Expanding a Database of Portuguese Tweets

Gaspar Brogueira[1], Fernando Batista[1], João P. Carvalho[2], and Helena Moniz[3]

1 **Laboratõrio de Sistemas de Língua Falada – INESC-ID, Lisboa, Portugal**
ISCTE-IUL – Instituto Universitário de Lisboa, Lisboa, Portugal
gmrba@iscte.pt,fmmb@iscte.pt

2 **Laboratõrio de Sistemas de Língua Falada – INESC-ID, Lisboa, Portugal**
Instituto Superior Tẽcnico (IST), Lisboa, Portugal
joao.carvalho@inesc-id.pt

3 **Laboratõrio de Sistemas de Língua Falada – INESC-ID, Lisboa, Portugal**
FLUL/CLUL, Universidade de Lisboa, Lisboa, Portugal
helena.moniz@inesc-id.pt

Abstract

This paper describes an existing database of geolocated tweets that were produced in Portuguese regions and proposes an approach to further expand it. The existing database covers eight consecutive days of collected tweets, totaling about 300 thousand tweets, produced by about 11 thousand different users. A detailed analysis on the content of the messages suggests a predominance of young authors that use Twitter as a way of reaching their colleagues with their feelings, ideas and comments. In order to further characterize this community of young people, we propose a method for retrieving additional tweets produced by the same set of authors already in the database. Our goal is to further extend the knowledge about each user of this community, making it possible to automatically characterize each user by the content he/she produces, cluster users and open other possibilities in the scope of social analysis.

1998 ACM Subject Classification H.3.1 Content Analysis and Indexing: Linguistic processing

Keywords and phrases Twitter, corpus of Portuguese tweets, Twitter API, natural language processing, text analysis

Digital Object Identifier 10.4230/OASIcs.SLATE.2014.275

1 Introduction

Twitter is one of the most widely used and well-known social networks worldwide. It allows for rapid communication and experience sharing among its users by providing an infrastructure for sending and receiving messages, containing 140 characters at most. About 646 million users produce approximately 400 million tweets everyday [12, 5]. Twitter is a source of information potentially useful for research in various fields, not only because of the amount of information produced, but also because the access to the data is facilitated for the scientific community by a number of APIs (Application Programming Interfaces).

This work aims at expanding a database of tweets that was collected over eight consecutive days, restricted to geolocated tweets produced in Portuguese regions, and written in Portuguese. The existing data was retrieved using the *statuses/filter* API that allows to fetch tweets with a low latency. The restriction to Portuguese regions was achieved by specifying geographic coordinates that define a number of rectangles covering the mainland and also the Portuguese archipelagos Azores and Madeira. The restriction to the Portuguese language was achieved by using the "*lang*" attribute that is automatically assigned by Twitter. An

3rd Symposium on Languages, Applications and Technologies (SLATE'14).
Editors: Maria João Varanda Pereira, Josẽ Paulo Leal, and Alberto Simões; pp. 275–282

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

overview of the dataset suggests that collected tweets are mostly produced by young people expressing very personal content, often describing family bonds and school activities and concerns. Furthermore, from a preliminary inspection to our 8-day sample we can also say that the data is fairly spontaneous, obviously coded in a written form.

This paper proposes a methodology that is now being used for expanding the database, thus allowing to further characterize the involved community in detail. It is our believe that the current database *per se* is an important resource to characterize part of the Portuguese Twitter community. 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; and among others. Due to their spontaneous nature, tweets may be eventually used for training spontaneous models for Automatic Speech Recognition (ASR) to cover the absence of models trained specifically with spontaneous data, since models are commonly trained on newspapers and broadcast news. Therefore, future work will tackle the use of tweets to train language models and to evaluate such models in different spontaneous speech domains. The extended version of the database will provide additional data that can be used for extending tasks, such as better assessment of age, twitter usage patterns over the time, vocabulary usage per author, amongst others.

This paper is organized as follows. Section 2 describes the related work in terms of twitter data collections, specially targeting the Portuguese language. Section 3 describes the existing corpus and presents a number of statistical elements concerning the tweet content. Section 4 proposes a methodology for extending the existing data and analyses the properties of the recently appended data. Finally, 5 presents the conclusions and overviews future trends.

2 Related Work

Previous studies involving Portuguese tweets are still scarcely found in the Literature. However, a considerable number of recent work have used Portuguese language Tweets in Sentiment Analysis related tasks [11, 3, 4]. [11] uses a database of 1700 tweets to evaluate the impact of different preprocessing techniques and negation modeling in the tweet sentiment classification. [3] 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 [9] 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, [6] provides a link to databases of parallel corpora that also include Portuguese language[1].

Finally, an architecture for automatic collecting tweets with a predefined delimited geographic region is proposed by [7]. Their architecture uses a MySQL database for tweets storage and a Twitter Streaming API for accessing and collecting an unlimited number of tweets.

3 Data Analysis

The data here analyzed corresponds to an 8-day period and was collected between March 14th and 21th 2014 [2]. The stream API *statuses/filter* was configured for retrieving geolocated

[1] `http://www.cs.cmu.edu/~lingwang/microtopia/`

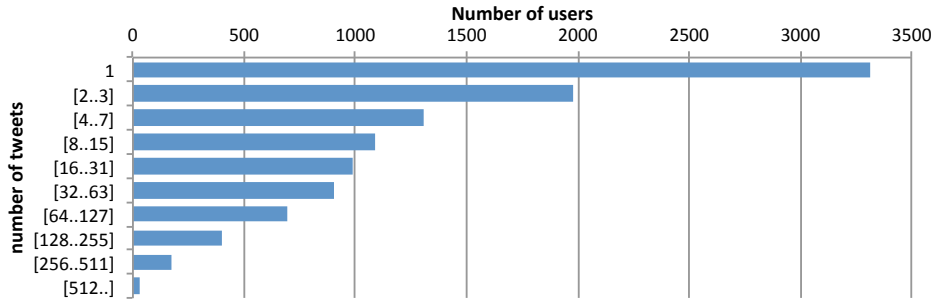

Figure 1 Number of users that have produced a certain quantity of tweets.

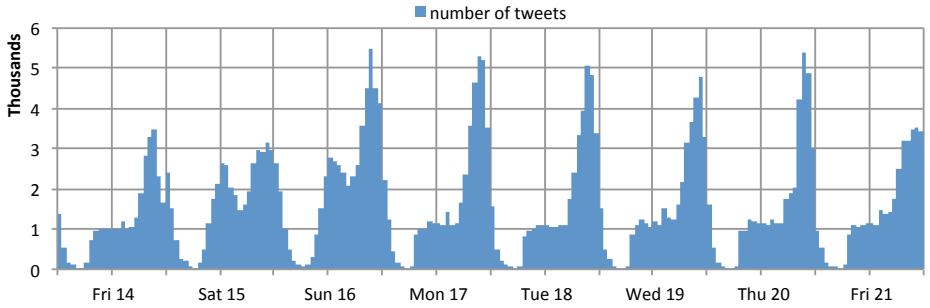

Figure 2 Distribution of tweets by hour.

tweets produced in Portugal. A total of 307K tweets were collected, corresponding to a daily average of about 48K tweets. The set of tweets being analyzed was produced by 11391 distinct users. Figure 1 represents the number of users that produced tweets in a given interval. Most of the tweets are, in fact, produced by distinct users. However, more than half of the users have produced more than one tweet during the period in analysis. The number of tweets varies considerable per user, ranging from 1 to around 1100 tweets in an 8-day period. This behavior justifies our approach described in Section 4 for extending our database, *i.e.*, a deeper knowledge of each user allows for a more suitable analysis of inter-users' traits.

The remainder of this section presents more detailed statistics concerning the number of produced tweets and their temporal distribution, and then focuses on characterizing the content of each tweet.

3.1 Temporal Distribution of the Collected Tweets

It is known that the number of produced tweets is not linear in time. Figure 2 presents the activity by hour, depicting that during friday evenings the number of tweets is lower than during the remainder of the evenings, suggesting that Twitter usage is mainly domestic during the evenings, *i.e.*, users do not usually tweet as much when going out with friends. In fact, users are less active during the day. However 43% of all the user activity is performed during that period which represents a considerable proportion. Taking that into account, we have made attempts to further characterize this community of users. The content we have observed suggested 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

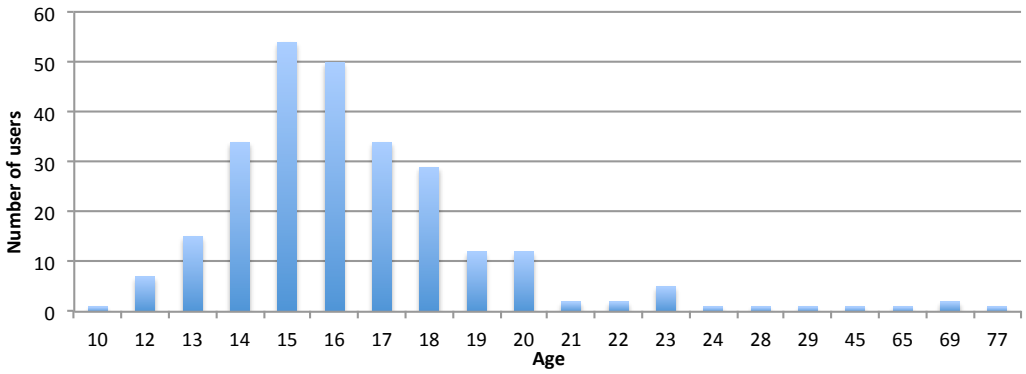

Figure 3 Distribution of the users age.

found that the user description associated to each user sometimes contains that information embedded in the text. Examples:

Paredes | 13 anos | Eminem
Ola tenho 14 anos e sõ sei dormir.
metro e meio de gente / 20 anos / ESTSP

In order to have an idea of the user age we have manually tagged about 265 users with their potential age, based on their own description. The resultant information is depicted in Figure 3, clearly revealing a predominance of teenagers and young adults, as expected by the previous analysis performed on the content. The extended version of the database described in Section 4 will provide additional data that can be used for better assessment of the age, twitter usage patterns over time, vocabulary usage per author, and age stylistic effects.

3.2 Content Analysis

One of the most interesting Twitter particularities is the 140 characters message length limit, leading to messages containing an expected low number of words. The data reveals a trend in the increase of the number of words per message along the day. The maximum value is attained around the hour of maximum Twitter activity.

Regarding tweets' content, Table 1 shows the most frequent trigrams from the 8-days in analysis. The trigrams' frequency is a key-factor for the understanding of the lexical selection

Table 1 Top trigrams.

Trigrams	Freq.
a minha mãe	1395
o meu pai	903
sei o que	746
o que ẽ	741
com a minha	704
tudo o que	634
toda a gente	621
com o meu	617
A minha mãe	601

Trigrams	Freq.
fim de semana	527
o que eu	459
todos os dias	450
a minha vida	444
que a minha	431
como ẽ que	428
porque ẽ que	424
que o meu	415
tenho de ir	412

Trigrams	Freq.
acho que vou	394
que ẽ que	386
a dizer que	382
dia do pai	376
ẽ que eu	373
ir para a	368
como Assunto do	353
e a minha	350
o meu irmão	345

Table 2 Top hashtags ordered by number of users that refer them.

Hashtag	users	Freq
#lt	631	1555
#lrt	616	1403
#np	529	1882
#twitteroff	238	439
#portugal	185	377
#carregabenfica	175	617
#lisbon	125	249
#lisboa	110	231
#love	102	132
#friends	94	132
#selfie	90	98
#nw	83	116

Hashtag	users	Freq
#sun	79	97
#somosporto	69	171
#me	61	103
#night	61	67
#porto	61	105
#beach	56	66
#happy	56	68
#valetudo	55	89
#sunset	54	57
#benfica	50	146
#excluidadasociedade	50	58
#sad	48	57

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 (not displayed in Table 1 for legibility issues), suggesting a strong activity of teenagers/young adults.

Another lexical clue 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 what regards verbal forms, either than being inflected in the first person, the selection of epistemic verbs ("sei"/know, "acho"/think) is a crucial indicator of the way speakers communicate their doubts and certainties, mostly associated with values of like/dislike. Clefts ("ẽ que"/it's ... that) are another very frequent structure selected by tweeters. These structures are used to focus particular constituents, as an emphatic structure.

From the above set of lexical-syntactic options of the tweeters we are able to characterize the 8-days sample tweets as very personal data, written in first person, communicating believes and emotions.

In line with the personal trait of tweets is the use of emoticons, which are pictorial representations associated with distinct emotions, allowing for the expression of feelings in e-contexts. Even though some authors claim that emoticons are used mainly by teenagers and young adults [1, 8] we feel that nowadays emoticons' use is widespread among age groups and no conclusions can be taken from this fact. Emoticons, such as :), :-), :3, ;), and :)) express joy, while emoticons such as :(, :$, :/, :((, :-(expressed sadness. The set of emoticons found in our database is similar to the ones reported by [10] for English tweets.

Only a small number of tweets include hashtags (about 4.3%). This is a rather low number when comparing to other Twitter data collections. [14] reports about 11% of tagged tweets for Portuguese, which includes not only European Portuguese but also Brazilian Portuguese. Moreover, Portuguese is one of the languages that uses fewer hashtags from the 8 languages analyzed. Finally, in a similar database of about 1.5 Million tweets written in Portuguese (including all varieties of Portuguese), collected during the same time period without restricting the location, and where most of the tweets are written in Brazilian Portuguese, such value corresponds to 10.2%.

The top most frequent hashtags are expressed in Table 2 and include #lt (Last or Latest Tweet), #lrt (Last or Latest retweet), #np (Now Playing) used whenever someone is listening

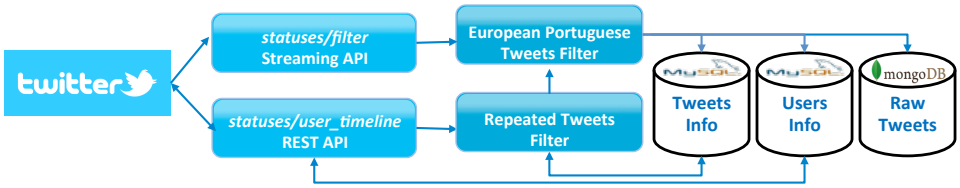

Figure 4 Diagram of the proposed data collection infra-structure.

to a song and wants to share it, and #twitteroff (Enough tweets for today). Nevertheless, we observed that the frequency of such hashtags was relatively low in the similar database mentioned in the previous paragraph of 1.5 Million tweets. #tl (position 69), #lrt (position 146), #np (position 109), #twitterof (position 643).

Another interesting fact concerns the number of retweets (RT), which is very low in our current database, corresponding to roughly 0.1% of the tweets. About 26% of the tweets from the 1.5Million tweets previously mentioned database are retweets. Two possibilities exist concerning this point. The first possibility is that geolocated tweets do not usually include retweets. Another hypothesis is related with the fact that tweets from our database are mostly personal are therefore not usually retweeted.

4 Database Expansion

Twitter APIs give free access to millions of tweets and can be classified into two broad categories, according to their design and access methods. The streaming API provides a continuous data flow of public tweets, where the data is continuously updated immediately after a request until an interrupt request. The REST API is based on the concept of *Representational State Transfer* APIs for web design permits to collect data for specific users. These two APIs accept different parameters and are suitable for quite different purposes. They are also equally restricted in terms of the type of data and the amount of data they can provide, so that a straight use of the available APIs is often not enough to obtain the required information. We propose a methodology that allows circumventing this restriction for collecting and storing Portuguese tweets

The proposed methodology is illustrated in Figure 4 and provides a way of collecting tweets produced in a given region and then expanding this initial set by collecting additional tweets, produced by the same set of users. It comprises two stages: i) collect and filter tweets containing geographic information, which are available through the streaming API statuses/filter. Tweets are filtered by several rectangles, defined by coordinates, corresponding to Portuguese regions, including the Portuguese archipelagos Azores and Madeira. At this point, tweets must also meet the restriction of being written in Portuguese, which is performed by checking the language attribute assigned by Twitter. For each successfully validated tweet, the tweet *user id* is extracted and stored in a MySQL database together with other tweet specific information, such as *tweet id*, and *creation date*; ii) for each of the stored *user id*s use the REST API for retrieving user timelines, allowing to retrieve at most the last 3200 tweets of each user [13].

The second stage entirely depends on the information stored about users and tweets in the first stage. However, the process of retrieving the latest tweets from a user timeline is currently limited to 45 requests per hour, each one returning a maximum of 200 tweets from that specific user. Consequently, fetching an entire user timeline may take 16 queries, which are also limited to 180 per 15 minutes windows [5]. Concerning this process, a number of

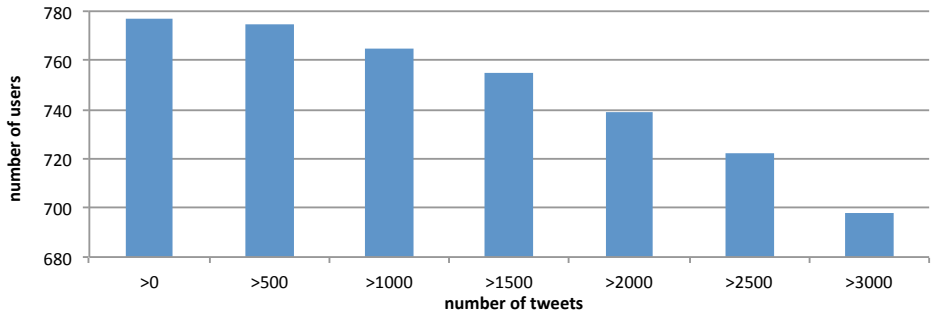

Figure 5 Number of users for which more than a certain amount of tweets was retrieved.

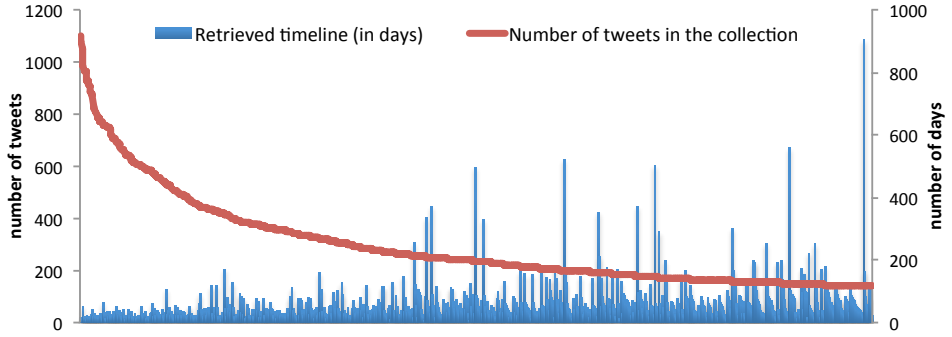

Figure 6 Relation between the time period covered by the retrieved timeline and the activity of a user during the 8-day period.

optimizations have been performed in order to avoid unnecessary queries. All information is stored in a MongoDB database, especially suitable for storing unstructured information, where duplicated entries are detected.

4.1 Analysis of the Recently Collected Data

By the time this document was written we had collected the user timeline of the 777 most active users during the 8-day period. In average, we have retrieved about 3167 tweets per user, totaling about 2.3 Million additional tweets. Figure 5 shows the number of users for whom the number of retrieved tweets achieved a given threshold.

The time period covered by the retrieved timeline varies according to the user activity, as expected. Therefore, the timeline of a very active user can go back as much as 8 days, while the timeline for a less active user can go back more than 900 days. Figure 6 illustrates such relation and supports the idea that more information about a less active user can nevertheless be retrieved by using our proposed approach.

5 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 a 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. We have found that such small community is mostly composed of young users who use this social network to exchange personal messages. The paper describes a method for expanding the database with related tweets, produced by the same community of users, by combining different Twitter APIs. The proposed methodology is now being applied, and by now allowed to collect 10 times more tweets than the original ones, corresponding to less than 10% of all the users in the database.

In a near future, we expect to complete retrieve the user timeline of our existing users, in order to further characterize the community of users. We also aim at studying different time periods, such as vacations where scholar subjects would not be so frequent.

Acknowledgments. This work was supported by national funds through FCT – Fundação para a Ciência e Tecnologia under projects PTDC/IVC-ESCT/4919/2012 (MISNIS) and PEst-OE/EEI/LA0021/2013, and under Grant SFRH/BPD/95849/2013.

References

1 A. Brito. O discurso da afetividade e a linguagem dos emoticons. *Revista Eletrônica de Divulgação Científica em língua Portuguesa, Linguística e Literatura*, 9, 2008.

2 G. Brogueira, F. Batista, J. P. Carvalho, and H. Moniz. Towards a characterization of tweets geolocated in Portugal. In *PROPOR 2014*, 2014 (submitted).

3 Tiago Daniel Sá Cunha. Sentiment analysis on Twitter's Portuguese language. Technical report, Faculdade de Engenharia da Universidade do Porto, 2013.

4 Eduardo Santos Duarte. Sentiment analysis on Twitter for the Portuguese language. Master's thesis, Faculdade de Ciências e Tecnologia, UNL Lisboa, 2013.

5 Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter Data Analytics*. Springer, New York, NY, USA, 2013.

6 Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL'13. Association for Computational Linguistics, 2013.

7 M. Oussalah, F. Bhat, K. Challis, and T. Schnier. A software architecture for twitter collection, search and geolocation services. *Knowledge-Based Systems*, 37(0):105–120, 2013.

8 Yanghui Rao, Qing Li, Xudong Mao, and Liu Wenyin. Sentiment topic models for social emotion mining. *Information Sciences*, 266(0):90–100, 2014.

9 José Carlos Santos and Sérgio Matos. Predicting flu incidence from Portuguese Tweets. In Ignacio Rojas and Francisco M. Ortuño Guzman, editors, *IWBBIO*, pages 11–18. Copicentro Editorial, 2013.

10 Tyler Schnoebelen. Do you smile with your nose? Stylistic variation in Twitter emoticons. *Working Papers in Linguistics*, 18(14), 2012.

11 Marlo Souza and Renata Vieira. Sentiment analysis on twitter data for portuguese language. In *Computational Processing of the Portuguese Language*, volume 7243 of *Lecture Notes in Computer Science*, pages 241–247. Springer Berlin Heidelberg, 2012.

12 Statistic Brain. Twitter statistics `http://www.statisticbrain.com/twitter-statistics/`, 2014.

13 Twitter. Documentation, `https://dev.twitter.com/docs/`, 2013.

14 Wouter Weerkamp, Simon Carter, and Manos Tsagkias. How people use twitter in different languages. In *Proceedings of the ACM WebSci'11*, Koblenz, Germany, 2011.