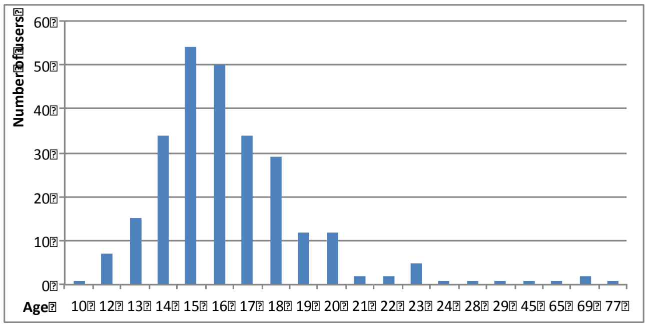

Figure 3. Distribution of the users age.

The content we have observed suggests that tweets were mostly produced by teenagers. So, we have made an attempt to characterize the involved community in terms of age, which is not a trivial task because that information is not clearly provided within the tweet content. We have found that the user description associated to each user sometimes contains that information embedded in the text. In order to have an idea of the user age we have manually tagged about 265 users, based on their description. The resultant information is depicted in Figure 3, clearly revealing the predominance of young authors, namely teenagers and young adults.

Also concerning the tweet text content, ngrams frequency is a key-factor for the understanding of the lexical selection used by tweeters. We can say that Portuguese tweets are mostly focused on personal messages based on family bonds, as illustrated in the selection of words from the same semantic field – “mãe”/mother; “pai”/father; “irmão”/brother; “irmã”/sister. Moreover, there is also a semantic field associated with school, encompassing vocabulary such as “teste”/test; “a minha turma”/my class; “aula”/lesson, suggesting a strong activity of teenagers/young adults. Another lexical cue to the characterization of the personal component of the tweets is the use of first person pronouns (subject “eu”/I; object “me”/me; and possessives “minha”/my; “meu”/my).

In line with the personal trait of tweets is the use of emoticons. Even though some authors claim that emoticons are used mainly by teenagers and young adults [8] we feel that nowadays emoticons’ use is widespread among age groups. The top 10 of most frequent emoticons found in the dataset are: :), :c, :(, :3, :-), :o, ;), :D, (@, :p. The set of emoticons is similar to the ones reported by [9] for English tweets.

4. CONCLUSIONS AND FUTURE WORK

The information produced by a community through a social network provides means to characterize such community over a vast number of perspectives. The interactions between the community members provide information that until now were very difficult to discover.

On Twitter, the interaction between users is carried out by small messages that can be used to express everything, from personal feelings to serious news. We have described database of collected geolocated tweets produced in Portugal. Our current database, containing tweets from 8 consecutive days aggregates about 300 thousand messages written in Portuguese and produced in Portugal. We have characterized the collected data and we think it is a valuable resource for studying part of the Portuguese community that is now using social networks.

5. ACKNOWLEDGMENTS

This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia under project PTDC/IVC-ESCT/4919/2012 and Grant SFRH/BPD/95849/2013.

6. REFERENCES

[1] S. Kumar, F. Morstatter, H. Liu. 2013. Twitter Data Analytics. Springer, New York, NY, USA.

[2] Brain, Statistic. 2014. Twitter Statistics, http://www.statisticbrain.com/twitter-statistics/.

[3] Twitter. 2013. Documentation, https://dev.twitter.com/docs/.

[4] M. Souza, R. Vieira. 2012. Sentiment analysis on twitter data for Portuguese language. Computational Processing of the Portuguese Language. Lecture Notes in Computer Science. Springer Berlin Heidelberg, vol. 7243, pp. 241-247.

[5] T. D. S. Cunha. 2013. Sentiment analysis on twitter's Portuguese language. Faculdade de Engenharia da Universidade do Porto.

[6] J. C. Santos, S. Matos. 2013. Predicting incidence from Portuguese tweets. IWBBIO, pp. 11-18.

[7] W. Ling, G. Xiang, C. Dyer, A. Black, I. Trancoso. 2013. Microblogs as parallel corpora. Proceedings of the 51st Annual Meeting on Association for Computational Linguistics.

[8] A. Brito. 2008. O discurso da afetividade e a linguagem dos emoticons, Revista Eletrônica de Divulgação Científica em Língua Portuguesa, Linguística e Literatura, number 9.

[9] Tyler Schnoebelen. 2012. Do you smile with your nose? stylistic variation in twitter emoticons. In Working Papers in Linguistics, volume 18. University of Pennsylvania.