A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media

Douglas Cirqueira*, Márcia Pinheiro*, Antonio Jacob Jr.†, Fábio Lobato‡ and Ádamo Santana*
*Institute of Technology
Federal University of Pará
Belém, Brazil
{douglas.cirqueira, marcia.pinheiro}@itec.ufpa.br, adamo@ufpa.br
†Technological Sciences Center
State University of Maranhão
São Luís, Brazil
antoniojunior@professor.uema.br
‡Engineering and Geosciences Institute
Federal University of Western Pará
Santarém, Brazil
fabio.lobato@ufopa.edu.br

***Abstract*—Online Social Networks have been increasingly adopted by web users interested in sharing their opinions and thoughts about restaurants, bars, and products they have visited or bought. This scenario enables new analyses to companies and institutions that seek information on how their audience perceives them, and which aspects should be improved. One technique widely used in this type of study is Sentiment Analysis (SA), which allows the automatic mining of opinions on a particular topic. However, this approach faces challenges in social networks, due to the informal nature of the posts and the lack of attention to the grammatical rules found on user-generated content. In this context, this paper presents a literature review about methods and techniques used in the preprocessing of social media data for SA, in the context of Brazilian Portuguese. The results highlight some gaps in the literature and research possibilities, mainly to increase the accuracy of analyses for those platforms.**

***Index Terms*—Sentiment Analysis, Preprocessing, Text Mining, Data Mining, Social Media**

I. Introduction

The distance between customers and companies has been significantly reduced in the last years. The big social media data generated daily on Online Social Networks (OSN) provides a new landscape of opportunities for enterprises, which have a higher chance of influencing their fans on the media, and understanding their feedback for improving offers [1]–[4].

In face of these facts, academia and industry have been proposing techniques to deal with that scenario, including analytics tools in Social Customer Relationship Management (Social CRM) [2], [5]. Among such technologies, there is Sentiment Analysis [6].

All this context is even more intense in Brazil, the country known as the social media capital of the universe [7], given that the Brazilian population is the second most active on OSN in the world [8]. However, it is still noticeable a lack of Text Mining tools to preprocess data for SA tasks for the Portuguese language and context, without the translation issue [9], [10].

Therefore, this paper presents a literature review regarding preprocessing for SA in Brazilian Social Media.

II. Background

Sentiment Analysis stands for a text mining approach that aims to detect the polarity of a given document automatically, usually towards a positive, negative or neutral valence [6]. Preprocessing in SA and TM aims to treat and select the best features from a dataset for further mining of sentiment information in the data [11].

The user-generated content (UGC) from OSN is ungrammatical and informal in nature, which imposes challenges for preprocessing tasks [12]. For instance, treatment of emoticons with specific polarities, such as smileys or sad faces in a text [13], or the handling of lexico-semantic features such as Part-of-Speech tags [14].

The existing reviews for the Portuguese language scenario regarding TM or SA have been focused on detecting the most adopted algorithms and data sources, such as [15], [16]; revealing techniques and tools applied [9]; illustrating national research groups in the field [17]; and presenting the main datasets and methodologies for TM in Portuguese [18].

Therefore, although it was possible to identify reviews in SA and on the Portuguese context, no previous work was identified aiming to explore the preprocessing literature for SA, in the context of Brazilian Portuguese Social Media, which is the primary focus of this proposal.

III. Literature Review

The literature review has followed a systematic mapping process similar to [9], and inspired by [19]. The methodology is composed of the following steps: 1) definition of research questions; 2) literature search; 3) selection of papers based on inclusion criteria; 4) selection of papers based on quality criteria; 5) information extraction and mapping of papers.

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00008

The research questions raised to conduct the literature search were:

1) What are the most used preprocessing steps for SA in Brazilian Portuguese Social Media?
2) What are the main methodologies to implement preprocessing steps for SA in Brazilian Portuguese Social Media?

An automatic search was performed, based on search queries in English[1] and Portuguese[2] to increase the coverage of studies.

Table I illustrates the scientific databases adopted and numbers regarding papers retrieved. The search period has covered papers from 2012 to 2018. The search queries applied have provided a total of 4,049 papers.

Given the massive number of works retrieved, an inclusion filter was applied aiming to retrieve only scientific papers in Portuguese or English, published in official conference proceedings, including PhD plus Master theses focused on Brazilian Portuguese language and datasets from Social Media. Then, duplicates were removed and it was retrieved a total of 131 potentially relevant articles.

Finally, the quality filter was applied. This filter is defined as the selection of proposals only if they have performed some preprocessing before the SA application, which has provided a total of 61 relevant articles[3] to the synthesis of the literature.

TABLE I
SEARCH STATISTICS FROM THE DATABASES.

Search Database	English	Portuguese	Selected	Relevant
Science Direct	24	0	4	2
Scopus	463	7	62	22
Scielo	3	2	2	1
Capes	236	4	7	2
Google Scholar	2950	360	56	34
Total	3676	373	131	61

IV. RESULTS

A. What are the most used preprocessing steps for SA in Brazilian Portuguese Social Media?

Figure 1 presents the percentage usage for the top 25 steps most employed in the literature, out of 31 individual preprocessing steps found in total. Some of them are already well-known in Text Mining in general, such as tokenization and stopwords removal. However, it was identified the presence of methods explicitly targeting sentiment information, such as to interpret adverbs of intensity as sentiment boosters [20]. Moreover, it is important to highlight that the aim was not to detect tasks exclusively to Portuguese, but those that are applied in the Brazilian context. Then, it is believed that the lower implementation complexity, pushed by their broad adoption, justifies the prevalence of removal tasks and tokenization.

Regarding the frequency of adoption, the steps dealing with stopwords and hashtags were the most applied, mentioned in 46% and 36% of the studies, respectively. On the other hand, the 6 least implemented tasks found, and not available in the Figure 1, were the handling of uppercase, repetition of letters, exclamation, question marks, laughing, and greeting patterns in the text, all with one occurrence each. Besides, the greater complexity of implementation can still affect the application of rare tasks as phonetic misspelling, which depends on the linguistic structure of a language to be applied.

B. What are the main methodologies to implement preprocessing steps for SA in Brazilian Portuguese Social Media?

This work has identified a categorization schema to generalize methodologies to implement SA preprocessing. The categories detected are: **Transformation (TR)** transforms the textual content through a particular task, such as stemming; **Deletion (RM)** removal of specific elements, such as emoticons; **Expansion and Replacement (ER)** a term is expanded and replaced, such as abbreviations; **Correction and Replacement (CR)** a misspelled term is corrected and replaced; **Content Extraction (CE)** retrieves content from special terms, such as hashtags, and replaces the original term by its extracted content; **Identifier Replacement (RE)** replaces a term by a unique identifier; **Polarity Computation (PC)** a term is considered to compute the final polarity score, according to its sentiment or associated characteristic.

Table II provides examples of outputs from each methodology for some preprocessing tasks. The text input given is: ”Gr8! I dreamed I was wathing the avengers!! :) #Marvel”.

TABLE II
IMPLEMENTATION METHODOLOGIES FOR PREPROCESSING IN SA

M	Task	Output
TR	Stemming	Gr8! I dream I was wathing the aveng!! #Marvel :)
RM	Punctuation	Gr8 I dreamed I was wathing the avengers #Marvel :)
ER	Slang	Great! I dreamed I was wathing the avengers!! #Marvel :)
CR	Misspelling	Gr8! I dreamed I was watching the avengers!! #Marvel :)
CE	Hashtag	Gr8! I dreamed I was wathing the avengers!! Marvel :)
RE	Emoticons	Gr8! I dreamed I was wathing the avengers!! #Marvel positive
PC	Polarity Sum	+1 (”:)”) and +1 (”Gr8”)

In Table II, the Stemming task works transforming the words ”dreamed” and ”avengers” to their radical forms. The RM methodology removes the exclamation signs. The ER method expands the slang ”Gr8” to ”Great”. The CR methodology corrects the misspelling ”wathing” to ”watching”. The CE approach extracts the hashtag content ”Marvel”. The RE method replaces the emoticon by its polarity, which is ”positive”. Finally, the PC methodology takes positive polarity values from the emoticon and slang present in the text.

Finally, Table III summarizes all the preprocessing steps detected, descriptions, and methodologies adopted to their implementation in a SA pipeline. It is noticeable that most of the Brazilian authors found still do not explore all the possible approaches for some tasks, such as the investigation of the RE methodology for laughing, greetings patterns, exclamation marks, sentiment words and emoticons. Therefore, no previous

[1] (”sentiment analysis” OR ”opinion mining” OR ”opinion detection” OR ”sentiment mining”) AND (”portuguese” OR ”brazilian”)

[2] (”análise de sentimento” OR ”mineração de opinião” OR ”detecção de opinião” OR ”mineração de sentimento”) AND (”português” OR ”brasileiro”)

[3] https://gofile.io/?c=1szBYQ

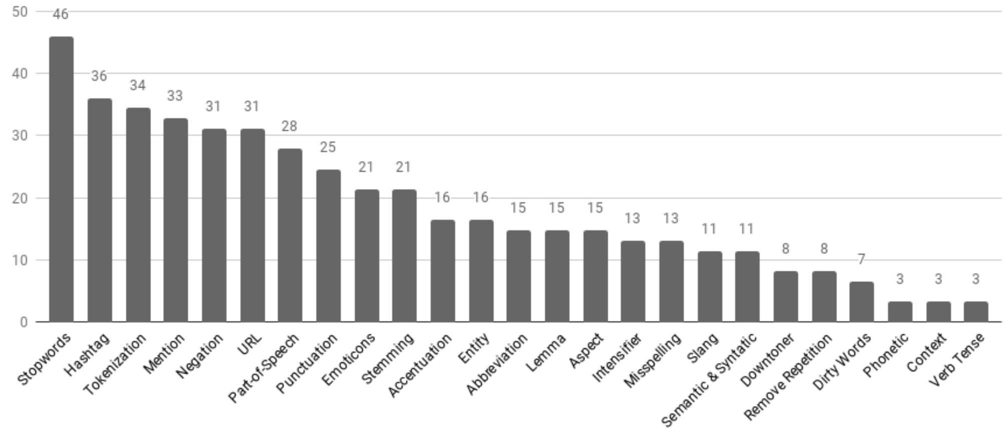

Fig. 1. Preprocessing steps found by usage frequency.

proposal for SA in Brazilian Portuguese has implemented all the preprocessing tasks detected in the Brazilian literature.

V. Conclusions

This paper has conducted a systematic review regarding the topic of preprocessing for SA for Brazilian Social Media data. Research questions were established to formalize the review scope. The results obtained have provided satisfactory feedback to the questions raised. Moreover, among the findings, it was noticed that the majority of studies are carried out implementing less complex and noise removal steps, such as the elimination of stop words and tokenization. Multiple methodologies to implement the same preprocessing tasks were also noticed. However, few authors explore this diversity of approaches. In general, each study presents specific preprocessing steps, with a focus on its application. This fact has reinforced the conclusion that there is not a complete and uniform framework for preprocessing Social Media data, with an emphasis on SA in the Brazilian Portuguese language.

As future work, it is intended to expand the research questions adopted in a new review, including aspects such as algorithms most employed. Besides, it is an aim to compare the results obtained with other literature reviews, not restricted to Brazilian Portuguese.

Acknowledgment

This research was supported by the the National Council for the Improvement of Higher Education (CAPES); the National Council for Scientific and Technological Development CNPq); and Amazon Foundation for Studies and Research (Fapespa).

References

[1] R. Agnihotri, R. Dingus, M. Y. Hu, and M. T. Krush, “Social media: Influencing customer satisfaction in b2b sales,” *Industrial Marketing Management*, vol. 53, pp. 172 – 180, 2016.

[2] F. Lobato, M. Pinheiro, A. Jacob, O. Reinhold, and Á. Santana, “Social CRM: Biggest Challenges to Make it Work in the Real World,” in *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers*, W. Abramowicz, R. Alt, and B. Franczyk, Eds. Cham: Springer International Publishing, 2017, vol. 263, pp. 221–232.

[3] G. R. T. de Almeida, F. Lobato, and D. Cirqueira, “Improving Social CRM through eletronic word-of-mouth: a case study of ReclameAqui,” in *XIVWorkshop de Trabalhos de Iniciação Cientfíc*, 2017.

[4] W. Silva, Á. Santana, F. Lobato, and M. Pinheiro, “A Methodology for Community Detection in Twitter,” in *Proceedings of the International Conference on Web Intelligence*, 2017, pp. 1006–1009. [Online]. Available: http://doi.acm.org/10.1145/3106426.3117760

[5] D. Cirqueira, M. Pinheiro, T. Braga, A. Jacob Jr, O. Reinhold, R. Alt, and Á. Santana, “Improving relationship management in universities with sentiment analysis and topic modeling of social media channels: learnings from ufpa,” in *Proceedings of the International Conference on Web Intelligence*. ACM, 2017, pp. 998–1005.

[6] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[7] The Wall Street Journal, “Brazil: The Social Media Capital of the Universe,” 2013.

[8] We are Social, “Digital in 2017: global overview,” 2018.

[9] B. A. Souza, T. G. Almeida, A. A. Menezes, F. G. Nakamura, C. M. Figueiredo, and E. F. Nakamura, “For or against?: Polarity analysis in tweets about impeachment process of brazil president,” in *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*, ser. Webmedia ’16. New York, NY, USA: ACM, 2016, pp. 335–338.

[10] D. Cirqueira, A. Jacob, F. Lobato, A. L. de Santana, and M. Pinheiro, “Performance Evaluation of Sentiment Analysis Methods for Brazilian Portuguese,” in *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers*, W. Abramowicz, R. Alt, and B. Franczyk, Eds. Cham: Springer International Publishing, 2017, pp. 245–251.

[11] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[12] S.-A. Bahrainian and A. Dengel, “Sentiment analysis using sentiment features,” in *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03*. IEEE Computer Society, 2013, pp. 26–29.

[13] G. S. Solakidis, K. N. Vavliakis, and P. A. Mitkas, “Multilingual sentiment analysis using emoticons and keywords,” in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, vol. 2. IEEE, 2014, pp. 102–109.

[14] K. Schouten, F. Baas, O. Bus, A. Osinga, N. van de Ven, S. van Loenhout, L. Vrolijk, and F. Frasincar, “Aspect-based sentiment analysis using lexico-semantic patterns,” in *International Conference on Web Information Systems Engineering*. Springer, 2016, pp. 35–42.

[15] E. Souza, D. Castro, D. Vitório, I. Teles, A. L. Oliveira, and C. Gusmão, “Characterizing user-generated text content mining: A systematic map-

TABLE III
PREPROCESSING TASKS IDENTIFIED IN THE LITERATURE REVIEW, AND ADOPTED METHODOLOGIES TO THEIR IMPLEMENTATION.

Task	**Description**	**TR**	**RM**	**ER**	**CR**	**CE**	**RE**	**PC**
Tokenization	Split a text in tokens	●						
Stopwords	Remove stopwords		●					
Accentuation	Remove accentuation		●					
Punctuation	Treat or remove punctuation, such as stop and comma marks	●	●					
Intensifiers	Treat specific terms as signs of sentiment boosting, such as adverbs							●
Downtoner	Treat specific terms as sentiment detractors							●
Uppercase	Treat uppercase characters as signs of sentiment boosting							●
Remove Repetitions	Remove repeated characters from words		●					
Treat Repetitions	Treat repeated characters as signs of sentiment boosting							●
Exclamation	Remove or treat as signs of sentiment boosting		●					●
Question	Remove question mark		●					
Abbreviations	Treat abbreviations in a text by their expansion			●				
Slangs	Treat slangs in a text by their expansion with or no polarity information			●				
Dirty Words	Remove obscene words from a text		●					
Emoticons	Treat or remove representations of emotional expressions		●					●
Misspelling	Correct misspellings				●			
Phonetic Misspelling	Correct misspellings concerning phonetic structure in a language				●			
Negation	Invert polarity in a text based on negation particles, such as "not"							●
Stemming	Reduce words by removing affixes for normalization	●						
Lemmatization	Reduce words based on their morphological analysis for normalization	●						
Part-of-Speech	Extract syntactic function of words	●						
Entity	Identify sentiment targets	●						
Aspect	Detect target aspects of sentiment within an entity	●						
Context	Consider semantic context to analyze a text	●						
Semantics & Syntactic	Evaluate semantic dependency and relationships among terms							
Verb Tense	Consider verbal tense in a text to influence in sentiment scoring							●
Hashtag	Remove hashtags, or replace their content or the character "#"		●			●	●	●
Mention	Handle references to users within an OSN		●			●		
URL	Remove hyperlinks, or treat them as a special character		●				●	
Laughing Pattern	Treat laughing patterns in a text		●					
Greetings	Identify greetings in a text		●					

ping study of the portuguese language," in *New Advances in Information Systems and Technologies*. Springer, 2016, pp. 1015–1024.

[16] E. Souza, D. Vitório, D. Castro, A. L. Oliveira, and C. Gusmão, "Characterizing opinion mining: A systematic mapping study of the portuguese language," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2016, pp. 122–127.

[17] T. A. S. Pardo, C. V. Gasperin, H. M. Caseli, and M. d. G. V. Nunes, "Computational linguistics in brazil: An overview," in *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, ser. YIWCALA '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1–7. [Online]. Available: http://dl.acm.org/citation.cfm?id=1868701.1868702

[18] M. da Silva Conrado, A. D. Felippo, T. A. Salgueiro Pardo, and S. O. Rezende, "A survey of automatic term extraction for brazilian portuguese," *Journal of the Brazilian Computer Society*, vol. 20, no. 1, p. 12, May 2014.

[19] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Tech. Rep. EBSE 2007-001, 2007.

[20] L. V. Avanço and M. d. G. V. Nunes, "Lexicon-based sentiment analysis for reviews of products in brazilian portuguese," in *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*. IEEE, 2014, pp. 277–281.