Data mining of social manifestations in Twitter: An ETL approach focused on sentiment analysis

Marcela Mayumi Mauricio Yagui
Graduate Program on Informatics
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
marcelayagui@ufrj.br

Luís Fernando Monsores Passos Maia
Graduate Program on Informatics
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
luisfmpm@ufrj.br

ABSTRACT
The objective of this study was to analyze sentiments of users of online social network twitter to understand how people manifested toward the article published by the magazine Veja on 04-18-16 entitled "bela, recatada e do lar" (beautiful, demure and from home) in an attempt to understand how this behavior evolved in two weeks and to assess which events had aroused greater reaction from people. To this end, a data mining technique known as sentiment analysis was used with the help of the ETL (Extract, Transform & Load) methodology and the Naive Bayes probabilistic learning algorithm. Moreover, the null hypothesis was formulated and tested to see whether two events that took place during the collection period influenced, in fact, the polarity of analyzed sentiments in the generated database.

CCS Concepts
• Information systems→Data mining • Information systems→Sentiment analysis • Networks→Social media networks.

Keywords
Sentiment analysis; data mining; information retrieval; Twitter; Naive Bayes.

1. INTRODUCTION
With the increasing phenomenon that is digital inclusion in the last decade and consequent widespread access to the Internet, transmission of information has become a progressively faster process and this has encouraged the emergence of the Information and Communication Technologies (ICT) and sophisticated online social networks (OSN).

These technologies enabled the use of electronic resources that favored the enhancement and celerity of the information transfer process. Therefore, are potentialized the resources of access, dissemination, cooperation and diffusion of information and of knowledge [19].

The OSN arose from the ability to approximate people regardless of distance, and many other factors that caused intrinsic changes in the communication paradigm after the advent of the Web.

The OSN are able to disseminate information in an amazing speed, which made easier mobilization to a cause. The self-communication of masses is the technological platform of the autonomy culture. From this autonomy, words, criticism and dreams of social movements extend to great part of society [7].

The mobilization of virtual crowds in OSN is a phenomenon that has been studied in order to predict behaviors, events and happenings, having its origins in the same rules that govern the mobilization of real crowds, varying, however, in size, speed and power [4].

In current OSN the old boundaries of space and time do not exist and techniques such as Information Retrieval and Collaborative Filtering assist in the process of monitoring user navigation and use of information.

The monitoring of the OSN is a crucial element in the process of learning about the behavior and opinions of users as a way of predicting events and happenings. We can consider that to enhance a certain experience of use is one of the primary objectives which must be realized by the data analyst. Moreover, the diffusion and popularization of OSN brings new opportunities in monitoring for the extraction of information which can be used in other fields such as politics, communication and scientific marketing [16].

The monitoring of OSN requires that metrics and resources be defined to verify the extent of their use and the electronic behavior of users though messaging and Web data Mining [20]. Data mining refers to the discovery of new information in function of patterns or rules embedded in large amounts of data. In practice, data mining needs the use of algorithms, information retrieval techniques, artificial intelligence, pattern recognition and statistics to be effectively carried out on large files and databases and in order to search for correlations between different data that permit acquire beneficial knowledge to an organization or an individual [9]. One of the areas in which data mining is strongly applied is in Decision Making, which requires filtering of relevant information and the provision of probability indicators [5].

In this work we applied the data mining technique known as sentiment analysis[1] [2] with support from the Naive Bayes probabilistic algorithm [24] to extract information about the users' opinions in the OSN twitter[2] during the social arousal "bela, recatada e do lar" (beautiful, demure and from home). Twitter

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

[1] Sentiment analysis consists in the attempt to predict the meaning of an opinion expressed in text, based on data mining.

[2] http://twitter.com/

plays an important role in this type of analysis as an open data source and being one of the most currently used OSN in the world. According to the company, in 2011 twitter was already accessed by over 200 million users producing a volume of 110 million posts per day[3]. The social network was chosen because it the microblogging system most widely used in the world, allowing its users disseminate short messages with a limit of 140 characters in real time to all linked to its network [22].

2. BASIC CONCEPTS

With the expansion of internet, the OSN now play a critical role in people's lives. Through OSN, they are provided with a communication channel and direct access to knowledge, which favors the exchange of experiences and information among people, which can be used, among other things, to share digital media (text, images, videos, audios, etc), exchange knowledge, maintain contacts, keep institutional history, conduct advertising and virtual marketing, support decision making, implement and establish interaction with teachers and researchers, enabling the interconnection and exchange of experience between organizations of the three sectors, among many other purposes [1].

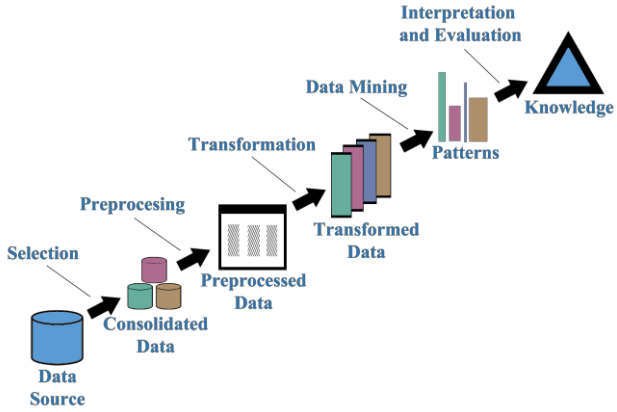
In compliance to these facts, information gathering and recognition of behavior patterns in these social networks have become an increasingly vital necessity to support decision making. Data mining assists in this process and can be regarded as a major step, in a larger process, known as Knowledge Discovery from Data (KDD) [14].

2.1 KDD and data mining

In KDD occurs the preprocessing of data (which includes all stages of preparation of data, such as funding mechanisms, organization and processing of data) and post-processing (phase that occurs on results obtained after data mining). In this sense, according Fayyad [10] "KDD is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". It is, therefore, the data analysis which it can be transformed into useful information, quickly understood by the user and that may bring some benefit to its possible decision-making.

The term KDD is seen as the process of taking temporal data collected from strategic information, therewith, making possible to attempt predictions about a problem or future events. Data mining is a very important step of KDD, since mining may occur in many modalities, i.e. there are different techniques and algorithms that can be applied to collected data, yielding knowledge in different ways. Although it is common to use the terms Data Mining and KDD with the same meaning, according to Fayyad [10], KDD is the process of extracting knowledge from data as a whole, and Data Mining, as only a stage particular to KDD, being in this way as the extraction of patterns in data realized with use of specific algorithms.

Due to the large volume of data stored by the most diverse organizations, data that were previously discarded, because it was believed to have no strategic value, are now availed. Also, there has been a growing concern in the capture, treatment and classification, as people are realizing the importance that these data will have in decision-making. In this context, data mining techniques are widely used because they made possible extraction oriented in steps, capable of producing information at a strategic level through robust methods of information retrieval [18].



Source: FAYYAD, 1996. (Adapted)

Figure 1. The KDD Process

2.2 ETL Methodology

To be effective, data mining must be accompanied by methodologies that assist in the process of extraction, transformation and loading of data. The ETL (Extraction, Transformation, Loading) process consists in a set of tools or algorithms, whose main functions are the acquisition of data from multiple data sources or from the multiple types of a single source, requiring different tools adapted to each source, cleaning and processing of data, where the data undergo modifications based on business rules and are uniformed, and loading of such data in environments such as data warehouses, data marts or data staging areas, in a organized and structured way [6].

3. RELATED WORKS

The following section is intended to present references about data mining in OSN and show several studies that were performed using methodologies similar to those applied in the present work.

3.1 Data mining in online social networks

Data mining has been used in various fields of knowledge. Recently with the boom of the OSN and microblogs such as Twitter, several studies have been conducted in the sense to quantify and qualify behaviors of sharing and messaging in this social network.

In the study performed by Kwak and Lee [15] twitter was used to explore how the scientific papers are shared in this OSN to understand the capabilities and limitations of the current practice of measuring or predicting their scientific impact on the web. In this study, the authors explored the dynamic sharing of papers on Twitter during a period of 135 days to understand how actively this social media reacts to papers and how well the online social responses reflect the established reputation in research communities.

In another recent study [13], the authors Saeed-Ul Hassan and Uzair Ahmed Gillani proposed a new metric to calculate the scientific impact of academic publications based on their

[3]http://www.forbes.com/sites/oliverchiang/2011/01/19/twitter-hits-nearly-200m-users-110m-tweets-per-day-focuses-on-global-expansion/

disclosure in OSN like twitter: the alt-index (similar to the h-index[4], but which considers the number of citations in social media, in order to capture the social impact of academic publications). In the study, data and references of tools and social networks such as Google Scholar, Twitter, Mendeley, Facebook, GooglePlus, CiteULike, Blogs and Wiki were collected in the period from 2010 to 2014. After collection those were used to calculate the Pearson correlation coefficient[5] among the alt-index and h-index, and among social quotes and academic quotes and they verified a high correlation between indicators that capture social impact, as well as an average similarity between the two indicators. In addition, quotes on social networks like twitter and facebook were about 40% more numerous than traditional quotes, indicating the strong power of influence that these networks have in the opinion formation and disseminating knowledge.

As we can see data mining in OSN can be performed during certain time intervals to provide useful and enriching information for data analysts, revealing "hidden impacts" and offering access to opinions of a wider audience, such as students, the government itself and the public in general.

3.2 Opinion mining and sentiment analysis

Another trend that emerged from data mining recently is sentiment analysis or opinion mining [23] which refers to the broad area of text mining, natural language processing and computational linguistics which involves the computational and semantic study of feelings, opinions and emotions expressed in texts. This type of mining enables detection of whether an opinion expressed in a text is positive, negative or neutral. It also can detect characteristic emotions like happiness, sadness or anger. This type of data mining has been rapidly becoming one of the areas of natural language processing most investigated in the last years [17].

Twitter has become quite popular in carrying out this type of analysis for its unique and characteristic format of microblog, allowing sending and sharing of text messages, images and videos quickly and through a network of millions of users, and exceptionally, for allowing text messages of, but limited to, 140 characters, which is determinant in how people can express themselves in this network and facilitates the work of mining and pattern recognition.

In this sense, several studies have been conducted focusing on sentiment analysis to classify emotions on twitter platform. One of the pioneers was the study of Bollen et al. [3], where the oscillation of sentiments expressed in messages from the OSN Twitter was analyzed in political and socio-economic events whose context could, somehow, be impacting in the identified fluctuations. In this study, classification of tweets generated six different classes of humor, created based on the comparison between the identified terms in the messages and terms that were previously defined and associated to each of the categories.

In another related study [8] the representation of sentiments was realized by classifying tweets based on hashtags, to separate desired categories, and in emoticons contained in the text messages, which assisted in training the classification algorithm, among other factors that were used in the extraction of feelings, such as punctuation and strings that formed keywords.

The classification of tweets in positive or negative polarities has also been previously studied by Pak and Paroubek [22] that in 2010 applied the Naive Bayes probabilistic classification algorithm to perform the procedure.

The Naive Bayes classifier also proved itself efficient for analyzing texts in Portuguese, as shown by Ferreira [11] where several methodologies of classification were used to analyze content in Portuguese and the Naive Bayes classifier demonstrated the most satisfactory results.

Another important work to be mentioned is the study of Nascimento et al. [21] where it is possible to observe the polarity of sentiments of the population in relation to reports disclosed by the media. In the methodology three topics in Portuguese language were previously selected for data collection, then the authors performed the manual labeling of content to analyze it through the use of three methods, one of them being the Naive Bayes. Again, the most satisfactory results were presented by this method.

Based on the studies previously indicated, sentiment analysis appears to show greater efficacy when the Naive Bayes classification method is used which, apparently, is the analysis technique that presents the best results for small text like those found on twitter social network.

Thus, this work presents a study that aims to analyze user behavior of the OSN twitter during a popular movement which was widely spread by the media, correlating facts and opinions to raise questions and formulate hypotheses. To this end, the study is based on data mining with support from the ETL methodology to, through the use of algorithms developed in python programming language, such as the Naive Bayes probabilistic classifier, the use of data retrieval techniques, pattern recognition and statistical concepts, perform sentiment analysis and tracking behavior patterns that lead to desired responses.

4. METHODOLOGY

The objective of this study was to analyze sentiments of users in the OSN twitter to understand how people manifested toward the article published by the magazine Veja on 04-18-16 entitled "bela, recatada e do lar"[6] (beautiful, demure and from home), to understand how this behavior evolved in two weeks and to assess which events had aroused greater reaction in people.

This process was carried out in three phases. In the first one, tweets were collected between days 04-22 and 05-03 to create the database that would be subsequently analyzed. In the second one, there was the classification of the messages from twitter, according to their nature in relation to the popular movement. The third one consisted in observation of the collected data and formulation of the research hypothesis, to investigate whether events that occurred during the time period in which the data were

[4] Index that quantifies the number of articles produced versus the number of citations in these articles. An h-index equal to 5 means a researcher has at least 5 publications quoted 5 or more times.

[5] Statistical concept that verifies the similarity between two variables of a metric scale.

[6] http://veja.abril.com.br/noticia/brasil/bela-recatada-e-do-lar

collected impacted positively or negatively on the users' behavior. These three phases will be explained below.

4.1 Data collection and pre-processing of the messages

Data collection was performed using the ETL methodology. It took place between April 22 and May 3, beginning four days after the start of the manifestation and ending after finding that the movement had reached its critical mass and was losing strength and scope, fact that was duly proven posteriorly after the analysis of the volume of tweets per day. This process was divided into three steps:

(i) Extraction – Between days 04-22 and 05-03 the data were collected from public posts from users of Twitter using the corresponding streaming[7] API. The keywords "bela recatada e do lar" and the hashtag "#belarecatadaedolar" were used to separate categories of desired tweets. Data were collected from a crawler implemented using the Node-Red[8] tool of the IBM Bluemix[9] service.

(ii) Transformation – In this stage messages were cleared so that there remained only data that corroborated with the research. For such, only messages in Portuguese were considered, being eliminated from the messages: special characters, punctuations and stop words.

(iii) Load – At this stage data were stored in the online database service Cloudant that has an implementation of Couchdb[10], which is a NoSQL database that works with JSON documents. This service was chosen because of a great advantage to use: data in the cloud are collecting tweets 24h a day automatically. On the other hand, this service presents a disadvantage when carrying out the analysis: the API service limits the amount of documents that can be read at a time. Because of this disadvantage, at the end of collecting all the data were exported from the platform to a JSON document and then saved in MongoDB database on the local machine, making it possible to access the data whenever greater speed was necessary because they are stored locally.

At the end of this process the data of 43,647 tweets were collected and stored.

4.2 The classification

In this second phase the messages were classified as POSITIVE, NEGATIVE and NEUTRAL according to their nature regarding the manifestation. In this process an implementation of the Naive Bayes probabilistic algorithm [12] in the Python language (version 2.7) was used, which was chosen from among different types of algorithms for mining existing data because of being the best that reflects the results for statistical demonstrations and also that facilitates the interpretation of the extracted information. The probability calculation realized by the algorithm can be expressed by the following formula:

$$P[H|E] = \frac{(P[E|H]\ P[H])}{P[E]} \quad (1)$$

where H represents the observed event, E the evidence, P[H] is the probability of the event before the evidence is seen, and P[H |E] is the probability of H given event E (i.e., the probability of H after the event E is seen). P[E] represents the sum of the iteration of both classes with the distribution estimation of the terms of the classes [12].

The classification process was divided into three steps:

(i) Manual Classification - a sample of 550 random tweets was selected to be classified manually and subsequently stored in different files for the positive, negative, and neutral. The classification criteria were as follows: (a) positive messages - jokes, memes and manifestation of support to women. (B) negative messages - retweets of messages and news in clear disapproval of the subject of the magazine, negative impressions in general, insults and offenses. (C) neutral – messages impossible to identify, such as messages whose contents are only links, photos or videos. The result of this manual classification was 155 tweets positive, 153 negative, and 242 neutral.

(ii) Training and testing - in this step we selected 2/3 of the tweets that were manually classified for training and 1/3 for testing. These data were used to train the learning process of the algorithm and verify its accuracy with regard to the body of the messages formed in the database. This test showed that the algorithm presented an accuracy of 78.6% in the classification, being very close to the human success rate, whose subjectivity analysis capability of a text oscillates between 72% and 85%, which is a very reasonable value.

(iii) Finally, after the learning phase of the algorithm the classification of all 43647 tweets from the database was performed.

4.3 Formulation of the research hyphotesis and observation of collected data

The main focus of the experiment was to evaluate the extent to which facts and events occurred during popular manifestations in social media can trigger impacts on the opinions and sentiments of people, that reflect in the form of emotions and positive and negative sentiments. Toward this, the following null hypothesis was formulated: Events and their disclosure influenced the polarity of tweets during the movement "bela, recatada e do lar".

The null hypothesis is a statement that corresponds to a particular aspect of the population, which will be accepted or rejected by statistical tests appropriated to the collected sample. The alternative hypothesis is a variation of the null hypothesis and corresponds to a statement that is "different", "higher", or "lower" than the defined by the null hypothesis. It will be accepted case the null is rejected.

Based on these precepts an investigation on the internet was led during the period in which data collection was carried out to monitor events and news published in journals and blogs that may have triggered "peaks" in the volume of twittered messages.

This research led to two distinct events, one in favor of the movement "bela, recatada e do lar", and another against.

[7] https://dev.twitter.com/streaming/public

[8] http://nodered.org

[9] https://console.ng.bluemix.net

[10] http://couchdb.apache.org

On day 04-23, the second day of collection, the first event[11] happened, groups of women mobilized against the enforced standard "bela, recatada e do lar", and protested in Brasília (Brazilian capital city), a clear demonstration of support for the feminist movement on twitter. This fact triggered a sharp increase in the volume of tweets compared to the previous day.

On day 04-25, the wife of a famous Brazilian priest spoke out against the feminist movement on Twitter, starting a campaign in favor of the conservative woman standard. The fact, however, only gained media notoriety at the end of day 04-27, gaining Internet publicity in 04-28, which was recorded as the second event[12]. Again, it was noted an increase in the volume of tweets, although lower than day 04-23, which had been gradually decreasing so far, indicating that the movement was already losing strength. The results of these analyses and the observed peaks in volume of tweets can be verified in the following section.

5. FIRST RESULTS OBTAINED

This section presents the first results obtained. Analyses of tweets related to the theme "Bela, recatada e do lar" were performed to ascertain if the influence of protests and their dissemination in journals and blogs increases the volume of tweets. After these analyses it was verified whether these events influenced in a negative or positive form the sentiment of the tweets. In preliminary analyses the numbers indicate a total 43647 collected tweets. Regarding the volume of tweets per day, the numbers indicate a progressive decrease in the rate of posts over time. The frequency distribution of the values in Table 1 better illustrates this result.

Table 1. Table of Frequencies for the volume of tweets

Factors	Frequency	Rel. Freq.	Perc. Freq.	Acum. Freq.
Day 1	6870	0,157399134	15,7399134	15,7399134
Day 2	9726	0,222833184	22,28331844	38,02323184
Day 3	7243	0,165944968	16,59449676	54,6177286
Day 4	3460	0,079272344	7,927234403	62,544963
Day 5	2441	0,055925951	5,592595138	68,13755814
Day 6	2192	0,050221092	5,022109194	73,15966733
Day 7	3446	0,078951589	7,895158888	81,05482622
Day 8	1804	0,041331592	4,133159209	85,18798543
Day 9	1442	0,03303778	3,303778037	88,49176347
Day 10	1605	0,036772287	3,677228676	92,16899214
Day 11	1942	0,044493321	4,449332142	96,61832428
Day 12	1476	0,033816757	3,381675717	100
Total	43647			

The frequencies of the table above may also be expressed by a line graph, providing a clearer vision of the distribution for the volume of tweets per day, as shown in Figure 2.

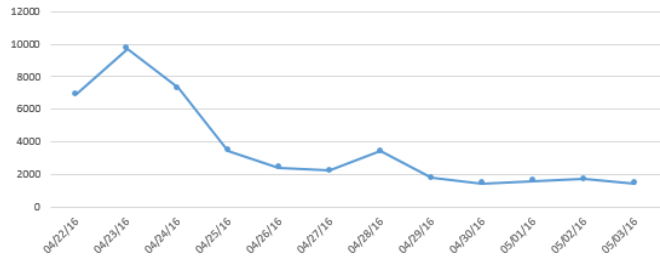


Figure 2. Volume of tweets per day

According to the graph of tweets per day, it is possible to observe two peaks in the distribution. These peaks correspond to the dates that occurred the two influencing events mentioned above, the feminist protest in 04-23 and the disclosure in online news of a campaign in favor of the conservative stereotype in 04-28. Figure 3 demonstrates these facts.

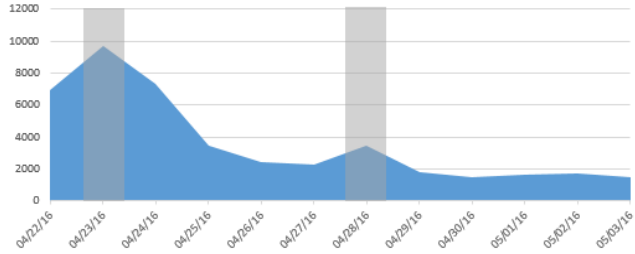


Figure 3. Volume of tweets per day with the dates of the events marked

Regarding the geographical distribution of the 43,647 collected tweets, only 1,548 (3.54%) were georeferenced, distributed in 460 cities. The low number occurs because to be georeferenced, tweets need to be published by means of a device that contains GPS, requiring that the user activates this feature. Without it, tweets rely solely on the location filled in by the user in his profile, which in most cases is inaccurate and does not match the real location. Table 2 describes the geographical distribution of messages during the movement, showing the cities with highest incidence of tweets.

Table 2. Tweets per city

City	Number of tweets
Rio de Janeiro	407
São Paulo	285
Porto Alegre	98
Belo Horizonte	94
Brasília	88
Curitiba	71
Fortaleza	64
Recife	55
Belém	48
Niterói	44
Salvador	44

Figure 4 illustrates in a bar graph the top 10 accumulated frequencies by city, as were expressed in Table 2.

[11]http://www.correiobraziliense.com.br/app/noticia/cidades/2016/04/23/interna_cidadesdf,528825/mulheres-protestam-contra-imposicao-do-padrao-bela-recatada-e-do-lar.shtml

[12]http://www.opovo.com.br/app/politica/2016/04/28/noticiaspoliticas,3608806/mulher-de-malafaia-rebate-campanha-bela-recatada-e-do-lar.shtml

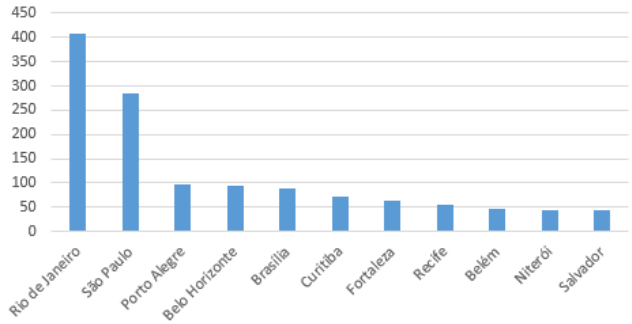


Figure 4. Tweets volume per city

6. SENTIMENT ANALYSIS RESULTS

In this section classification results of the collected tweets, according to sentiment, are presented. To better represent these results, values were assigned from 0 to 1 for classes of sentiments. The number 0 indicates that the polarity of the tweet is negative, 0.5 neutral, and 1 positive. Table 3 shows the descriptive summary obtained through the calculated measures of central tendency and dispersion. From the table, it can be observed that the low standard deviation indicates the data dispersion is small (only three classifiers with approximate values). Yet a slightly negative asymmetry or to the left indicates the mean is less than the mode. In the case of this sample, it happens the number of tweets classified with the neutral sentiment is higher, followed by the negative sentiment. The positive sentiment has the lowest proportion. The Kurtosis is positive and indicates the curve is Leptokurtic, or in other terms, presents an elevated top, meaning the data are located around the mode. This occurs because of the number of cases classified with the neutral sentiment (mode of the distribution).

Table 3. Descriptive summary

Mean	0,487055
Median	0,5
Mode	0,5
Standard Deviation	0,239794
Variance	0,057501
Asymmetry	-0,072602
Kurtosis	1,327155

Regarding to the sentiment associated to each tweet, approximately 10% were classified as positive, 13% negative, and 77% as neutral, according to the frequency indicated by Figure 5.

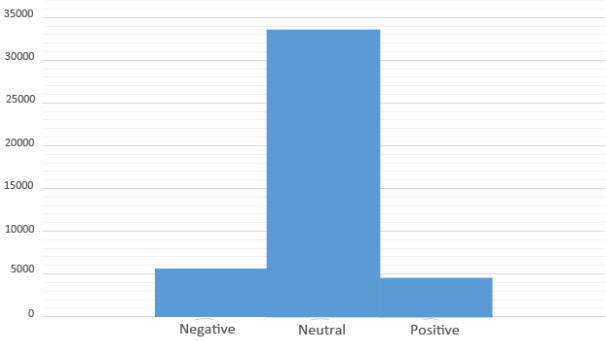


Figure 5. Histogram of classified sentiments

With respect to the total classified, the number of tweets aggregated by feelings within a day period is expressed by the frequency Table 4.

Table 4. Table of Frequencies of the sentiments per day

	Negative	Neutral	Positive	Total
Day 1	928	5239	703	6870
Day 2	1226	7428	1072	9726
Day 3	885	5620	738	7243
Day 4	448	2658	354	3460
Day 5	316	1880	245	2441
Day 6	282	1717	193	2192
Day 7	480	2580	386	3446
Day 8	226	1385	193	1804
Day 9	215	1079	148	1442
Day 10	183	1265	157	1605
Day 11	244	1544	154	1942
Day 12	166	1184	126	1476
Total	5599	33579	4469	43647

These values are also demonstrated graphically in Figure 6.

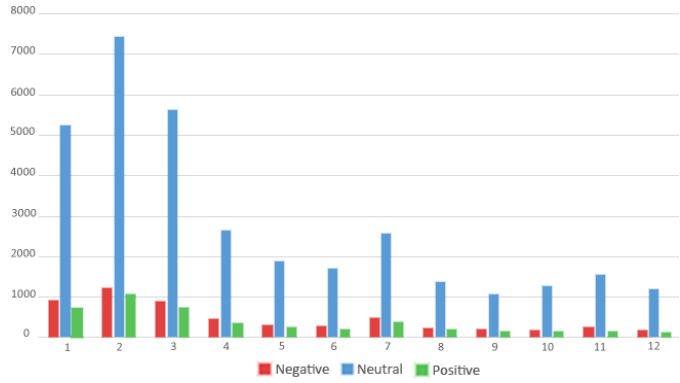


Figure 6. Classified sentiments per day

Figure 7 correlates to the geographic distribution with the sentiments. From the graph, it can be observed the city with the highest percentage of tweets with the negative feeling is Fortaleza, with 20%. In the other hand the highest rate of neutral tweets is 86% and belongs to the city of Curitiba. Finally, the highest proportion of positives is in Recife with 20%.

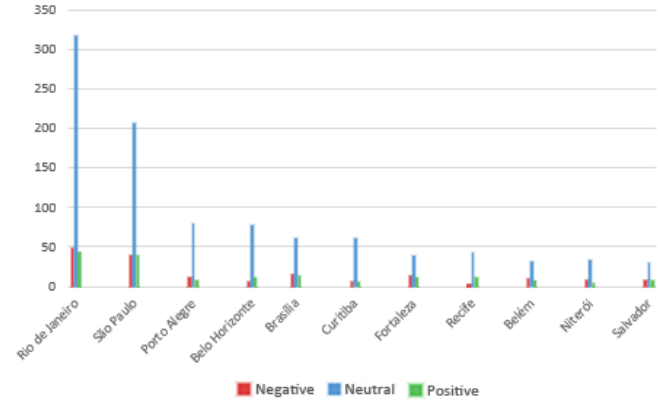


Figure 7. Classified sentiments in cities that published the most

Another interesting result of sentiment analysis that deserves to be mentioned is the ranking of tweets that most repeated due to a great number of shares (retweets) propagated. In this ranking, three of the five messages most repeated had a negative attribution, totaling 5539 negative tweets among the ones that most spread. Table 5 expresses these results.

Table 5. Table with ranking of retweets

Number of retweets	Tweet	Sentiment
3081	n aguento mais bela recatada e do lar em legenda de fotos façam parar	Negative
2073	A única diferença entre eu e Marcela é que aqui em casa A PRESIDENTE SOU EU e meu namorado é a primeira dama kkkkkk #belarecatadaedolar	Positive
1494	Eu e essa tal de Marcela Temer somos TAO PARECIDAS kkk #belarecatadaedolar #sqn https://t.co/Qb13vo82VS	Negative
1296	"bela, recatada e do lar https://t.co/nzvN89I9aZ"	Neutral
964	"Veja traça perfil de Marcela Temer: ""bela, recatada e do lar"". Voltamos aos anos 60!"	Negative

7. TESTS OF VERIFICATION OF THE HYPOTHESIS

The first clue to verify if the null hypothesis should be accepted demonstrates an increase in the volume of tweets in the days of the events. Additional statistical tests were realized in order to have an affirmative answer to the acceptance or rejection of the null hypothesis.

Considering the non-normality of the data, the non-parametric tests of correlation of Spearman and Kendall were used to verify if the null hypothesis can be accepted. Both tests correlate variables Sentiment and Day and yield as a result a value between -1 and 1. If the result is closer to the interval extremities, the variables are strongly correlated (if close to 1, the correlation is direct, is close to -1, the correlation is reverse). If the result is tending to zero, i.e., towards the middle of the interval, the correlation is weak. Observing Tables 6 and 7, both tests indicated a poor correlation between Sentiment and Day. Thus, the null hypothesis was rejected since there is no correlation between the day from the events and sentiment of the posted messages.

Table 6. Result of the Spearman test

	Sentiment	Day
Sentiment	1	-0,005256324
Day	-0,005256324	1

Table 7. Result of the Kendall test

	Sentiment	Day
Sentiment	1	-0,004430351
Day	-0,004430351	1

8. FINAL CONSIDERATIONS

As stated earlier, the main objective of this study was to analyze sentiments of users of OSN twitter to understand how people manifestated toward an article published by the magazine *Veja* on 04-18-2016 entitled "bela, recatada e do lar" in an attempt to understand how this behavior evolved in two weeks and to assess which events had aroused greater reaction from people. To this end, a data mining technique known as sentiment analysis was used with the help of ETL methodology and the Naive Bayes probabilistic learning algorithm. Moreover, the null hypothesis was formulated and tested to see whether the two events during the collection period influenced, in fact, the polarity of analyzed sentiments in the generated database. Up against performed studies and obtained results the following considerations can be made:

(i) The Naive Bayes probabilistic algorithm, implemented in python programming language presented an accuracy of 78.6% in the classification, which is a very satisfactory result if we consider the human success rate, whose subjectivity analysis capability of a text oscillates between 72% and 85%,

(ii) Based on statistical analysis associated to the volume of tweets per day, numbers indicate a progressive decrease in the rate of posts over time. However, it can be seen that two events generated peaks in the volume of tweets in days 04-23 and 04-28. Thus, there are strong indications that isolated events can have impacts on popular movements orchestrated on OSN.

(iii) Popular manifestations in OSN such as twitter are difficult to be studied and understood, as they tend to lose strength and scope quickly in the absence of facts and events that "boost" the movement.

(iv) The number of neutral sentiments is disproportionately larger than the positive and negative. This is mainly due to the fact that large numbers of users reproduce impartial message content or no content, such as retweets of online news and links. Example: "bela, recatada e do lar https://t.co/nzvN89I9aZ"

(v) In tests of the validation phase, due to non-normality of the data, non-parametric tests of correlation of Spearman and Kendall were used to verify if the null hypothesis could be accepted. The results showed a weak correlation between sentiments and day of the events that corroborate the hypothesis. As a result, the null hypothesis was rejected because there is no correlation between the data from the events and the sentiment of posted messages.

9. CONCLUSION

This work presented an empirical study in order to better understand the relationship between popular movements in OSN and sentiments and opinions expressed by their users. Between 04-22 and 05-03, data collection was carried out during the movement “bela, recatada e do lar” for posterior application of statistical investigations and sentiment analysis in collected data. From this it became possible to assess: (i) the influence of external events in volume of tweets; (ii) the pattern of sentiments expressed by users during the popular movement; (iii) the impact external events cause in the sentiments and emotions of users.

Based on the obtained results it was possible to observe that isolated events that occur during the social movement apparently generated fluctuations in the volume of messages. However, these events had little or no correlation with the polarity of sentiments expressed by users, fact duly corrabated in the section of the hypothesis verification test. Furthermore, the pattern of sentiments shown by users presented a great number of neutrals, followed by negatives and positives. This indicates that most users were indifferent to the movement.

Thus the experiment demonstrated, through modern data mining techniques, information retrieval and statistical analyzes, such as opinion mining and sentiment analysis, can be used to identify behavior patterns in popular manifestations in the social media.

For future works the inclusion of the categories sarcasm and emoticons should be considered to help create larger sets and more effective training and tests.

10. ACKNOWLEDGMENTS

This research is supported by Capes, CNPQ and PPGI/UFRJ.

11. REFERENCES

[1] Alonso, A. and Anastassakis, Z. 2013. Ambiente colaborativo Comum: produção e troca de percepções sobre o espaço urbano e seu uso. *Anais do Colóquio Internacional de Design - Edição 2013: Design Para os Povos* (Belo Horizonte, 2013), 70–79.

[2] Balamurali, A.R., Joshi, A. and Bhattacharyya, P. 2011. Harnessing wordnet senses for supervised sentiment classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), 1081–1091.

[3] Bollen, J., Van de Sompel, H., Hagberg, A. and Chute, R. 2009. A principal component analysis of 39 scientific impact measures. *PloS one*. 4, 6 (2009), e6022.

[4] Bonabeau, E. 2004. The perils of the imitation age. *Harvard Business Review*.

[5] Camilo, C.O. and Silva, J.C. da 2009. *Mineração de dados: Conceitos, tarefas, métodos e ferramentas*. Technical Report #Technical Report #RT-INF_001-09. Universidade Federal de Goiás.

[6] Campos, S.R. 2013. *Validação de dados em sistemas de data warehouse através de índice de similaridade no processo de ETL e mapeamento de trilhas de auditoria utilizando indexação ontológica*. Universidade de Brasília.

[7] Castells, M. 2013. *Redes de indignação e esperança: movimentos sociais na era da internet*. Zahar.

[8] Davidov, D., Tsur, O. and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. *Proceedings of the 23rd international conference on computational linguistics: posters* (2010), 241–249.

[9] Elmasri, R. and Navathe, S.B. 2005. *Sistemas de banco de dados*. Pearson Addison Wesley.

[10] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. eds. 1996. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence.

[11] Ferreira, M. da C. da S. 2012. *Classificação Hierárquica da Atividade Económica das Empresas a partir de Texto da Web*. Universidade do Porto.

[12] de França, T.C. and Oliveira, J. 2014. Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. *Anais do XXXIV Congresso da Sociedade Brasileira de Computação* (Brasília, 2014), 128–139.

[13] Hassan, S.-U. and Gillani, U.A. 2016. Altmetrics of "altmetrics" using Google Scholar, Twitter, Mendeley, Facebook, Google-plus, CiteULike, Blogs and Wiki. *arXiv:1603.07992 [cs]*. (Mar. 2016).

[14] Jiawei Han, Micheline Kamber and Jian Pei 2011. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

[15] Kwak, H. and Lee, J.G. 2014. Has Much Potential but Biased: Exploring the Scholarly Landscape in Twitter. *Proceedings of the 23rd International Conference on World Wide Web* (New York, NY, USA, 2014), 563–564.

[16] Li, Y.-M. and Li, T.-Y. 2011. Deriving marketing intelligence over microblogs. *Proceedings of the 44th Hawaii International Conference on System Sciences* (2011), 1–10.

[17] Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*. 5, 1 (2012), 1–167.

[18] Lucas, A. de M. 2002. *Utilização de técnicas de mineração de dados considerando aspectos temporais*. Universidade Federal do Rio Grande do Sul.

[19] Maia, L.F.M.P., Costa, R.J.M. and Cruz, S.M.S. 2015. Uma Proposta de Biblioteca Digital de Trabalhos de Conclusão de Curso. *Anais da II Escola Regional de Sistemas de Informação do Rio de Janeiro* (Rio de Janeiro, 2015).

[20] Maia, L.F.M.P., Yagui, M.M.M., Quispe, F.E.M., Oliveira, G.S., Leonardo, J.S. and Cruz, S.M.S. 2014. Combinando Dados de Clickstream e Análise de Redes Sociais Para Identificação do Comportamento Eletrônico dos Petianos da Região Sudeste. *Anais do XIX Encontro Nacional de Grupos do Programa de Educação Tutorial* (Santa Maria, 2014).

[21] Nascimento, P., Aguas, R., De Lima, D., Kong, X., Osiek, B., Xexéo, G. and De Souza, J. 2012. Análise de sentimento de tweets com foco em notícias. *Anais do XXXII Congresso da Sociedade Brasileira de Computação* (Curitiba, 2012), 16–19.

[22] Pak, A. and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (Valletta, 2010), 1320–1326.

[23] Sarlan, A., Nadam, C. and Basri, S. 2014. Twitter sentiment analysis. *Proceedings of the 6th International Conference on Information Technology and Multimedia* (Putrajaya, 2014), 212–216.

[24] Yousefpour, A., Ibrahim, R. and Abdull Hamed, H.N. 2014. A Novel Feature Reduction Method in Sentiment Analysis. *International Journal of Innovative Computing*. 4, 1 (2014).