SentiProdBR: Building Domain-Specific Sentiment Lexicons for the Portuguese Language

Tiago de Melo

Universidade do Estado do Amazonas (UEA) - Manaus, AM - Brazil
tmelo@uea.edu.br

Abstract. *Online reviews are readily available on the Web and widely used for decision-making. However, only a few studies on Portuguese sentiment analysis are reported due to the lack of resources including domain-specific sentiment lexical collections. In this paper, we present an effective methodology using probabilities of the Bayes' Theorem for building a set of lexicons, called SentiProdBR, for 10 different product categories for the Portuguese language. Experimental results indicate that our methodology significantly outperforms several alternative approaches of building domain-specific sentiment lexicons.*

1. Introduction

Sharing of user experiences on web-based opinions platforms, such as Amazon.com, is becoming widespread. This results in a huge number of available reviews which can be a valuable source of knowledge for decision-making. As a result, manufacturers can obtain rapid feedback to improve the quality of their products, and customers can make purchasing decisions based on others' opinions. Despite the benefits of such reviews, extracting useful information represents a significant challenge due to the large scale and distinct characteristics. Sentiment analysis can provide a feasible and valuable way to automatically scan through reviews and classify them into different sentiment polarities with strength indications [Xiang et al. 2019].

Numerous studies have focused on sentiment analysis for the English language. However, there is a need for further research in other languages such as Portuguese. The lack of word processing tools and annotated data for experiments appear as a challenge for sentiment analysis in this language. Another issue is the lack of suitable Natural Language Processing (NLP) resources for Portuguese such as specific lexicons for general use and lexicons for sentiment analysis. In addition, as reported by Pereira [Pereira 2021], there is still room for sentiment analysis method development for the Portuguese language that explore linguistic specificities. For example, to the best of our knowledge, there is no method or data collection with domain-specific sentiment lexicons for Portuguese.

In this paper, we propose a methodology to automatically build a domain-specific sentiment lexicon, called SentiProdBR, in an unsupervised way and without prior knowledge. To this end, we collected 342,815 product reviews from 10 different categories published on Amazon.com, and the domain-specific sentiment scores were calculated using probabilities of the Bayes' Theorem of each word as introduced in [Labille et al. 2017, Huang et al. 2020]. We compared the performance of our methodology with three popular sentiment lexicons in Portuguese. Results showed that the proposed method significantly outperformed the baselines. Furthermore, results obtained indicated that the proposed method can be used in real applications, achieving a F1-score of 0.838 on average.

Our main contributions can be summarized as follows. First, we propose a methodology to automatically build domain-specific sentiment lexicons via probability theory for the Portuguese language. Secondly, we empirically demonstrate that creating an accurate method for unsupervised building sentiment lexicon tasks is possible. Finally, we are making the sentiment lexicons with 32,019 terms in product-specific domains available[1] to the research community.

The remainder of the paper is organized as follows. Section 2 provides a review of related work on sentiment lexicons approaches. Section 3 presents the methodology used in our research. Section 4 includes an experimental evaluation and discussion of the proposed method. Finally, Section 5 concludes this work.

2. Related Work

Several studies discuss building sentiment lexicons. This section focuses on some of the relevant techniques conducted on the creation of sentiment lexicons in general purpose and domain-specific paradigms for the Portuguese language. ReLi is a domain-dependent lexicon composed of a set of 1,600 reviews from 13 Portuguese books published on the Internet. The lexicon was manually annotated with opinion information by Freitas et al. [Freitas 2013], and it contains 609 entries (385 positives and 224 negatives). OpLexicon is a sentiment lexicon with 32,191 entries (24,475 adjectives and 6,889 verbs), based on journalistic texts and film reviews written in Brazilian Portuguese [Souza and Vieira 2011]. They generate a list composed of the adjective's name and polarity, which assign ones of two values: 1 and -1. Vilares et al. [Vilares et al. 2018] proposed a method to automatically generate SenticNet for various languages, including Portuguese, and obtained BabelSentic. They use statistical machine translation tools to create sentiment lexicons for each target language. Our study differs from these approaches by creating the unsupervised domain-specific lexicons.

3. Materials and Methods

3.1. Data of Domain-Specific

To build SentiProdBR, we collected 342,815 user reviews from Amazon[2] for 10 different categories submitted from 2012 through 2021. Reviews are rated from 1 to 5 stars. We consider reviews rated 1-star and 2-star to be negative, whereas 4-star and 5-star reviews are considered positives. 3-star reviews are considered neutral and are ignored.

Previous sentiment analysis studies [Almatarneh and Gamallo 2018] are focused on adjectives as the primary subjective content source in a text. Following this, we carry out pos-tagging using spaCy[3] to identify adjectives, which we filter results by. Table 1 presents a summary of statistics from the dataset.

3.2. Building Lexicons Algorithm

Algorithm 1 outlines our building lexicon algorithm. The algorithm accepts as input a set of reviews $\mathcal{R}$ and yields as output a set $\mathcal{L}$ of lexicon pairs, where each pair is comprised of a word $w_i \in \mathcal{R}$ and a polarity $p \in \{positive, negative\}$.

[1] http://tiagodemelo.info/datasets.html

[2] https://www.amazon.com.br

[3] https://spacy.io

Table 1. Statistics of user reviews for each domain.

Domain	#Products	#Reviews		Average Reviews per Products	#Words		Average Words	
		#POS	#NEG		%POS	%NEG	#POS	#NEG
Automotive	829	11,634	1,727	16.12	81.27%	18.73%	12.6	19.6
Baby	635	17,351	1,820	30.19	85.25%	14.75%	11.7	19.3
Books	747	120,042	5,414	167.95	93.89%	6.11%	20.1	29.1
Cellphones	271	36,566	1,804	141.59	91.97%	8.03%	15.8	28.0
Fashion	1,267	15,607	4,258	15.68	69.62%	30.38%	11.1	17.8
Food	994	13,988	1,604	15.69	83.30%	16.70%	11.0	19.2
Games	699	46,692	4,370	73.05	86.31%	13.69%	14.9	25.2
Laptops	71	3,298	690	56.17	76.07%	23.93%	20.2	30.4
Pets	701	6,383	735	10.15	86.31%	13.69%	15.1	20.8
Toys	1,196	29,312	3,018	27.03	85.82%	14.18%	13.4	21.5

Algorithm 1: Building Lexicon Algorithm

Input: Set of reviews $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$;
Output: Lexicons pairs $\mathcal{L} = \{\langle w_1, p\rangle, \langle w_2, p\rangle, \ldots, \langle w_m, p\rangle\}$, where word $w_i \in \mathcal{R}$ and $p \in \{positive, negative\}$;

1 **let** $\mathcal{R}^+$ be the set of positive reviews $R^+ \subseteq R$;
2 **let** $\mathcal{R}^-$ be the set of negative reviews $R^- \subseteq R$;
3 **let** $\mathcal{W}$ be the set of words $w \in \mathcal{R}$;
4 **let** $p(+|w)$ be the probability of word w being positive;
5 **let** $p(-|w)$ be the probability of word w being negative;
6 **let** τ be the threshold;
7 **foreach** $w_i \in \mathcal{W}$ **do**
8 **if** w_i *is not adjective* **then**
9 continue;
10 $p(+|w_i) = \frac{p(+)\times p(w_i|+)}{p(w_i)}$;
11 $p(-|w_i) = \frac{p(-)\times p(w_i|-)}{p(w_i)}$;
12 $score(w_i) = p(+|w_i) - p(-|w_i)$;
13 **if** $score(w_i) \geq \tau$ **then**
14 $\mathcal{L} \leftarrow \mathcal{L} \cup \{\langle w_i, positive\rangle\}$;
15 **else**
16 $\mathcal{L} \leftarrow \mathcal{L} \cup \{\langle w_i, negative\rangle\}$;
17 **return** $\mathcal{L}$

The algorithm iterates through the words $w_i \in \mathcal{W}$ (Loop 7- 16), where words that are not adjectives are discarded (Lines 8 and 9). In Line 10, the algorithm calculates the probability $p(+|w_i)$ of word w_i of being positive, where $p(+)$ is the proportion of words belonging to the positive class (+), i.e., the quotient of the number of words in the positive reviews $\mathcal{R}^+$ and the total number of words appearing in all reviews $\mathcal{R}$. In Line 11, the algorithm calculates the probability $p(-|w_i)$ of the same word w_i being negative, where $p(-)$ is the proportion of words belonging to the negative class (-), i.e., the quotient of the number of words in the negative reviews $\mathcal{R}^-$ and the total number of words appearing in all reviews $\mathcal{R}$.

In Line 12, $score(w_i)$ produces scores in the range from 1 to -1, where 1 indicates that w_i is absolutely positive and -1 indicates that w_i is absolutely negative. The formulas in Lines 10 and 11 do not consider that there are much more positive reviews than negative ones. Our assumption is that the polarity of words tends to be positive due to the greater

number of positive reviews compared to the number of negative reviews. Therefore, we added a weight factor τ to consider the frequency of words within 5 and 4 star classes. If $score(w_i) \geq \tau$, then w_i is considered positive. Otherwise, w_i is considered negative. The weight factor τ was chosen empirically, as discussed in the next section.

3.3. Baselines

To evaluate SentiProdBR, we used three popular lexicons for the Portuguese language: a) OpLexicon; b) BabelSentic; c) ReLi, as mentioned in Section 2. To the best of our knowledge, there are no available sentiment lexicons for Portuguese in product-specific domains evaluated in our experiments.

4. Experiments

4.1. Experimental Setup

We evaluate SentiProdBR using sentiment analysis for each domain dataset, compared against baselines. We compute the review score by summing up each term's score in the review from its domain-specific lexicon, then normalizing for length. If the resulting score is positive, then the review is deemed positive, and vice versa.

Figure 1 shows an example of user review along with the score of sentiment lexicons found for the terms *horrível* (horrible) and *decepcionado* (disappointed). The score is calculated as the average of the sentiment lexicons. For example, the review in Figure 1 would have a score of -0.729 ($\frac{-0.822-0.636}{2}$) and would be classified as negative.

$$\textit{Notebook } \overbrace{\textit{horrível}}^{\text{-0.822}}\textit{. Estou } \overbrace{\textit{decepcionado}}^{\text{-0.636}}\textit{.}$$

(Horrible notebook. I'm disappointed.)

Figure 1. Example of computing score.

To measure the performance of each approach, we used three commonly adopted measures in previous works [Labille et al. 2017, Labille et al. 2016, Deng et al. 2017]; namely, *precision*, *recall*, and *F1-score*. Precision is the ratio of correctly predicted polarity of user reviews to the total predicted polarity of user reviews. Recall is the ratio of correctly predicted polarity of user reviews to the total of user reviews in each dataset. Finally, F1-score is the harmonic mean of precision and recall. The metrics are defined as $P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F1 = \frac{2\times P\times R}{P+R}$, where TP means the number of user reviews was identified correctly; FP means the number of user reviews was identified incorrectly; and FN means the number of user reviews without any lexicon.

4.2. Experimental Results

We evaluated SentiProdBR lexicons versus three baselines and reported our results in Table 2, which shows the precision, recall, and F1-Score averaged across all 10 domains. As shown, lexicons of SentiProdBR are more accurate than both generic lexicons. Our domain-specific lexicons achieve an F1-score of 0.838 on average, which is an improvement of 28.33% over OpLexicon and an improvement of 28.92% over BabelSentic. This validates our assumption that some words are associated with different sentiments and sentiment strengths depending on the domain. ReLi is a domain-specific lexicon; it

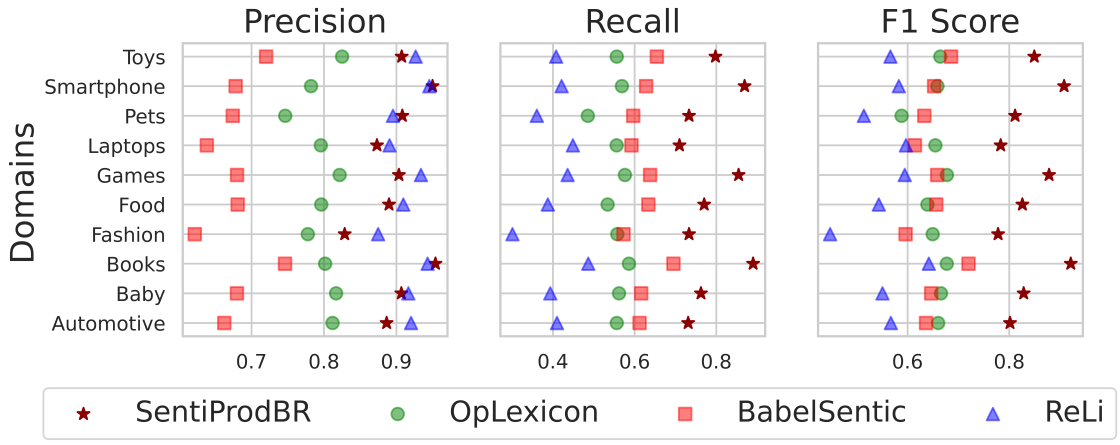
Table 2. Evaluation across all domains (average).

	Precision	Recall	F1 Score
SentiProdBR	0.901	**0.785**	**0.838**
OpLexicon	0.797	0.554	0.653
BabelSentic	0.678	0.624	0.650
ReLi	**0.916**	0.405	0.560

achieves the best precision on average. However, ReLi's lexicons set is small and hence, presented a low F1-score. We adopt $\tau = 0$ for these experiments.

We further evaluated the performances of each lexicon against each domain and reported the results in Figure 2. ReLi and SentiProdBR achieved close results in terms of precision. The good performance of ReLi, in terms of precision, is due to its small set of lexicons. However, the recall achieved by ReLi is quite low. SentiProdBR achieved better results than others in all domains for recall and F1-score metrics. Our best domain-specific lexicons reached 0.92 for F1-score in the domain *Books* against BabelSentic's score of 0.67. Conversely, our lowest domain-specific F_1 score was achieved in the category of *Fashion* products with 0.78 versus the second best method, OpLexicon, that achieved 0.65. We believe this is due to the fact that the *fashion* category is comprised of several subcategories, such as shoes, clothes and jewelry, and whose lexicon is much different from each other.

Figure 2. Metrics of all approaches on all domains.



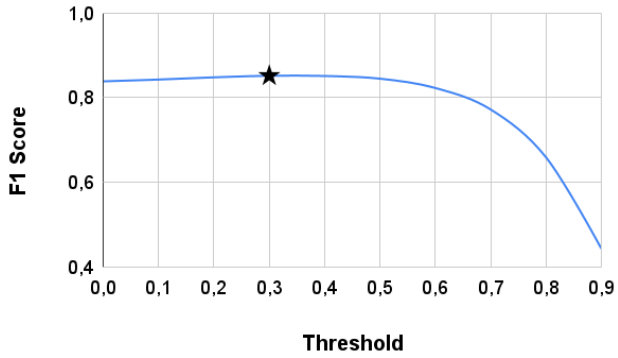
4.3. Estimating Factor τ

In Section 3.2, we proposed using weight factor τ to consider the user review frequency within each star class. Our assumption is we should be stricter with the most frequent classes. To obtain the best threshold, we perform experiments with different factor τ values. The results are presented in Figure 3, where we plot F_1 score averaged across all 10 domains when varying factor τ from 0 to 0.9. As shown, $\tau = 0.3$ produces the best average across all domains, with approximate F1-score gains of 0.02, when compared to default $\tau = 0$.

5. Conclusions

In this study, we proposed a methodology to build domain-specific sentiment lexicons for Portuguese. Our work differs from the traditional approaches by creating the unsupervised

Figure 3. Influence of the factor τ across all 10 domains.



domain-specific lexicons. To achieve this goal, we employed probabilities to calculate the sentiment strength of each word. We evaluated our method with three baselines, showing the efficacy of our methodology. Another advantage is that we do not have to adapt our lexicon from generic lexicons. In addition, we have applied our method for an extensive dataset and generated SentiProdBR as a large domain-specific sentiment lexicon for 10 different categories. In future work, we plan to experiment with using deep learning and word embeddings for sentiment lexicon creation.

References

Almatarneh, S. and Gamallo, P. (2018). A lexicon based method to search for extreme opinions. *PLOS ONE*, 13(5):1–19.

Deng, S., Sinha, A. P., and Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, 94:65–76.

Freitas, C. (2013). Sobre a construção de um léxico da afetividade para o processamento computacional do português. *Revista Brasileira de Linguística*, 13(4):1031–1059.

Huang, M., Xie, H., Rao, Y., Feng, J., and Wang, F. L. (2020). Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Information Sciences*, 520:389–399.

Labille, K., Alfarhood, S., and Gauch, S. (2016). Estimating sentiment via probability and information theory. *KDIR*, 2016:121–129.

Labille, K., Gauch, S., and Alfarhood, S. (2017). Creating domain-specific sentiment lexicons via text mining. In *Workshop Issues Sentiment Discovery Opinion Mining*, pages 1–8.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Souza, M. and Vieira, R. (2011). Construction of a portuguese opinion lexicon from multiple resources. *Simpósio Brasileiro de TI e da Linguagem Humana*.

Vilares, D., Peng, H., Satapathy, R., and Cambria, E. (2018). Babelsenticnet: a commonsense reasoning framework for multilingual sentiment analysis. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1292–1298. IEEE.

Xiang, R., Jiao, Y., and Lu, Q. (2019). Sentiment augmented attention network for cantonese restaurant review analysis. In *Proceedings of WISDOM'19: Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'19)*.