Twitter Stream Analysis in Spanish

María D. R-Moreno
Departamento de Automática
Universidad de Alcalá
Alcalá de Henares, Madrid, Spain
mdolores@aut.uah.es

Álvaro Cuesta
Departamento de Automática
Universidad de Alcalá
Alcalá de Henares, Madrid, Spain
alvaro.cuestac@gmail.com

David F. Barrero
Departamento de Automática
Universidad de Alcalá
Alcalá de Henares, Madrid, Spain
david@aut.uah.es

ABSTRACT

Social Networks have opened to companies and politicians' new ways to understand what the clients and citizens are looking for. That is, companies need to understand what is happening in the market in order to be competitive. And politicians need to better understand what the people are worried about if they want to comply with the wishes of their voters. Until now, a significant amount of resources were dedicated to collect a small set of consumers or citizens opinions to conduct focus groups and surveys in pursuit of consumers or social insights. With the rise of social networks, things have changed. Companies and politicians now have the ability to gather more data than ever before on a large number of users in near real time and at a much lower cost.

In this paper we present the architecture we are building on top of *Twitter* in order to extract and analyze the mood of the users against some events. In particular, we have analyzed the impact that the "Boston Marathon" event produced in the public opinion. During the observed time frame we have observed little social iteration and a high number of retweets in the Spanish-speaker twitter community.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

Social networks, sentiment analisys, Twitter, tweets

1. INTRODUCTION

In the last decade Social Networks have opened to companies and politicians new ways to understand what the clients and citizens are looking for. Traditional marketing campaigns included questions such as why the people are using the product X the way it is now, why the people are not buying the companies? brand or why the citizens do not agree with some political decisions. But this information is not enough if companies want to survive in competitive environments and politicians want to comply with the wishes of their voters.

Companies need to understand what is happening in the market place. More importantly, marketers need to understand the *why* behind consumer behaviours. And politicians also need to understand what are the citizens thinking about special events with high social impact or what is the general feeling against certain decisions. Just by knowing the answer to those questions companies will ultimately identify their consumers and politicians can better understand what the people are worried about and try to solve their problems.

Until now, a significant amount of resources were dedicated to collect a small set of consumers or citizens opinions to conduct focus groups and surveys in pursuit of consumer or social insights. With the rise of social media, things have changed. Companies and politicians now have the ability to gather more data than ever before on a large number of users in near real time and at a much lower cost. Social networks platforms, or simply social networks, are revolutionizing the way people can communicate, thanks to the new way of sharing interests and activities in different areas such as politics, economy, religion, hobbies, etc. A social network is a social structure made up of a set of actors (i.e. individuals or groups) and a complex set of relationships between these actors. The social network perspective provides a clear way of analyzing the structure of the whole social entities [**?**]. The study of these structures allows us to identify local and global patterns, locate influential entities, and examine network dynamics.

We can extract a lot of information about what is going on specific events or comments on products, but simply grouping and mining the information, not only about social relationships in terms of network theory, but also about the social impact and the general opinion of their members, as well as where the opinions are produced. Among the large number of social networks available on the Internet, there is one that contains a collection of characteristics that makes it especially interesting from the research perspective: Twitter. It is a platform that gathers people's opinions about a

wide spectrum of topics, and most of them are open, allowing a fascinating field of research. Not surprisingly, there is a large corpus of literature devoted to Twitter analysis [**?**] such as applying Twitter to, among other tasks, events detection [**?**] or public health [**?**].

Twitter allows their members to send and read text-based messages up to 140 characters long, known as *tweets*. Although users follow each other in a graph-like manner, these follows serve only for subscription purposes, while *tweets* remain publicly available online for anyone to read or reply. The vast amount of rapidly changing data found both on the Internet and specifically on social networks has led to a growing desire of knowledge extraction without manual intervention. The popular nature of these services is ideal for the discovery of trends and mass-opinion. The discovery and systematic analysis of knowledge is useful for both individuals and organizations.

Twitter analysis has been based on a complete set of techniques that allow, for instance, the early detection of trending topics [**?**]. One important and powerful mathematical tool used on Twitter analysis is Graph Theory [**?**, **?**] and the associated graph metrics [**?**], that let a deeper analysis of the relationships found on Twitter. Another relevant type of analysis applied to Twitter is sentiment analysis, that tries to quantify the emotional response to a given topic [**?**, **?**, **?**]. It allows us to determine the sentiment of thousands or even millions of posts by classifying them as positive, negative or neutral. However, the classification of data by sentiment represents only the first step towards discovering the consumer and citizens insight that marketers and politicians all seek to find.

Then, the purpose of the article is to describe the first step to develop a framework to automatically gather data from Twitter streams for further sentiment analysis. This tool will integrate sentiment analysis in Spanish, and to this end we captured a data stream in Spanish related to the Boston terror attack. This framework has been tested extracting tweets that contained the sentence in Spanish "Maratón de Boston" (which means "Boston marathon") along one week, and performing a basic analysis of some elemental statistics. The goal is to characterize the activity in Spanish that the Boston terror attack generated on Twitter.

The paper is structured as follows. Next section describes the architecture of the application we have developed to gather and analyze Twitter activity. Then, we report the data acquisition process followed by the description of the captured data about an event with high social impact. The paper finishes with some conclusions and future work.

2. ARCHITECTURE DESCRIPTION

Twitter offers several ways to access its data. Under the perspective of data analysis we should stress two: The Search API and the Streaming API. The first one is used to query Twitter about its content such as users or a keyword. This API imposes a limitation to the number of queries that a user is allowed to query (200 per hour), and therefore doing data-intensive analysis through the Search API is complicated. On the contrary, the Streaming API provides a real-time stream of tweets through an HTTP connection. Of course, this API also poses some limitations but even with those limitations it is easy to extract large amounts of data.

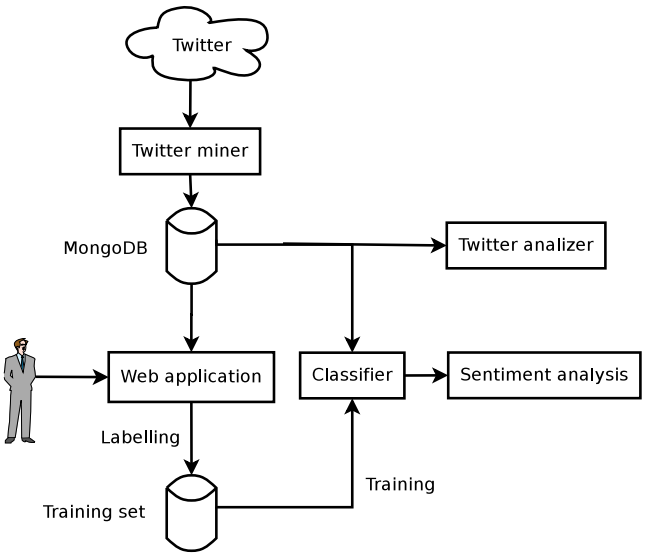

Figure 1: Architecture of the application with its subsystems: Twitter miner, Twitter analyzer, web application and sentiment analysis.

We have developed a Twitter analysis tool based on the Twitter Stream API. There are many programming languages that provide interfaces to this API: Java, C++ or R, just to mention some of them. In our case, given its simplicity, we have selected Python for data extraction. Data analysis, which is described in the next section, is performed with R because of its powerful statistical and advanced plotting functionalities.

The stream that Twitter provides can be filtered by a string, which might be a plain word, a hashtag (a keyword that begins with '#'), or a user name (which begins with '@'). A serious limitation in the Stream API is that Twitter only allows us to filter for each IP address, difficulting the data extraction in case there were need to gather data from several filters.

Roughly speaking, we can distinguish two steps on Twitter analysis: Data extraction and data analysis. In our tool those tasks are implemented by three independent (but related) subsystems, illustrated in Fig. 1, and described next:

1. **Twitter miner**. This subsystem is in charge of extracting data from Twitter using the Streaming API. Once data is extracted, Twitter miner stores data locally for further processing. These two tasks (extraction and storage) are performed in real-time.

2. **Twitter analyzer**. It is a collection of R scripts that access the database and generates reports with a collection of descriptive statistics and graphics such as time series.

3. **Sentiment analysis**. Subsystem based on a classifier aimed to analize sentiments linked the Twitter stream.

The trainer takes a collection of tweets, labels them according to the criteria of a human supervisor and then construct a classifiers. This subsystem is not discussed in this paper. The sentiment analisys subsystem is composed by the following elements:

(a) **Web application**: A web application designed to let a human trainer label tweets sampled from the main database. There are three labels that the trainer may choose that refer the intention of the person that wrote the tweet. The three classes are *positive*, *negative* and *neutral*. Positive and negative labels denote positive or negative feelings, while neutral tweets are tweets that contain objective information, non-Spanish or unintelligible texts. Classified tweets are stored in a database to train a classifer.

(b) **Classifier**: A classifier that is fed with the labeled tweets coming from the web application. Once it has been trained, it takes tweets from MongoDB and classify them as *positive*, *negative* or *neutral*.

(c) **Sentiment analysis**: A reporting tool that integrates the classifier. It takes the output of the classifier to generate a report with the sentiment analysis.

Some words should be dedicated to storage. Twitter miner keeps tweets and their associated metainformation such as geolocalization, owner or timestamp in MongoDB, which is a non-SQL database. It provides increased performance in comparison to other classical SQL driven databases such as MySQL or MariaDB. For massive data storage without the need of complex queries to the database and advanced functionalities, MongoDB seems a better choice. Once the application is deployed in a server, data extraction may begin.

3. DATA EXTRACTION AND EXPLORATORY ANALYSIS

Data acquisition began shortly after the terror attack. The first bomb detonated on April 16, 2013, at 2:50 p.m., local east coast time, and our data extraction begun at 00:43 GMT+2, four hours after the first detonation. The time frame to capture data was exactly one week, from April 16 to April 23, which seems a reasonable amount of time to obtain a general perspective. The low activity observed at the end of this time window supports this decision.

In order to capture tweets related to the terror attack, we filtered the tweets containing the sentence "Maratón de Boston". Of course, the attack generated a large number of hashtags on Twitter, such as "#marathon" or "#bombing", however, most of these hashtags are written in English and we needed tweets in Spanish to develop the sentiment analysis subsystem. A logical choice would have been "Boston", but this word is used in several languages such as English, Spanish or French, just to mention some of them, and we wanted to filter tweets in Spanish. Geolocaliation is not a good solution since the Spanish-speaker population is widely dispersed in Europe and America. Therefore we filted using the sentence in Spanish "Maratón de Boston".

Table 1: Statistics of the dataset used in this study. The dataset was created gathering tweets filtered with the sentence in Spanish "Maratón de Boston" (Boston matathon) along one week, from April 16 2013, 00:43 GMT+2.

Tweets	28,894
Retweets	12,864
Tweets without retweets	16,030
Users	24,990
Words without stop words	345,179
Geolocated tweets	255
Mentions	1,223
Responses	852
Mentions (not responses)	371

Table 1 summarizes the dataset of tweets we captured [1]. The overall amount of information stored is 105MB, which contains the tweets and associated metainformation such as its owner, timestamp and geolocalization data.

As Table 1 shows, along one week just after the terror attack, 24,990 Twitter users generated or retweeted 28,894 tweets, which yields an average value of 1,15 tweets per user. The 44% of the tweets (exactly 12,864) are actually retweets. Therefore a remarkable amount of activity on Twitter are just retweets. More surprisingly the number of mentions is as low as 1,223, and the mentions that are responses to a previous mention only 852, the 5% of the tweets[2]. This fact suggests that the social activity on Twitter around the Boston terror attack is not as social as one could expect. An explanation to this might be found in the nature of the tweets: They might have been used just to express feelings instead of communicating with other users.

4. DESCRIPTIVE ANALYSIS

The evolution in time of the activity on Twitter might arise valuable information. This section is devoted to study this evolution through a set of time series that plot the amount of activity on Twitter. We do not go through more elaborated techniques such as sentiment or graph analysis. In order to provide a proper granularity of the data, all the figures in this section plots data grouped by hour.

Figure 2 reports the number of tweets captured along one week from the terror attack, measured in number of tweets per hour. It is clear that most activity is focused short after the terror attack. After one day, at April 17 we can observe that the activity lowers dramatically, and then it remains almost constant. However, there are activity oscillations that decay with time. A more detailed view of days 19 and 20 with the tweets grouped by second (not shown) shows two peaks that coincide in time with the pursuit and detention of the terror attack suspects. This high frequency peak is filtered by the grouping used in the figure. Reading random samples of the tweets located in those peaks reveal that, actually, many of them are related to the suspects pursuit.

[1]This dataset is freely available upon request to the authors.
[2]We have not considered retweets in this computation; in that case, the value would have been even lower.

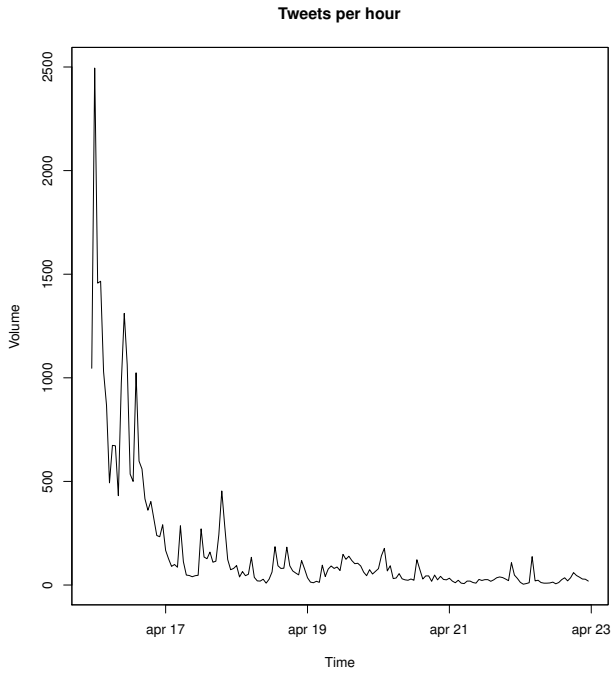

Figure 2: Number of tweets (original and retweets) grouped by hour containing the sentence in Spanish "Maratón de Boston". The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013, just one week.

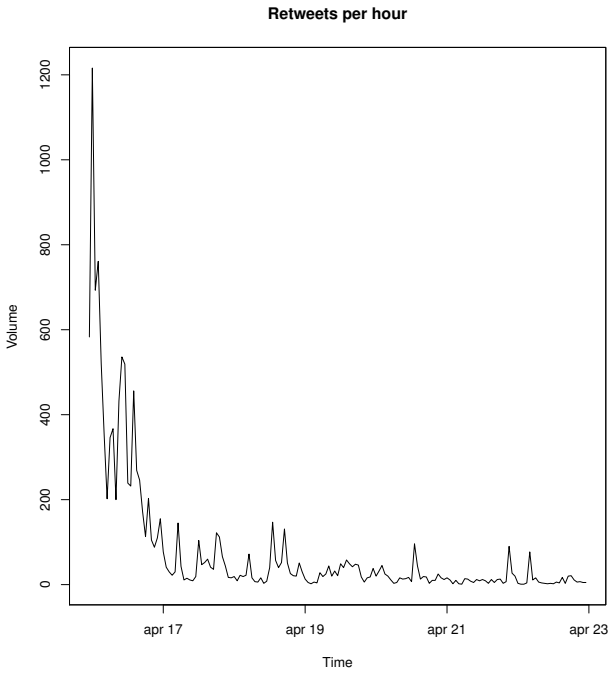

Figure 3: Number of retweets grouped by hour containing the sentence in Spanish "Maratón de Boston" (Boston Marathon). The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013, just one week.

To complement the perspective provided by Fig. 2, which includes original tweets and retweets as well, Fig. 3 reports a time serie that only contains retweets. As in the previous case, we observe that most retweets are originated just after the terror attack, and their number shrinks quite fast to remain almost constant with some small oscillations after April 17 and without those oscillations after April 19. One week after, the number of retweets containing the word "Boston" is very low. This behavior shows how the interest of Twiter users after the event decreases fast with time.

Figure 4 reports the number of tweets, excluding the retweets. Its behavior is very similar to the retweets. As in the previous case, we observe higher activity in the days close after the terror attack. If we compare Figs. 3 and 4 it turns out that much activity generated on Twitter around Boston terror attacks were retweets, which average roughly half of the activity on Twitter. It is consistent with the results shown in Table. 1.

Finally, Fig. 5 reports the tweets word count. To provide a fair measure, the figure excludes the retweets and stop words. The figure shows that, in average, the tweets contain around 12 words, and this value remains almost constant in the whole time serie. There are very few tweets with a small number of words. Twitter limits the number of characters to 140, and therefore there is obviously a higher limit of the number of words.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have described the architecture we are building on top of *Twitter* for sentiment analysis in Spanish. In particular, we have analyzed the impact that the "Boston Marathon" event produced in the public opinion. We have shown how the activity on Twitter is concentrated along one day after the attack with a rapid decay. The percentage of retweets along all the time serie, which is one week long, is quite high, around 44%, therefore much of the activity were not original tweets and many users just forwarded tweets. Notably, the amount of social iterations on Twitter was pretty low, less than 5% of the tweets were responses to a previous mentions, suggesting that users did not use Twitter as a communication channel, but rather as a platform to transmit a message. The dramatic event that motivated this dataset provides a good testbed for sentiment analysis.

6. ACKNOWLEDGMENTS

This work was partially supported by the Spanish CDTI project COLSUVH, leaded by the *Ixion Industry and Aerospace* company.

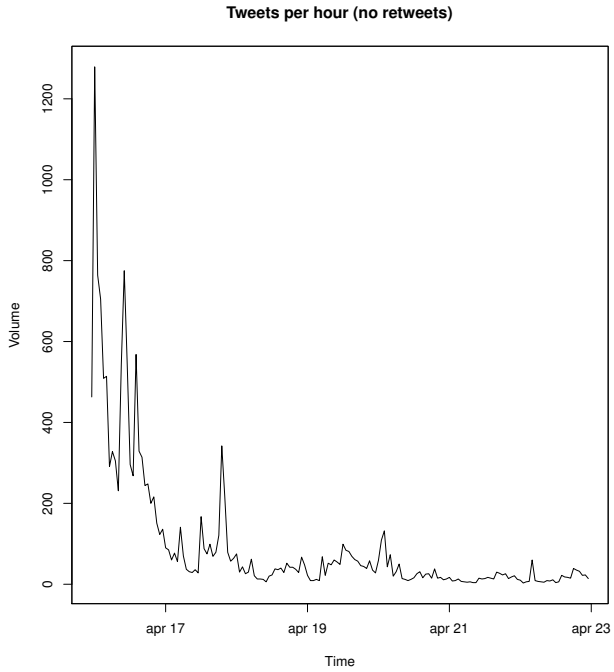

Figure 4: Number of original tweets (retweets are excluded) grouped by hour containing the sentence in Spanish "Maratón de Boston". The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013, just one week.

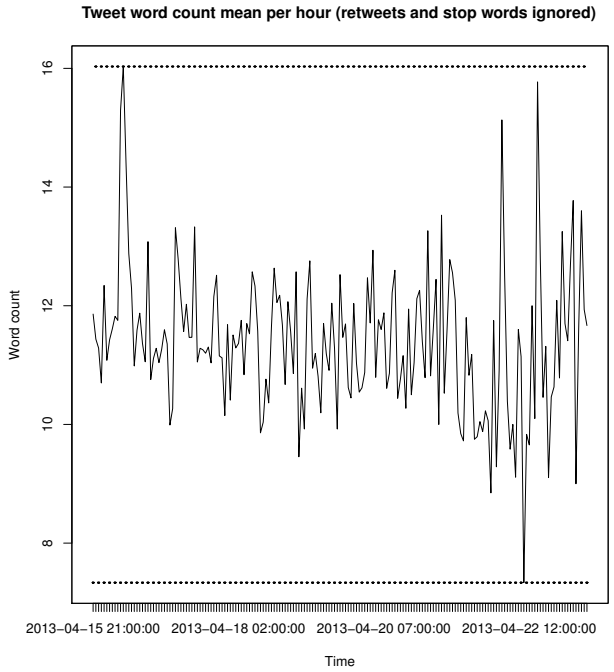

Figure 5: Number of tweets without retweets grouped by minutes containing the sentence "Maratón de Boston". The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013.