Social Data Analysis of Brazilian's Mood from Twitter

David N. Prata, Kleber P. Soares, Michel A. Silva, Daniela Q. Trevisan, and Patrick Letouze

***Abstract*—In this work, a software application was developed to analyze and visualize messages over *Twitter* social network, ranking the posts relatively to variations in moods within the Brazilian territory. Artificial intelligence techniques such as text mining and sentiment analysis were used for this purpose. The use of methods of machine learning allows determining the polarity (positive or negative) of *tweets* collected. Results were displayed in cartograms, through representations of *tweet's* geographic locations. Surprisingly, another study of *twitter's* mood from United States Nation showed similar results for the variation of moods throughout the day, hypothesizing a humor pattern for human beings during the period of 24 hours.**

***Index Terms*—Twitter, sentiment analysis, social data mining.**

I. Introduction

Twitter is a social networking and microblogging server for broad access and use, allowing users to send and receive messages in text up to 140 characters, known as *tweets*. These messages can be sent through many types of electronic devices connected to the internet, in real time.

Twitter's service is very popular in Brazil. The country is among the top five in subscribers using the social network that has over 500 million active users and 400 million *tweet*s sent per day [1]. Based on this large volume of data available and by the importance of extracting useful information, Twitter grows the researcher's interest in observing particular data from a large amount of data. A common example of this type of work is to search for reviews of a particular product or service. Difficult tasks for machine learning techniques such as natural language processing because of the large variation of linguistic terms, slang and emoticons presented in typed messages.

The sentiment analysis or opinion mining is a kind of data mining focused on meeting the demands for identifying and extracting emotions, opinions, or points of views, expressed in the texting messages. This work intends to analyze the sentiments expressed in messages sent on Brazilian territory during 24h. From datasets made up of posts collected from *Twitter* we could shape a visualization of these feelings. For this purpose, we developed a tool able to perform natural language processing and to classify messages as positive or negative, presenting their send locations on a map.

Some works [2] of moods in Twitter used the data set provided by [3] which is used for *tweet*s in English language composition database. To analyze the feelings in the Brazilian territory, we needed to collect *tweet*s in Portuguese language. Accordingly, we sorted the *tweet*s in manual and automatic mode, in positive and negative polarities, thus forming a training base to Brazilian Portuguese language. For this training base, a Naive Bayes classifier was trained to recognize the polarity of *tweet*s message with an accuracy of 79%, i.e. near to the average pairwise agreement rate between human text classifiers of 80% [4].

As a final result, the representation of feelings is illustrated on a cartogram of Brazil, showing the mood variations generated from federated states in the course of a day.

This paper is organized as follows: Section II provides information on the methodology used, with details of the Natural Language Processing (NLP), the creation of the training base, the collection of data streaming from Twitter in real time, the data classification and visualization of results obtained through the use of colored cartogram. In Section III, the results of the work are presented, and in Section IV, the conclusions.

II. Methods

The application was developed in Python, a high level program language, object-oriented, dynamically, and strongly typed language. For our purposes, we used the IDE JetBrains PyCharm.

A. Natural Language Processing

Natural Language Processing is an intersection field of study among different areas such as science and computing and linguistic, concerning the interactions between computers and human language. The goal of NLP is to understand natural language, allowing computers to extract meaning from natural language input.

The biggest challenges of NLP are the ambiguity presented in natural language texts, and the complexity of semantic information contained in a single sentence. In this work, we used the Natural Language Toolkit (NLTK) to perform the text pre-processing, removal of stopwords, and the use of stemming through RSLP Stemmer (Remover Suffixes of the Portuguese Language) for morphological normalization of texts.

A common element in *tweet*s is the emoticons. Emoticons are simple set of letterings, mostly containing only two characters. Before removing words with few characters, we added emoticons in a white-list for the classifier training set. Some most used emoticons examples are shown in Table I.

Another topic to consider are the misspelled *tweet*s, mostly in the form of repetitions of letters in words, these have been treated with the use of regular expressions. The goal is to create specific rules to replace the incidence of repeated

Manuscript received August 25, 2014; revised November 1, 2014. This work was supported in part by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

David N. Prata, Kleber P. Soares, Michel A. Silva, Daniela Q. Trevisan, and Patrick Letouze are with the Department of Computer Science of the Federal University of Tocantins – "Universidade Federal do Tocantins", Palmas-TO, 77.001-009 Brazil (e-mail: ddnprata@uft.edu.br).

letters in a word for just the word itself. For example, to find the word "cooooool" or "looooove", these would be treated for its "cool" and "love", respectively. By the use of regular expressions other eliminations were held, e.g., as part of the collected *tweet*s symbology of the service, at sign (@), sometimes used to indicate any user within the message, or the use of hashtags (#), in a simplified way to cite some issue. Both were removed with the use of regular expressions.

TABLE I: EMOTICONS EXAMPLES FROM A WHITE-LIST

Emoticon	Meaning
:)	Happy face
:D	lagh
:(	Sad face
;)	Flashing
:-)	Happy face (with nose)

B. Training Bases

Databases for sentiment analysis in Portuguese language are not easily found as in English [5]. After a few failed attempts to find a set of examples already sorted, we decided to build up a new base of examples gathered especially for this work.

Several different of sources were gradually adding content in order to improve the training base by performing tests during its development. The training is itself formed by a series of collected *tweet*s previously classified by polarity. Part of the training base was also collected from opinions sites where complaints or compliments are expressed about many products and services. One of these sites is the ElogieAqui <http://www.elogieaki.com.br/>, containing only praise comments, the site was used as a basis for the classification of positive polarity.

After the training base was completed, it was composed of 22862 sentences classified by polarity. From this amount, 9674 forms the 'positive' training base, and 13188 the 'negative' base. We decided to ignore the neutral class for this project because it was useless for our purposes.

C. Real Time Twitter Stream

At first glance, the collection of *tweet*s was performed using the Twitter Search API that allows queries against indexes of recent or popular *tweet*s and behaves similarly, but not exactly like, the search feature available on web or mobile Twitter clients, as found in <http://twitter.com>. Although the API (Application Programming Interface) has been very useful, the collection could not be performed in real time or with geolocated data. These facts lead us to discard its use in our design. The solution was to use the Streaming API with focus in completeness rather than relevance *tweet*s, enabling access to public data streams from Twitter.

D. Georefenced Data

To use the geographic component of virtual communities, it is necessary to transform pre-existing geographic information in a "usable" form [6]. Usable forms include latitude and longitude bounding box around geographic features, and polygonal representations (e.g., the shape of the map of Brazil). Along with the attribute information attached these data, such as a user name, population, etc.

One of the project requirements for the collection of *tweet*s was the need of the *tweet*s to be georeferenced. There are two types of variable location in a *tweet*: location and coordinates. The former is a free Twitter's field where the user can provide any location for his/her profile, which can be detrimental to the work georeferencing, for example, in the case of users using imaginary places or simply leave the field blank when creating his/her profile. One study [7] on the quality of information from Twitter showed that 16% of *tweet*s in English language does not have any information about the location field, and when filled out only 66% have a valid geographically entry. But even when not ruled out, there is likelihood for an incorrect location supplied by the user. In

In this case, our option was to use the coordinate's variable as a parameter to filter *tweet*s based on location. It provides a pair of coordinates, latitude and longitude. The format used by the variable Geo JSON is a format for encoding a variety of spatial data structures based on JavaScript Object Notation (JASON). Thus, it could enable the collection of *tweet*s sent only within the Brazilian territory.

E. Classification

One of the most common tasks in machine learning is to identify a record or object, and allocate it to a pre-established class [8]. The purpose of the classification algorithm is to find some similarity and correlation of the attributes with the class. The classification process can use it when trying to predict the class of a new example. Algorithms normally used to perform the classification procedure. Support Vector Machine (SVM), Naive Bayes, Maximum Entropy and algorithms based on neural networks [9], besides the k-nearest neighbors (k-nearest neighbors or kNN) cited by [10] and [11] are good algorithms to perform the sentiment analysis among others in the literature.

Works like [12] showed that Naive Bayes is a simple algorithm and taken as a starting point for the classification task as it has achieved good overall performance and are much faster than SVM in the application. In fact, the cost to train SVM for large data set is a clear disadvantage. In comparison with the SVM both kNN and Naive Bayes are very simple and easily assimilated. The Naive Bayes classifier is superior in terms of CPU and memory consumption, as shown by [13]. In many cases their performance is very close to the most complicated techniques. This fact has been decisive to choose it as main classification method in this work. The implementation class *nltk. classify. naive ayes* used was present in NLTK. All classifier training was done manually, using the training base, already labeled, divided between two text files containing terms labeled as "positive" and "negative". A list of the most effective features to distinguish the polarity of words is generated as shown in Fig. 1.

After the training phase, the classifier was ready to be tested. It follows a sample input to be analyzed:

"Feliz com meu resultado na prova de habilita ção de hoje" This message goes to the classifier, already trained, separating the terms presented in the phrase, ignoring stopwords and making use of stemming, the remaining terms:

['feliz', 'resultado', 'prova', 'habilita ção', 'hoje']

Thus the classifier checks its dictionary of terms (feature set), from which he was trained, recognizing the presence or absence of the terms in the document. Example snippet from the dictionary:

```
{'contains (prova)':              True,
'contains (perdido)':             False,
'contains (=))':                  False,
 'contains (feliz)':              True,
 'contains (parab éns)':          False,
 'contains (resultado)':          True,
'contains (habilita ção)':        True,
                ...}
```

The output of this example is "positive, +1," because, "feliz" is a very effective feature, with a score of 13 on the likelihood ratio, to distinguish the polarity of sentiments, as can be seen in listing as the most effective characteristics. Precision was used as metric for evaluating returning a value of 0.798, or about 80%.

```
Most Informative Features
           contains(bom) = True           positi : negati =     20.3 : 1.0
             contains(:d) = True          positi : negati =     17.2 : 1.0
          contains(super) = True          positi : negati =     13.4 : 1.0
           contains(amor) = True          positi : negati =     13.4 : 1.0
           contains(*-*) = True           positi : negati =     13.4 : 1.0
          contains(feliz) = True          positi : negati =     13.0 : 1.0
        contains(comprei) = True          positi : negati =     12.0 : 1.0
         contains(demais) = True          positi : negati =     10.9 : 1.0
 contains(parab\u00e9ns) = True          positi : negati =     10.9 : 1.0
           contains(boa) = True           positi : negati =     10.9 : 1.0
             contains(:() = True          negati : positi =      9.2 : 1.0
          contains(boas) = True           positi : negati =      8.8 : 1.0
           contains(amo) = True           positi : negati =      8.4 : 1.0
           contains(kkk) = True           positi : negati =      7.8 : 1.0
         contains(certo) = True           positi : negati =      7.8 : 1.0
```

Fig. 1. List of the most effective features

At the end of the process, the output of the classifier was combined to form a geographic reference file of comma separated values (CSV), as shown in Fig. 2.

```
1  negativo, -1,-28.4087922,-55.005077
2  positivo, +1,-8.06259389,-34.88818219
3  positivo, +1,-5.7780471,-35.2765069
4  positivo, +1,-22.90357067,-47.05822254
5  positivo, +1,-15.84760695,-47.96690912
6  negativo, -1,10.4682472,-66.8001126
7  negativo, -1,41.2767491,-8.375511
8  positivo, +1,-1.05093883,-46.76257723
9  positivo, +1,-5.12593917,-39.72980535
10 negativo, -1,-7.7813308,110.3885321
```

Fig. 2. CVS file containing the polarity of *Tweet*s and geotagging.

F. Visualization

The idea of the cartogram was used for the visualization of feelings in Brazilian territory with color grading where you can follow through the shapes and colors of the CSV data file obtained in the step described in the previous section. In this representation of the data collected stage, the GIS software Quantum GIS (QGIS) was used for the creation of cartograms from geographic coordinates collected from *tweet*s. As seen in Fig. 2 of the previous section, we can observe the output polarity assigned to each *tweet*, by analyzing the numerical assignment generated in relation to which sense it belongs, and its latitude and longitude.

QGIS make the latitude and longitude data easily integrated into axes to which they belong on the location within the map coordinates of Brazil. A file in shapefile format / ESRI in Reference System WGS 1984 in lat / long was used for this purpose.

To represent the polarity of sentiments and the amount of *tweet*s, a palette of colors was provided by divergent Color Brew 2 [14]. As seen in Fig. 3, the positive *tweet*s were mapped in green color and the red color was used to the negative *tweet*s, similar to the work in [15].

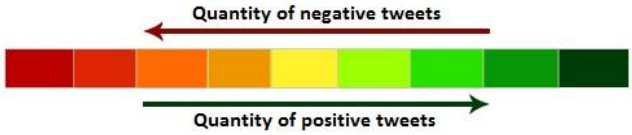

Fig. 3. Color palette used to represent polarity feelings and the quantity.

The cartogram generated maps the number of *tweet*s as variable, the more higher the number of *tweet*s with certain feeling, the more darkly the color of the area on the map. The absence of color in a region indicates that no *tweet* was collected at that time.

III. Results

Our intention was to visualize the mood variance over the course of an entire day. For this purpose, we collected 38,061 (thirty-eight thousand and sixty-one) *tweet*s sent in Brazilian territory. The collect started at 00h:00min and finish at 23h:59min. Running the trained classifier, the messages were classified as positive or negative, and were marked on the map in a timely manner as seen in Fig. 4, where the red dots represent the negative *tweet*s and in green the positive ones.

Fig. 4. *Tweet*s classified and mapped ccording to their origin.

To better understand the moods for each region (federated state) and to visualize the time when the messages were sent, we designed 24 cartograms, each one representing a time of day collected. The polarity was represented by colors, where the highest amount of *tweet*s of a type, positive or negative,

sets the color green or red, and the greater the numerical the darker polarity is the representative color (see Fig. 5- Fig. 6).

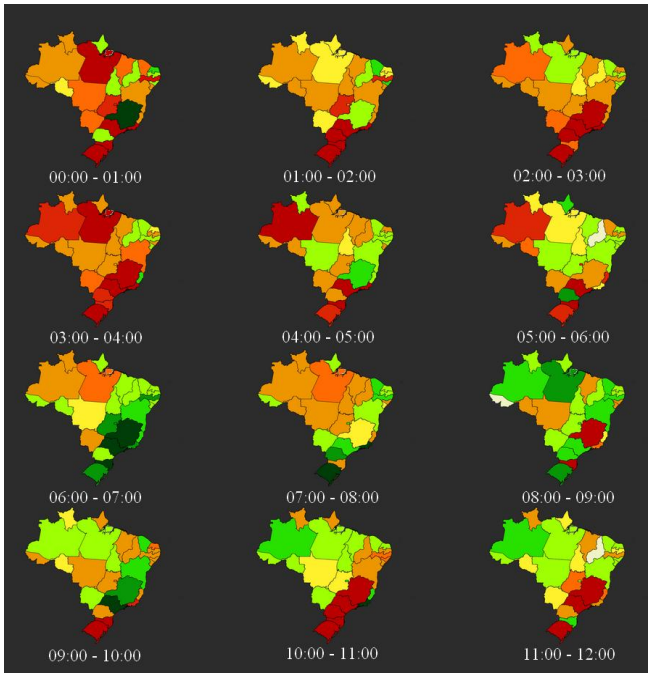

Fig. 5. Cartograms generated from the collection between 00h to 12:00h.

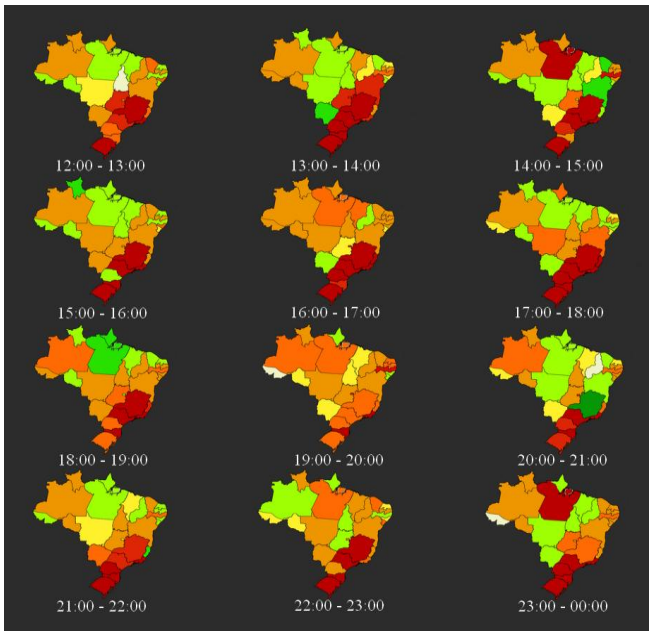

Fig. 6. Cartograms generated from the collection between 12h to 23:59h.

By the cartograms generated, we could observe a predominance of positive *tweet*s in the morning. To depict the positive moods compared to the negative ones, a new map was made, as shown in Fig. 7.

We can observe that the rate between positive under negative *tweet*s reaches its peak between 6am to 7am, in the morning, and continue during the morning, until 10am. It is followed by a large low during the hours between 11am to 1pm, where it returns to sag at around 2pm. Finally, we have a new high in the period of 4pm with fall again only around 9pm. Similar results were reported by [15] with the difference that there is a gradual decline in the number of positive *tweet*s after 6am until 3pm. The almost linear result from their work can be explained because of the much smaller sample space used in the present work.

Fig. 7. Graph showing the variation of mood over a day.

IV. Conclusion

In this work, a tool was developed to classify messages from the *Twitter* social network to assess variance in mood. For this purpose, we made use of techniques of textual mining and sentiment analysis, using machine learning techniques. The results were displayed in charts with their geographical locations using GIS techniques.

The accuracy achieved was 0.798, very close to the 80%, a result considered satisfactory since it is very close to human perception of feelings [4].

The final result was very close to those achieved by [15]. Although our work target only *tweet*s from Brazil, and the other study [15] collected only *tweet*s from United States Nation, the studies showed similar mood results for most of an entire common day.

This study conjectures a humor pattern for human beings during the period of 24 hours throughout the day.

References

[1] H. Sul, A. R. Dennis, and L. I. Yuan, "Trading on Twitter: The financial information content of emotion in social media," in *Proc. 2014 47th Hawaii International Conference on the System Sciences (HICSS)*, pp. 806-815, 2014.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter." in *Proc. the fourth ACM international Conference on Web Search and Data Mining*, pp. 65-74, 2011.

[3] C. Meeyoung *et al*., "Measuring user influence in twitter: The million follower fallacy," *Access our Sentiment Analysis APIs 14 Languages Covering the World*, vol. 10, pp. 10-17, 2010.

[4] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165-210, 2005.

[5] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. the International Conference on Language Resources and Evaluation*, vol. 2, 2010.

[6] B. Hecht and D. Gergle, *A Beginner's Guide to Geographic Virtual Communities Research*, Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, pp. 333-347, 2011.

[7] B. Hecht, H. Lichan, S. Bongwon, and H. C. Ed, "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles," in *Proc. the Sigchi Conference on Human Factors in Computing Systems*, pp. 237-246, 2011.

[8] J. Schmitt, "Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo," PhD Thesis, Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Ciência da Computação, 2005.

[9] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478-514, 2012.

[10] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2622-2629, 2008.
[11] G. Vinodhini and R. M. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, no. 6, 2012.
[12] D. Steinbruch, "Um estudo de algoritmos para classificação automática de textos utilizando naive-bayes," PhD Thesis. Pontifícia Universidade Católica, 2006.
[13] J. Huang, J. Lu, and C. X. Ling, "Comparing naive bayes, decision trees, and SVM with AUC and accuracy," in *Porc. the Conference on Data Mining ICDM. Third IEEE International*, pp. 553-556, 2003.
[14] M. Harrower and C. A. Brewer, "ColorBrewer.org: An online tool for selecting colour schemes for maps," *The Map Reader: Theories of Mapping Practice and Cartographic Representation*, pp. 261-268, 2011.
[15] A. Mislove. S. Lehmann, Y. Ahn, J. Onnela and J. N. Rosenquist. Pulse of the Nation: U.S. mood throughout the day inferred from Twitter. [Online]. Available: http://www.ccs.neu.edu/home/amislove/twittermood/ , 2010.

David Prata was born in Goiânia, Brazil on 18th September, 1965. Dr. Prata completed his bachelor of computer science in 1992. Then on, he went to complete his specializing in academician. He worked as an system analyst in Tocantins Government, being in charge for the accountability and financial systems. Later, he successfully completed his master degree in computer science from Campina Grande Federal University, with application research in education in 2000 year. He coordinates graduate and undergraduate courses in computer science at Alagoas Faculty in Maceio, Brazil. He was allotted to Federal University of Alagoas in 2006. Then, he moved to Federal University of Tocantins. His doctoral was developed in part at Carnegie Mellon University, USA, completed in 2008. He is currently coordinating a master degree in computational model. His research interests are education and ecosystems.

Michel A. Silva graduated in computer science from the Federal University of Tocantins in 2008. Currently he is pursuing a masters of computer modeling system by the Federal University of Tocantins and he works as a legislative assistant esp. programming - Legislature of the State of Tocantins, Brazil, mainly in the following areas: development, php, java, rails, c + +, usability, information architecture, data mining.

Daniela M. Trevisan holds a degree in information systems at University Lutheran Center Palmas (ULBRA). She concluded specialization in database by Catholic University of Tocantins and specialization in management information systems from Federal University of Lavras. She is currently an information technology analyst at Federal University of Tocantins. She has experience in the area of computer science, with an emphasis on the database, analysis and management information systems and data mining.