Data mining of social manifestations in Twitter: Analysis and aspects of the social movement "Bela, recatada e do lar" (Beautiful, demure and housewife)

MARCELA MAYUMI MAURICIO YAGUI[1]
LUÍS FERNANDO MONSORES PASSOS MAIA[1]
JONICE OLIVEIRA[1]
ADRIANA S. VIVACQUA[1]

[1]Graduate Program in Informatics (PPGI)
Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro - RJ - Brazil
[1](marcelayagui,luisfmpm)@ufrj.br, (jonice,avivacqua)@dcc.ufrj.br

Abstract. In recent years, the Online Social Networks (OSN) enabled the growth of many social movements in digital media, because they allow messages to be posted and shared instantly. In the political scope, the OSN are a means by which social groups been assembled themselves and defended their causes. Between April and May 2016, a Brazilian magazine published an article entitled "*Bela, recatada e do lar*" (beautiful, demure and housewife), whose repercussion mobilized several social groups and generated a virtual protest in nationwide scale. The goal of this research was to analyze behavior of users in the social network Twitter to identify how people reacted to the article "*Bela, recatada e do lar*". To achieve this, a network of shared messages (retweets) was built, where the centrality metrics Degree, Betweenness and PageRank were calculated to identify which users most influenced the social movement. Also, a data mining technique known as sentiment analysis was used with the aid of the ETL (Extract, Transform and Load) methodology and the Naïve Bayes probabilistic algorithm to study users behavior and opinion. Furthermore, an analysis of highlighted events was performed from most frequently tweeted hashtags. Results showed that: (i) users that had more influence in the social movement could be split into two main classes: one represented by users with high PageRank values, or in other words, users that published relevant content and were shared extensively by others; and another class represented by users with high Betweenness values, meaning that they acted in an influential manner only inside specific communities. (ii) In its majority, users expressed opinions against the conservative standard for women defended by the magazine article. (iii) Events that occurred in parallel to the social movement "*Bela, recatada e do lar*" apparently influenced the content and amount of published messages in the OSN.

Keywords: Twitter; Data Mining; Social Network Analysis; Sentiment Analysis; Naïve Bayes.

(Received March 20th, 2018 / Accepted May 15th, 2018)

1 Introduction

Online Social Networks (OSN) have profoundly marked human relations: people meet again, relations are strengthened and undone, business are done, companies advertise their stock. Social actors such as parents, grandparents, grandchildren, teachers, students, professionals (and many others) are present and active in interactions mediated by the OSN. A result of this behavior is the occurrence of diverse phenomena related to the interaction between human beings mediated by computers.

OSN enabled the emergence of many social movements in digital media and have been, alone, a driving force of a revolution in human communication. They not only supported the reorganization of family groups and social groups, but also potentialized them. Currently OSN like Twitter[1] strengthen resources of access, dissemination, cooperation, and diffusion of information and knowledge [9].

Additionally, social networks have become an important platform to announce political programs and are also a democratic platform in that social groups have organized themselves and defended their causes [22]. The symmetry of privileges and power of speech make the interaction more democratic and networks have become even more important carriers to minorities, that now see OSN as a means to be heard [19].

Social networks made possible the dissemination of information and ideas in a rapid and 'organic' manner: the same social actors that use their pages to share content are disseminated to other users of their own 'network of contacts'. A common phenomenon is 'viralization': the dissemination of a message to a huge number of users by means of re-transmission of the original message (sharing). This is a phenomenon particularly sought after by groups that wish to mobilize society to a cause [10].

The mobilization of virtual crowds in OSN is a phenomenon that has been studied with the goal of predicting behavior, facts, and events. Bonabeau [7], states that this virtual mobilization is governed by the same principles as real crowds, varying, in contrast, to size, speed and range. These variations combined with the non-structured character of disseminated information and a new linguistic culture, present a challenge for the data scientist that wishes to analyze social interactions, to comprehend and predict social behavior. In this sense, shared content mining on network seems to be a promising tool for comprehending and predicting the online mobilization phenomena.

OSN monitoring is a crucial element in the process of acquiring concepts in user behavior and opinions, as well as a potential form to prevent facts and events. Understanding popular virtual interaction and being able to predict the next movements has practical applications in other fields of knowledge, such as: politics, communication, and scientific marketing [14], it is important to take note of local characteristics, or in other words: events and news that can potentially influence a group (object) of study, its language (catchwords, colloquialisms, slang, etc.), and other characteristics, whose measurement data analysts still find difficult to perform objectively. However, the field has a basic set of metrics used to analyze social networks, as shown by [1, 2, 5, 6, 11, 21].

In this context, a data mining technique known as sentiment analysis [15] is becoming very popular as a monitoring method, in the sense to identify user behavior patterns during popular movements orchestrated in OSN. Sentiment analysis uses automatic mechanisms in which subjective information is extracted from messages written in natural language. Its application allows the creation of structured knowledge databases from non-structured data, extracted from popular virtual environments such as Twitter, to generate prediction models that support decision-making [18, 20].

In the field of Social Network Analysis (SNA), the patterns of information dissemination in OSN are also traceable from the social relations that are made in these virtual environments, more specifically, through dynamic sharing of messages [23]. Mapping these relations is achievable through a set of metrics applied to the graph structures formed from these social structures [2], where nodes represent users and edges represent their connections (sharing of messages). In a OSN like Twitter, the SNA enables tracking of data propagation in specific communities, as well as identification of influential profiles (through retweet network) [27].

The present research aims to generate tendency predictors from user posts collected in OSN, particularly in Twitter. Moreover, contributions are made to the available knowledge about data mining techniques, specifically, for this kind of data, and new datasets are provided among those published in the Portuguese language, which are a scarce resource, given that most datasets are in the English language.

2 Research questions

The goal of this research was to study the repercussion of the virtual protest "*Bela, recatada e do lar*" (Beautiful, demure and housewife) based on an analysis reported in a previous study [26]. In this sense, we sought to identify which profiles most influenced the social manifestations from its posted (and shared) messages in Twitter. We also investigated the temporal relation between events that happened throughout protests and the triggering of post frequency peaks. In this sense, three research questions were explored: (i) What opinions and feelings were most frequent among user profiles regarding the movement triggered on Twitter?; (ii) Which profiles most influenced the movement "*Bela, recatada e do lar*"?; And (iii) What were the most commented topics during the movement?

[1] `https://about.twitter.com/company`

3 Methodology

The analysis process consisted of four stages. Initially, tweets were collected throughout the days 04/22 and 05/03 to create a database that would be analyzed later. Second, a sentiment analysis in the Twitter messages was built, according to its nature in regard to the social movement. Third, a retweet network analysis was undertaken in a way to identify the influential profiles of the movement. Lastly, the construction of word clouds with the most frequent hashtags was performed, whose purpose was to analyze which events were most commented. These stages are further explained in the following sections.

3.1 Data collection and preprocessing of messages

Data collection was performed using the Extract, Transform and Load (ETL) process described in [26]. In this process, between days 04/22 to 05/03, posts of the OSN Twitter were acquired and subsequently preprocessed, where the following content in tweets was eliminated: special characters, punctuations and stop words. Data collection started four days after the beginning of the virtual manifestation and ended after we observed that the virtual movement achieved its critical mass and lost momentum. As a result of this process, a dataset of 43,647 messages related to the social movement "*Bela, Recatada e do Lar*" (in Portuguese language) was generated. For data collection, we used the software tool Node-Red[2] of the IBM service Bluemix[3].

3.2 Sentiment analysis

In this stage, a random sample of 1,500 messages was selected and manually marked with the labels 'positive', 'neutral', or 'negative', as in [26]. From 1,500 labeled posts, 523 were positive, 424 negative, and 553 neutral. The label qualifies how the author of the message manifested toward the virtual protest. Thus, messages labeled as 'positive' are those which expressed agreement with the protest (and, as a consequence, disagreement to the magazine's original article). Messages whose content was in disagreement to the movement were labeled as 'negative'. Construction and publication of this dataset is the main contribution of this paper.

In the following stage, training and testing of a predictor model using an implementation of the Naïve Bayes probabilistic algorithm [16, 20, 26, 27, 28] was carried out using the Python language (version 2.7), aiming to define a baseline for future comparisons.

The training strategy consisted in using the bag of words technique as input data elements. The chosen test mode consisted in separating 1/3 of the labeled messages for testing and the remainder 2/3 of acquired messages for training.

The use of bag of words with Naïve Bayes has the purpose to generate a performance measurement baseline for results comparison with other researchers. The simplicity of the implementation of Naïve Bayes and the bag of words technique (both well documented in the literature) allow studies to be more easily reproduced by another researcher, which also supported in our decision concerning the use of this algorithm for establishing the performance baseline.

This classification method achieved an accuracy of 81%. The result is auspicious, since it is consistent with the human accuracy index, whose ability to analyze text subjectivity ranges from 72% [24] to 85% [12, 17]. Lastly, after the learning algorithm stage, automatic classification of all 43,647 tweets of the database was performed.

3.3 Temporal analysis of tweets frequency

After the previous steps, the next stage was to investigate the relationship between key events that occurred during the manifestations and the behavior of posts with the hashtag #BelaRecatadaEDoLar or with that string. Thus, published news were collected from several newspapers and blogs. The purpose consisted in verifying whether there is a relationship between the media publications and volume of messages posted, especially peaks and declines on the volume. In this investigation two distinct events were selected, one in favor of the movement "*Bela, recatada e do lar*", and another against.

On day 04/23, the first event was held[4]. Groups of women mobilized against the coercion of the conservative view "*Bela, recatada e do lar*" and protested in Brasilia (capital of Brazil) in clear demonstration of support to the feminist movement on Twitter.

On day 04/25, the wife of a well-known Brazilian reverend manifested against the feminist movement on Twitter, starting a campaign against the feminist movement (and in favor of the conservative position expressed in the magazine's article). This fact, however, has drawn publicity (by media) only at the end of day 04/27, spreading across the internet as of 04/28 (selected day for the second event[5]).

In Figure 1 it is possible to observe two peaks in the distribution. These peaks correspond to the dates on

[2] http://nodered.org

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

[4] https://goo.gl/PxV45D

[5] https://goo.gl/8Xqy1f

which the two mentioned events occurred, the feminist protest on 04/23 and the online newscast of a campaign in favor of the conservative stereotype on 04/28.

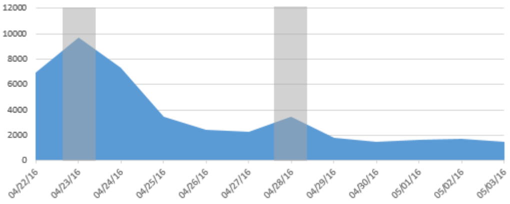

Figure 1: Volume of tweets per day (highlighting the peak days).

3.4 Analysis of the retweet network

For retweet network construction, data previously stored in MongoDB[6] was exported to the CSV format and read in the Gephi[7] tool. Gephi is an open source software that was chosen as an analysis tool because it allows a wide range of network analyses to be performed on large volume of social media data, as well as filtering, manipulating and customizing graphs. In addition, there are several plugins that extend the ability to explore and analyze data [3].

After loading the graph in Gephi, network node centrality metrics were calculated in order to identify the most influential users in the complete network (in the 12 days of the movement) and in key days of the movement. These metrics are:

(i) Degree Centrality - Defined as the number of edges connected to a given node of a network, allowing the Degree of involvement of a node to be analyzed. For directed graphs, the Degree can be divided into In-degrees (number of edges that connect in the input stream of a node) and Out-degrees (number of edges that connect to the output of a node). For non-directed graphs, Degree can be considered as the sum of the In-degrees and Out-degrees [2].

In this research, the Degree metric is equivalent to the number of times a user has been retweeted by another user. The interpretation we give is that a node's retweets frequency indicates its influence on the network.

(ii) Betweenness Centrality - Considering a path between two nodes, the Betweenness metric can be defined as the number of smallest minimum paths passing through this particular path node. They are nodes that can direct the flow of information, whether they facilite, hinder or even transform the data, and are considered node "bridges" between nodes/subgraphs. The importance of nodes with high Betweenness can be justified by the fact that the disconnection of these nodes can cause the partitioning of interconnected subgroups, breaking the network [2].

(iii) PageRank - PageRank is a probabilistic algorithm that quantifies the influence of a node based on the importance of neighboring nodes. It is used for sorting search results, like Google does, for instance, where the best placed pages are the ones with the best connections and the highest probability to be accessed on the network. Thus, each node's rank depends on the network structure and is calculated by simulating a random visitation of nodes. This metric can be interpreted as a good indicator of influence [2]. In this work, the PageRank computation is performed as follows: the probability (p) of a random user to visit a random node (and retweet it) in the retweets network is 0.85, where p is a damping factor ranging from 0 to 1, usually defined as 0.85 [8]. After defining and computing SNA metrics, word clouds were constructed to identify which hashtags were most disseminated during the social movement.

3.5 Construction of the word clouds

The construction of word clouds was carried out with aid of the data mining software Knime.

Knime[8] is a free software designed for data science that can be used to extract, analyze, transform and integrate miscellaneous data types [13]. With support for preloaded modules, discussion groups, and training manuals, this tool holds sundry built-in components that allow creation of ETL processes for data mining, machine learning, statistical analysis, and big data [4]. Knime was used for optimizing the creation of word clouds through an agile processing of the huge text volume present in the tweets.

The process of creating word clouds starts with the insertion of the tweets into the MySQL relational database. From this, data were read, renamed and loaded into the 'Text Processing' module, which processes and transforms the tweet's texts (Figure 2).

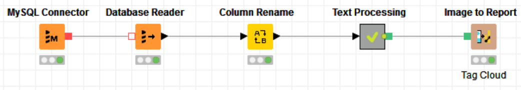

Figure 2: Word clouds creation process.

[6] https://www.mongodb.com/
[7] https://gephi.org/
[8] https://www.knime.org/

To trigger the text processing, each tweet is transformed into arrays of text documents associated to each line's identification. After this transformation, the process executes the 'Extract Hashtags' module, as can be seen in Figure 3.

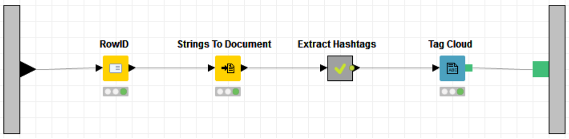

Figure 3: Text processing module.

The Extract Hashtags module is shown in Figure 4. In this module, hashtags are searched in documents by means of a preset regular expression. After the search, hashtags are filtered and inserted into a bag of words along with the document related to them. The process proceeds performing the calculation of relative Term Frequency (TF[9]) of the hashtags present in each document and inserts a new column containing the TF value. After the TF calculation, similar hashtags are counted and sorted in decreasing order.

Right after sorting, the 10 most popular hashtags are filtered, which are associated to their respective data (tweet, user, etc). Hashtags are then transformed into strings and the process returns the word cloud image where the most tweeted hashtags are displayed.

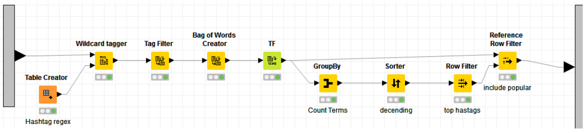

Figure 4: Extract Hashtags module.

4 Results

This section presents the results obtained. Analysis of tweets related to the theme "*Bela, recatada e do lar*" were carried out to identify which profiles had the greatest influence, to assess which events aroused greatest reaction in people and to identify what were the main feelings and opinions issued in Twitter in regard to the news published by the magazine Veja[10].

This analysis was an advancement of the analysis carried out in [26]. As will be shown on the results of the sentiment analysis (section 4.1), the percentage of tweets classified with the neutral feeling (approximately 77%) was very high, not allowing an accurate analysis nor correlation with the results of retweets analysis.

In current analyses, the same base of 43,647 tweets was used, however, this time a larger volume of messages was manually labeled, which generated a more efficient training and testing corpus. As a matter of fact, the results of the sentiment analysis have changed drastically in comparison to the previous study.

4.1 Results of the sentiment analysis

In this section the results of tweets classification by sentiment analysis are presented. The previous study's analysis had a sample of 550 tweets manually classified and presented a result of approximately 10% as positive, 13% as negative and 77% as neutral, as shown in Figure 5.

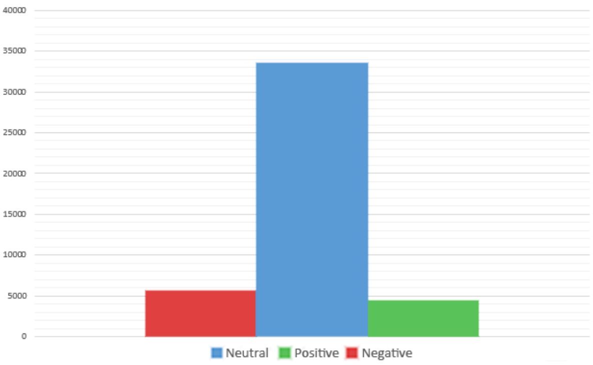

Figure 5: Classified tweets chart (previous study's sentiment analysis).

Because of the high percentage of neutral tweets achieved, a new round of sentiment analysis was performed considering a corpus of 1,500 tweets manually classified. The current results showed that 39% of the tweets were classified as neutral, 35% as positive and 26% as negative (Table 1 and Figure 6).

Table 1: Table of sentiment frequencies.

Sentiment	Frequency	Rel. Freq.	Perc. Freq.	Accum. Freq.
Negative	11,380	0.260728	26.0728	26.0728
Neutral	17,206	0.394208	39.4208	65.4936
Positive	15,216	0.348615	34.8615	100
Total	43,647			

[9] Term Frequency is calculated by dividing the frequency of each hashtag in a document by the total of terms in that document.

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

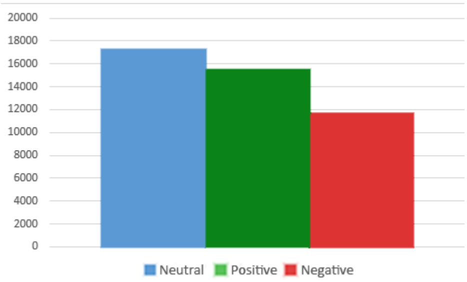

Figure 6: Classified tweets chart (current analysis).

Regarding the total classified, the number of tweets aggregated by 'sentiment per day' is expressed by Table 2 (of frequencies).

Table 2: Table of sentiment frequencies per day.

	Negative	Neutral	Positive	Total
Day 1	1,368	3,134	2,363	6,870
Day 2	2,413	4,114	3,199	9,726
Day 3	1,822	3,214	2,207	7.243
Day 4	1,006	1,704	950	3,460
Day 5	712	1,014	715	2,441
Day 6	641	359	1,192	2,192
Day 7	1,058	806	1,582	3,446
Day 8	498	181	1,125	1,804
Day 9	401	604	437	1,442
Day 10	519	656	430	1,605
Day 11	534	793	575	1,942
Day 12	408	627	441	1,476
Total	11,380	17,206	15,216	43,647

These values are also demonstrated graphically on Figure 7.

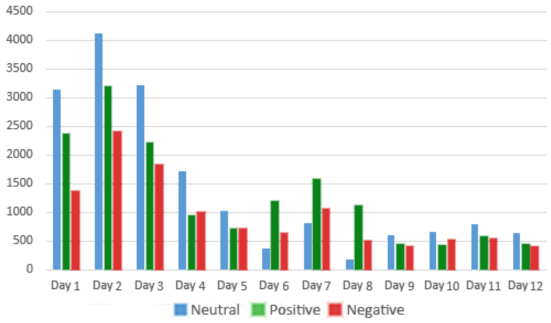

Figure 7: Sentiments classified per day.

Figure 8 correlates geographic distribution (performed in the previous study) with sentiments. From the chart, it can be observed that the city with the highest percentage of tweets with the negative sentiment is Salvador, with 41%. On the other hand, the highest rate of neutral tweets is 61%, and belongs to the city of Recife. Lastly, the highest proportion of positives is in the city of Brasilia with 52%.

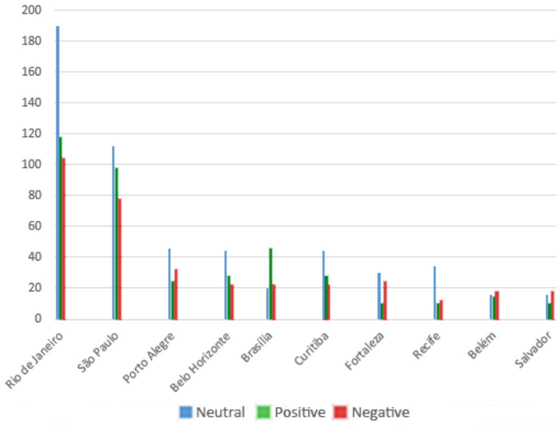

Figure 8: Classified sentiments in cities that most published.

According to the new results, neutral is still the dominant sentiment (17,206), although the volume of neutrals has reduced significantly in comparison to the old analysis (33,579). In contrast, unlike the previous analysis, where tweets with negative feeling (5,599) were slightly higher than positive ones (4,469), the current analysis shows predominance of positive tweets (15,216) in comparison to negative tweets (11,380). Therefore, an improvement in the manual labeling process led to a decrease in the amount of neutrals, which caused an increase in the difference between positive and negative quota (1,130 in the previous analysis and 3,836 in the current analysis).

Different results were also observed in the ranking of most shared messages (retweets). In this ranking, three of ten most retweeted messages have the positive sentiment attached, totaling 4,404 positive tweets among the most disseminated ones. Two of them have the negative sentiment (4,045 negative tweets) and most of them do not have any sentiments attached (4,707 neutral tweets). Table 3 expresses these results.

Based on these numbers it is possible to answer the first question of the research: What opinions and feelings were most frequent among users' profiles regarding the movement triggered on Twitter? The answer is that, if we disregard messages classified as neutral, where lack of information prevents the identification of a feeling attached to the tweet (tweets whose content are links, photos, images, etc.), positive opinions, ergo, supporting the movement and opposing Veja magazine's article were more frequent.

Table 3: Most retweeted messages.

Publisher	Original tweet	Tweet Translation	Retweeted	Sentiment
luscas	n aguento mais bela recatada e do lar em legenda de fotos façam parar	I can't stand the phrase "beautiful, demure and housewife" in photo caption anymore, make it stop.	3,081	Negative
mc_caroloficial	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	The only difference between me and Marcela is that, here at home, I AM THE PRESIDENT and my boyfriend is the first lady hahahahahaha #belarecatadaedolar	2,073	Positive
mc_caroloficial	Eu e essa tal de Marcela Temer somos TAO PARECIDAS kkk #belarecatadaedolar #sqn https://t.co/Qb13vo82VS	Me and this girl Marcela Temer are SO SIMILAR hahaha #belarecatadaedolar #sqn[11] https://t.co/Qb13vo82VS	1,494	Positive
jeantissociall	bela, recatada e do lar https://t.co/nzvN89I9aZ	beautifull, demure and housewife https://t.co/nzvN89I9aZ	1,296	Neutral
jose_simao	Veja traça perfil de Marcela Temer: "bela, recatada e do lar". Voltamos aos anos 60!	Veja delineates the profile of Marcela Temer[12]: "beautiful, demure and housewife". We are back to the 60's again!"	964	Negative
luanlovato	bela , recatada e do lar https://t.co/krI800zwUn	beautiful, demure and housewife https://t.co/krI800zwUn	934	Neutral
lucianagenro	Bela, recatada e do lar. https://t.co/7vhmlevEG7	Beautiful, demure and housewife https://t.co/7vhmlevEG7	903	Neutral
rockinrio	Enquanto a gente tem o Palco Mundo, as mulheres têm o mundo como palco ;) #belarecatadaedolar https://t.co/S1VUaCFbf9	While we have 'Palco Mundo'[13], women have the world as stage ;) #belarecatadaedolar https://t.co/S1VUaCFbf9	837	Positive
potterish	Bela, recatada e do lar. https://t.co/0LVBcpvCQI	Beautiful, demure and housewife. https://t.co/0LVBcpvCQI	789	Neutral
gifsdegatinhos	bela, recatada e do lar https://t.co/SOK9LRWIJb	beautiful, demure and housewife https://t.co/SOK9LRWIJb	785	Neutral

4.2 Results of retweet network analysis

In this section, we try to answer the second research question: Which profiles most influenced the movement "*Bela, recatada e do lar*"? To answer this question, we analyzed the retweets network using Gephi.

Figure 9 shows the non-directed retweets graph created from the 12 days of collection, composed of 21,246 nodes and 20,026 edges, where: (i) pink nodes are users who published the original message and were retweeted, representing 16.54% of the total; (ii) blue nodes are the users who retweeted the messages, representing 81.47% of nodes and; (iii) green nodes are users whom this relationship occurs in both cases (bidirectional edges), and hence, falling under items (i) and (ii) representing 1.99% of users.

From the network, the three centrality metrics (Degree, Betweenness and PageRank) were calculated to identify which users most influenced the movement.

In this study, Degree is the number of times a user

[11] Brazilian sarcastic slang.

[12] She is the first lady of Brazil and granted an interview to the magazine Veja. The article based on the interview caused controversy, which led to the virtual protest.

[13] 'World Stage'. Main stage at the Brazilian music festival 'Rock in Rio'.

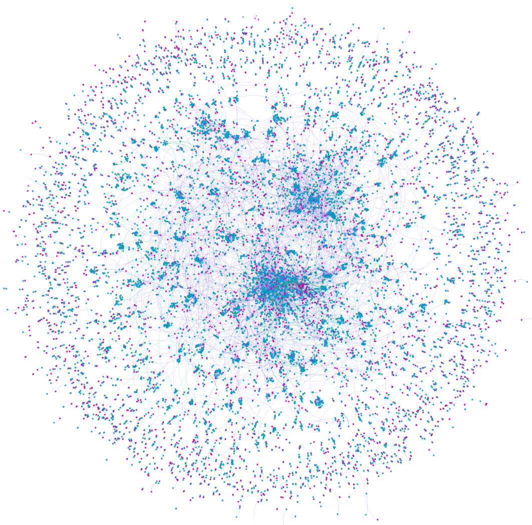

Figure 9: Retweets network of the social movement "*Bela, recatada e do lar*".

Table 4: Degree of top 10 users.

User	Degree
BolsonaroZuero	454
ClaudiaLeitte	392
Antoniotabet	374
Gordaopvcs	351
Frasesdebebadas	340
MidiaNINJA	284
Leo_oguarda	264
HugoGloss	237
luscas	224
iLovePut*****	222

was retweeted by another user, implying that the more a profile has been retweeted, the bigger the influence it has on the network. Table 4 shows top 10 users according to Degree. It can be observed that this result covers a variety of account types, including humorous profiles, singers, bloggers, among others.

Regarding the Betweenness metric, users that have highest values have greater probabilities as to influence neighboring nodes. In other words, from this metric it is possible to understand how information flows through the network, and observe which profiles connect communities. Table 5 shows top 10 users according to Betweenness.

Table 5: Betweenness of top 10 users.

User	Betweenness
andandrea	0.17524
Leo_oguarda	0.06613
gordaopvcs	0.05899
gifsdegatinhos	0.04990
BolsonaroZuero	0.04389
antoniotabet	0.04309
MidiaNINJA	0.04088
maarinolasco	0.04045
luscas	0.03836
nadiardgs	0.03680

For PageRank, nodes with the highest values can be considered key users on the network because other important users interact with them. In other words, it is likely that these users are highly active in the OSN and possibly are recognized as important profiles by other users. Table 6 shows the top 10 profiles according to PageRank.

Table 6: PageRank of top 10 users.

User	PageRank
BolsonaroZuero	0.00866
ClaudiaLeitte	0.00804
Gordaopvcs	0.00725
antoniotabet	0.00724
frasesdebebadas	0.00714
Leo_oguarda	0.00544
HugoGloss	0.00500
MidiaNINJA	0.00497
iLovePut*****	0.00472
Falandocarioca	0.00463

According to Willis et al. [25], key users that belong to a movement can be inferred from the correlation of two metrics (in this case, Betweenness and PageRank) through statistical methods. That said, based on the methodology defined by [25], Figure 10 presents the scatter plot created, where the X and Y axis represents, respectively, the metrics Betweenness and PageRank. Each point in the diagram represents a profile that is within the interval of the 60 most retweeted.

Each of the diagram's four quadrants has a particular interpretation that defines how each user exerts his influence: (i) the lower left quadrant (low PageRank and low Betweenness) has common users since the profiles contained therein tend to have no specific role; (ii) the upper left quadrant (high PageRank and low Betweenness) has users who are likely to exert influence by starting discussions and sharing tweets that others will retweet. This is justified because they tend to be located in one of the nuclei of the network; (iii) the lower right quadrant (low PageRank and high Betweenness) has users that are important to a particular public/community, as they are considered bridges be-

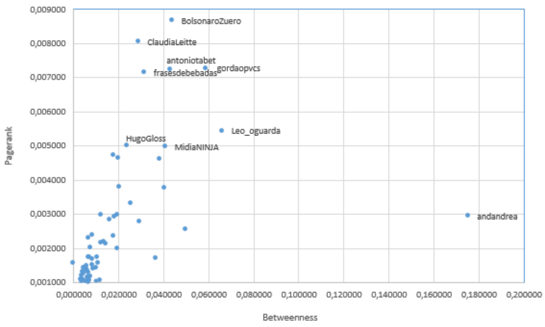

Figure 10: Scatter plot of 60 most retweeted users.

tween content production and the audience in which they are connecting; (iv) lastly, the upper right quadrant (high PageRank and high Betweenness) has users who share characteristics of (ii) and (iii) combined. They are rare profiles that have a strong influence in the online community. According to Figure 10, one can visually identify the movement's key users. The user @andandrea is the only one that fits in item (iii) indicating its influence within specific communities. On the other hand, the profiles @BolsonaroZuero, @ClaudiaLeitte, @antoniotabet, @gordaopvcs, @frasesdebebadas, @HugoGloss, @Leo_oguarda, and @MidiaNINJA have characteristics of item (ii) and represent influential users in regard to content publishing.

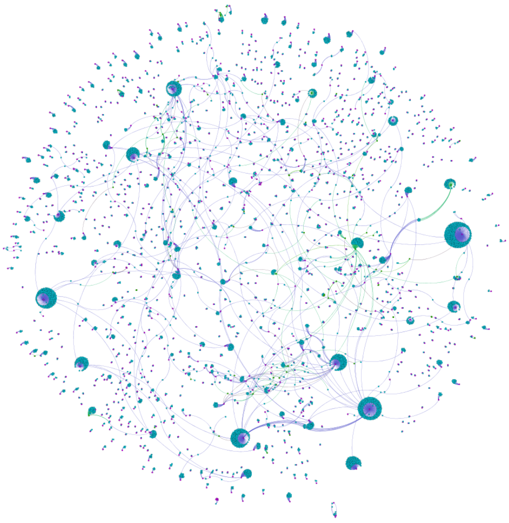

Figure 11: Retweets network of day 04/23.

The same analyses were performed to the days 04/23 and 04/28 (days of the mentioned events), in order to identify which were the key users in those days and establish correlation between events and influential users.

Figure 11 shows network status in 04/23. 6,179 nodes and 5,412 edges are marked, distributed among the following profile types: users who retweeted (blue nodes), representing 80.74%, users who initially posted tweets (pink nodes), representing 18.13%, and users who did both (green nodes), representing 1.13%.

Yet in regard to the metrics, Degree, Betweenness, and PageRank, they were also calculated separately for day 04/23. Table 7 shows the top 10 users (most retweeted) on day 04/23 according to Degree.

Table 7: Degree of top 10 users on day 04/23.

User	Degree
frasesdebebadas	313
BolsonaroZuero	253
Leo_oguarda	196
antoniotabet	170
ronanoliveira_	139
MidiaNINJA	118
MafiaLMonster	104
luscas	92
thedebnam	88
nadiardgs	87

Table 8 shows the top 10 users on day 04/23 according to Betweenness.

Table 8: Betweenness of top 10 users on day 04/23.

User	Betweenness
Leo_oguarda	0.12313
bemvindo_ator	0.07183
gordaopvcs	0.07065
antoniotabet	0.06899
prazerpv	0.06225
j_livres	0.05634
Ligiapfeffer	0.05558
Hstylessucker	0.05393
PerrysOverdose	0.05362
_ThomasConti	0.05356

Table 9 shows the top 10 users on day 04/23 according to PageRank.

The same interpretation shown in Figure 10 can be made to correlate the analyzed metrics. The scatter plot in Figure 12 shows the main profiles of the second day's collection, day of the feminist protest (04/23) that took place in Brasilia. The users @frasesdebebadas, @BolsonaroZuero, and @antoniotabet are sorted to item (ii) high PageRank and low Betweenness, being considered as influential nodes regarding the content publishing, and @Leo_oguarda was sorted to item (iv) high PageRank and high Betweenness, thus was influential within

Table 9: PageRank of top 10 users on day 04/23.

User	PageRank
frasesdebebadas	0.02318
BolsonaroZuero	0.01809
Leo_oguarda	0.01441
antoniotabet	0.01202
ronanoliveira_	0.00936
MidiaNINJA	0.00815
MafiaLMonster	0.00778
Luscas	0.00669
Thedebnam	0.00647
Nadiardgs	0.00597

the communities and in regard to the content publishing.

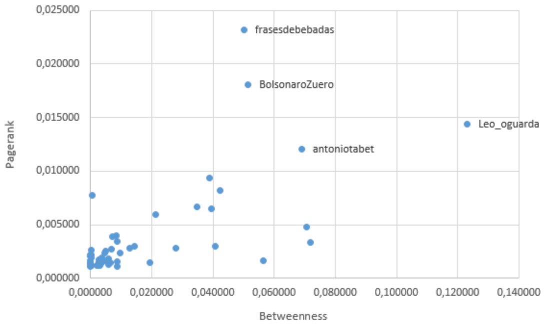

Figure 12: Scatter plot of day 04/23.

Once more, the previous procedure was performed for 04/28.

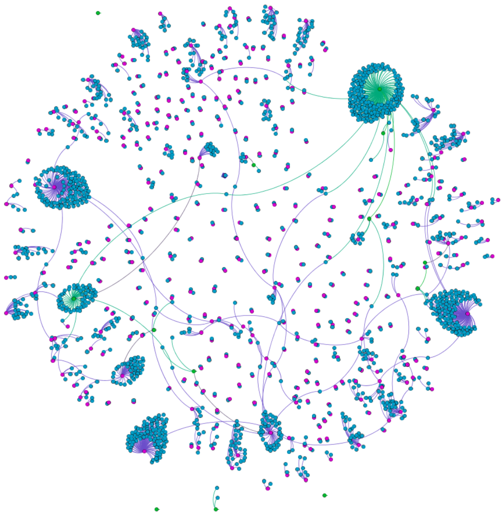

Figure 13: Retweets network of day 04/28.

The graph of Figure 13 has 1,933 nodes, 1,612 edges, where 8.52% are publishers (pink nodes), 80.9% are retweeters (blue nodes) and 0.83% published and retweeted messages (green nodes).

Table 10 shows the top 10 users (most directly retweeted) on day 04/28 according to Degree.

Table 10: Degree of top 10 users on day 04/28.

User	Degree
AnaVilarino1	178
ptadeu_pt	122
implicante_org	120
icrinvel	95
Sofimagal	57
JornalismoWando	46
Sybylla_	43
siteRDTPop	27
mariasilviaeduc	23
FreakOne_	22

Table 11 shows the top 10 users on day 04/28 according to Betweenness.

Table 11: Betweenness of top 10 users on day 04/28.

User	Betweenness
AnaVilarino1	0.04512
ptadeu_pt	0.02784
Sofimagal	0.01707
jcotrin	0.01323
JornalismoWando	0.01166
robsongfreire	0.01047
GolpeNuncaMais	0.01046
crispino	0.00685
implicante_org	0.00529
leitao_patricia	0.00453

Table 12 shows the top 10 users on day 04/28 according to PageRank.

Table 12: PageRank of top 10 users on day 04/28.

User	PageRank
AnaVilarino1	0.03936
ptadeu_pt	0.02761
implicante_org	0.02752
icrinvel	0.02206
Sofimagal	0.01295
Sybylla_	0.01006
JornalismoWando	0.01005
siteRDTPop	0.00649
mariasilviaeduc	0.00546
FreakOne_	0.00534

Lastly, the scatter plot (Figure 14) was created to identify key profiles on 04/28, day in which the wife of a famous reverend started the Twitter campaign in favor of the view "*Bela, recatada e do lar*". As can be seen in the diagram, four users stood out on the observed day. The profiles @implicante_org and @icrinvel can be interpreted as elements of item (ii) (high PageRank and low Betweenness), being relevant in content production. Also, the users @pt_tadeu and @AnaVilarino1 were sorted to item (iv) (high PageRank and high Betweenness), which means that they are considered important nodes and generally exert strong influence in the publication of content and in the community.

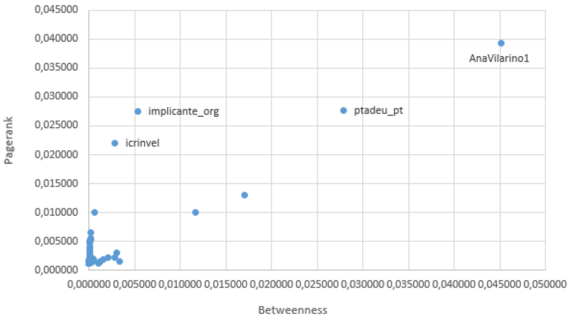

Figure 14: Scatter plot of day 04/28.

4.3 Word clouds results

The last analyses consists on the construction of words clouds that aim to answer the third research question: What were the most commented topics during the movement?

Through this method it was possible to identify which hashtags were most frequently disseminated during the movement. Taking this into account, we built three word clouds, one from the whole dataset (12 days), one from day 04/23 and one from day 04/28 (days of the potentially influential events), disregarding #BelaRecatadaEDoLar as it is a reserved hashtag in this study.

Figure 15 shows the word cloud created from the most used hashtags in the movement. Besides the hashtag #belarecatadaedolar, several others stood out regarding the usage frequency and can be useful to figure out new secondary events that may also have influenced content and volume of tweets.

The hashtag #SOSCoup (beginning on 04/21) was a topic against the impeachment process of Dilma Rousseff (Brazilian president from 2011 to 2016).

#BTSisonFIRE refers to the release (on 05/01) of the videoclipe ’Fire’ from the South Korean band Bang-Tan Boys (BTS).

#askmagcult is a topic created to ask and answer questions about an event called Magcon (beginning not identified).

#LEMONADE is related to the release of the sixth album of the singer Beyoncé (on 04/23).

Another trend is #VEDA, which means "Vlog Everyday in April", that is, when using the hashtag, each vlogger commits to publish one video per day in the month of April.

#METGala is an annual fundraising event for the Metropolitan Museum of Art in New York that brings together a number of today’s celebrities and artists.

Other hashtags related to the social movement "*Bela, recatada e do lar*" also stood out, such as: #bela (stands for ’beautiful’), #recatada (stands for ’demure’), and #mulhernaogostadehomemque (stands for ’girls don’t like boys who’).

Furthermore, the hashtag #sqn (slang to ’*só que não*’, which stands for ’said no one ever’) also stood out, showing a hint of sarcasm on tweets related to the conservative standard raised by the magazine Veja.

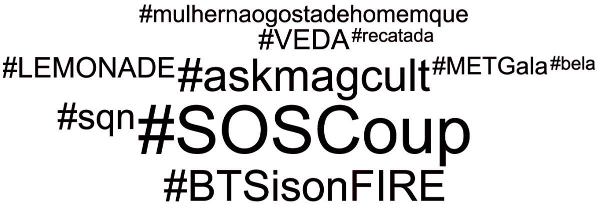

Figure 15: Word cloud of the social movement "*Bela, recatada e do lar*".

Concerning the word cloud of day 04/23 (Figure 16), the following hashtags stood out: Zé de Abreu (Brazilian artist) spat on a couple after a discussion about politics at a restaurant in São Paulo on the night of 04/22. Since then, several users on Twitter spoke against the actor and asked for his resignation, using the hashtag #DemiteoZehdeAbreu (’fire Zé de Abreu’).

#EtaMundoBom, a brazilian soap opera that aired from 01/18 to 08/26 of 2016 also appeared associated to the analyzed tweets.

#youngmaland, a popular hashtag in the OSN Instagram and that refers to selfies, mostly of cosplay, and taken at the South Korean theme park Youngma Land, is also associated.

The hashtags #vejamachista (stands for ’Veja magazine is male chauvinist’), #bela, and #recatada, related to the social movement were also used on published tweets. The hashtag #VEDA, along with the sarcastic hashtag #sqn, also stood out on day 04/23.

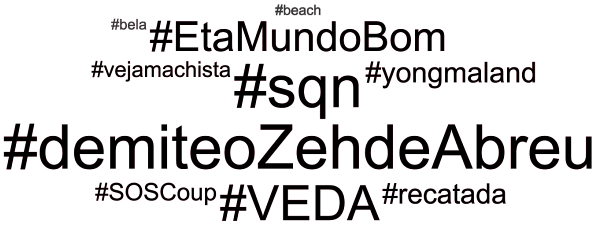

Figure 16: Word cloud of day 04/23.

On day 04/28, the newspaper profile #OGlobo was mentioned, along with the following hashtags related to social movement: #veja, #mulherperfeita ('perfect woman'), #bela, #MENTIRA ('lie'), #feminista ('feminist'), #luta ('fight'), and #marcelatemer (hashtag directed to Marcela Temer).

Hashtags #VEDA and #sqn were also featured on day 04/28, being widely used in tweets related to the social movement, as can be seen in Figure 17.

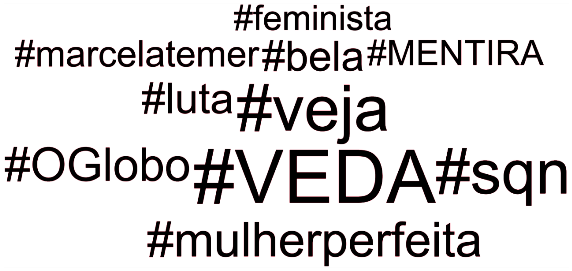

Figure 17: Word cloud of day 04/28.

In general, most hashtags refer to topics that are not directly related to Twitter protests. Yet on days of the events most of the hashtags, in fact, were related to the movement. This reinforces the hypothesis of the previous study [26], that the two events (on days 04/23 and 04/28) caused peaks on tweets frequency and influenced users opinions.

As we could see, the hashtag #sqn stood out throughout the analyzed period (including the days of both events). Therefore, there is a strong indication of sarcasm in tweets directed to the conservative standard raised by the magazine Veja.

The hashtag #VEDA, similarly, also gained prominence during the period when the virtual protest took place. For tweets that used this hashtag, it was noticed that messages contained only a link to a Youtube video.

Events like those previously described seem to be directly related to the high number of neutrals identified in sentiment analysis.

5 Related work

As mentioned in [15, 18, 20], sentiment analysis refers to a wide field of text mining, natural language processing and computational linguistics that comprises the computational and semantic study of sentiments, opinions, and emotions expressed in texts.

Through this method, it is possible to identify opinions of positive or negative polarity in many kinds of text, for instance, web documents or OSN posts.

As reported in [20], this kind of text mining has been becoming quite popular in recent years and seems to present better results when using the Naïve Bayes classification method, especially when employed in small texts such as those found in the OSN Twitter. The works of [15, 26], applied the Naïve Bayes method to analyze user sentiments in the OSN Twitter successfully. In [15] the classification method used a dataset of 216 manually marked messages to train the algorithm. In [26] this dataset was of 550 manually marked messages. The present work used the same classification methodology as [15, 26], however, our method consisted in a dataset of 1,500 marked messages, which guaranteed better results.

Regarding the analysis of influential profiles in the OSN Twitter, the work of Willis et al. [25] identified how networks of influence are formed among users who published about the 2012 London Olympics. Basically, the authors used the metrics Betweenness and PageRank and calculated the correlation between these two variables in order to identify what kind of influence users exerted alongside the Olympic Games (e.g. influence on content or in specific communities), and how intense was that influence. It was also possible to evaluate the influence of individual tweets and how long these tweets remained active after their publication. As a result, key users and key social interactions of the event were identified. Corporate profiles shaped the network and affected its connectivity and personal profiles increased interactions and discussions in the network.

The authors Weitzel et al. [23] analyzed the retweets network in the health domain under the perspective of topological structure and ego networks, with the aid of a nodes' classification mechanism based on retweet weights. Studying several network centrality metrics, the authors concluded that measures such as PageRank indicate a user's popularity and measures such as Betweenness indicate the key position of a node within the network.

The present work used an influence detection method similar to that of Willis et al. [25] and supported by the idea of Weitzel et al. [23], applying scatter plots on the retweets network to find key users through the correlation among PageRank and Betweenness. Unlike the work of Willis et al. [25] who studied an event of global scale such as the Olympic Games, our work focused on studying the interactions on a virtual protest of smaller scale, in a shorter time span and in Portuguese language, which may have narrowed the results. Nevertheless, while the related works used a single approach to map opinions or profiles engagement in Twitter, the present work combines three different approaches, allowing us to map three kinds of influencing factors: opinions, profiles and topics, in a combined way.

6 Discussion and conclusions

As previously mentioned, this work consists of an extension of an earlier study [26], whose main objective was to analyze the impacts of an article published by the magazine Veja in the OSN Twitter. The "experimental design" of the original study was kept in order to analyze the temporal relationship between facts and events that occurred during the social manifestations and tweet frequency peaks for over two weeks, highlighting two events, one against (day 04/23) and another in favor (04/28) of the conservative women standard.

In this manner, we sought to carry out complementary analyses to identify which profiles had the greatest influence and which topics had most "viralized" during the period in which the social movement remained active on Twitter. Also, a new sentiment analysis was performed, since the results of the analysis performed in the previous study were insufficient to answer the considered hypothesis.

Based on these precepts, this new study tried to answer three research inquiries: (i) "What opinions and feelings were most frequent among users' profiles regarding the movement triggered on Twitter?"; (ii) "Which profiles most influenced the movement "*Bela, recatada e do lar*"?" and (iii) "What were the most commented topics during the movement?"

To answer the first question, the new sentiment analysis was performed, where a new manual classification process labeled 1,500 tweets, enhancing the training and test corpus used by the algorithm. This significantly improved the results of the sentiment analysis, reducing the number of neutrals from 33,579 (previous analysis) to 17,206 (current analysis). The number of tweets classified by the algorithm as positive was 15,216 and as negative 11,380. Also, according to the retweets ranking (Table 3), shared messages with positive feeling were more frequent throughout the protests in Twitter. That is, ignoring the tweets classified as neutral by the algorithm, or, in other words, messages without content or that only contains links, photos, images, among others, we observed the predominance of positive opinions on the movement (in favor) and against to the conservative view presented by the magazine.

To answer the second question, a retweet network was built in the Gephi software and the Degree, Betweenness and PageRank centrality measures were applied in order to identify key users of the movement. These profiles are represented by nodes that are usually very active in the network, that initiate discussions and publish content that is very much shared by other users (Degree and PageRank), or even play an important role in disseminating content, being bridges between clusters or specific communities in the network (Betweenness).

Therefore, according to criteria defined in the work of [25], and through correlation between the two analyzed metrics (PageRank and Betweenness), it was verified that: (i) Over the course of two weeks (total period analyzed), the profile @andandrea exerted influence within specific communities and the users @BolsonaroZuero, @ClaudiaLeitte, @antoniotabet, @gordaopvcs, @frasesdebebadas, @HugoGloss, @Leo_oguarda, and @MidiaNINJA were characterized by influence related to publishing of content, fact supported by the scatter plot in Figure 10.

On day 04/23, the profiles @frasesdebebadas, @BolsonaroZuero and @antoniotabet were relevant in the creation of content. On the other hand, the profile @Leo_oguarda was sorted as influential within some communities and in regard to the content publishing, being a key user on the analyzed day, fact supported by the scatter plot of Figure 12.

Lastly, on day 04/28, the profiles @implicante_org and @icrinvel were also considered to be influencers of content. On the other hand, the profiles @pt_tadeu and @AnaVilarino1 are rare profiles that exerted great influence during the analyzed day, initiating discussions and being highlighted in publication of content and its dissemination to the communities of Twitter, fact supported by the scatter plot of Figure 14.

To answer the third question, word clouds were built. From them it was possible to identify the most frequently published topics (hashtags) over the two weeks and in the peak days, where several secondary events that impacted the volume of tweets were observed.

These analyses demonstrated that many of the featured topics in the analyzed period were not related to social movement. However, on days of the highlighted events, most of the main topics have a direct relation with the protests and could have caused peaks in the tweets frequency, also influencing users' opinions as stated in the hypothesis of [26].

The results of some combined analyses also reveal other interesting data. For example, the hashtags #sqn and #VEDA were trends throughout the period of the virtual protest. In the case of the hashtag #VEDA, these were tweets where the user only inserted a link to Youtube in the message. The hashtag #sqn, almost always associated with sarcastic tweets, often also contains only a link (usually associated to an image). Example: `https://goo.gl/MAFeGm`.

Combined with sentiment analysis, results like this may explain the high number of neutrals identified by the algorithm.

Based on these results, there are strong indications that several smaller events have caused impacts during this popular movement orchestrated in Twitter.

Thus, it was possible to analyze profiles that influenced different groups of users and profiles that started discussion topics and most disseminated content, which were identified in section 4.2. It was also possible to observe that secondary events occurred during the social movement apparently generated oscillations in the volume of messages, as shown in section 4.3. In addition, among the opinions expressed by the users, most of them were neutral (39%), followed by positive (35%) and negative (26%). The present study also showed that, combined, SNA techniques, data mining, and statistical analyses are more effective on identifying behavioral patterns in popular manifestations of social media than individually.

As a main contribution of this work, a new dataset is offered to the scientific community on the Portuguese language. The Dataset was manually labeled and a performance baseline was established using Naïve Bayes. As a result, it was possible to better understand the users' behavior, the relationships raised and the boundary between popular topics in Twitter and the opinions expressed by users.

For future works, we expect to create a retweet network with directional edges. The inclusion of the categories sarcasm and emoticons is also considered to assist in the creation of more effective sets for training and testing.

References

[1] Balamurali, A. R., Joshi, A., and Bhattacharyya, P. Harnessing wordnet senses for supervised sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1081–1091. Association for Computational Linguistics, 2011.

[2] Barrat, A., Barthélemy, M., and Vespignani, A. *Dynamical Processes on Complex Networks*. Cambridge University Press, Oct. 2008. Google-Books-ID: TmgePn9uQD4C.

[3] Bastian, M., Heymann, S., Jacomy, M., and others. Gephi: an open source software for exploring and manipulating networks. *Proceedings of the Third International ICWSM Conference*, 8:361–362, 2009.

[4] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31, 2009.

[5] Bollen, J., Mao, H., and Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, volume 11, pages 450–453, 2011.

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

[7] Bonabeau, E. The perils of the imitation age. *Harvard Business Review*, 82(6):45–54, 2004.

[8] Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[9] Castells, M. Communication, power and counter-power in the network society. *International journal of communication*, 1(1):29, 2007.

[10] Castells, M. *Networks of outrage and hope: social movements in the Internet Age*. Polity Press, Cambridge, UK ; Malden, MA, 2012.

[11] Davidov, D., Tsur, O., and Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 241–249. Association for Computational Linguistics, 2010.

[12] Golden, P. Write here, write now. *Research Live*, May 2011.

[13] KNIME. KNIME | About KNIME, 2017.

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

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

[16] Pak, A. and Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, volume 10, pages 1320–1326, Valletta, 2010.

[17] Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

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

[19] Stieglitz, S. and Dang-Xuan, L. Political Communication and Influence through Microblogging–An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior. pages 3500–3509. IEEE, Jan. 2012.

[20] Vinodhini, G. and Chandrasekaran, R. M. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

[21] Wasserman, S. and Faust, K. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[22] Wattal, S., Schuff, D., Mandviwalla, M., and Williams, C. B. Web 2.0 and politics: the 2008 US presidential election and an e-politics research agenda. *MIS quarterly*, pages 669–688, 2010.

[23] Weitzel, L., Quaresma, P., and de Oliveira, J. P. M. Measuring node importance on twitter microblogging. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 11. ACM, 2012.

[24] Wiebe, J., Wilson, T., and Cardie, C. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

[25] Willis, A., Fisher, A., and Lvov, I. Mapping networks of influence: tracking Twitter conversations through time and space. *Participations: Journal of Audience & Reception Studies*, 12(1):494–530, 2015.

[26] Yagui, M. M. M. and Maia, L. F. M. P. Data mining of social manifestations in Twitter: An ETL approach focused on sentiment analysis. In *Proceedings of the XIII Brazilian Symposium on Information Systems*, Lavras, 2017.

[27] Yagui, M. M. M., Maia, L. F. M. P., Ugulino, W., Vivacqua, A., and Oliveira, J. "Bela, Recatada e do Lar": Base de Dados e Aspectos do Movimento Social Ocorrido na Rede Social Online Twitter. In *Proceedings of the XXXVII Congress of the Brazilian Computer Society: #ComputaçãoParaTudoeParaTod*s*, pages 1585–1594, São Paulo, 2017.

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