SentiLexBR: An Automatic Methodology of Building Sentiment Lexicons for the Portuguese Language

Tiago de Melo

Universidade do Estado do Amazonas, Brazil
tmelo@uea.edu.br

Abstract. User reviews are readily available on the Web and widely used for sentiment analysis tasks. Sentiment lexicons plays an important role in sentiment analysis, where each sentiment word is given a sentiment label (positive or negative) or score (1 or -1). However, a sentiment lexicon may express different sentiment polarity according different domain. In addition, only a few studies on Portuguese sentiment analysis are reported due to the lack of resources including domain-specific sentiment lexical corpora. In this paper, we present an effective methodology, called SentiLexBR, using probabilities of the Bayes' Theorem for building a set of sentiment lexicons. An unsupervised algorithm is proposed to automatically identify sentiment lexicons with their polarities for the Portuguese language. Experimental results on user reviews datasets in 12 different domains indicate the effectiveness of our methodology in domain-specific sentiment lexicon generation for Portuguese. In addition, the sentiment lexicon produced by SentiLexBR also significantly outperforms several alternative approaches of building domain-specific sentiment lexicons.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text analysis; I.7.2 [**Document and Text Edition**]: Languges

Keywords: Natural Language Processing, Portuguese Language, Sentiment Analysis, Sentiment Lexicon

1. INTRODUCTION

Online user reviews posted on web-based opinions platforms, such as Amazon.com and Tripadvisor.com, are 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, companies can use the feedback provided by reviews of their business to measure the level of satisfaction related to specific services or products [Amora et al. 2018], and customers can make purchasing decisions based on others' opinions [de Melo et al. 2019]. 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 [Birjali et al. 2021; Chaturvedi et al. 2018]. However, the linguistic resources available for sentiment analysis in other languages, such as Portuguese, are still limited [Oliveira and Melo 2021]. 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.

This work is an extended version of SentiProdBR [de Melo 2021] published in the 36th Brazilian Symposium on Databases (SBBD'21) with: (*i*) related work section expanded; (*ii*) news experiments considering verbs and not only adjectives; (*iii*) increasing number of datasets and baselines in experiments. In this paper, we propose a methodology to automatically build a domain-specific sentiment lexicon, called SentiLexBR, in an unsupervised way and without prior knowledge. To this end, we collected 917,895 user reviews from 12 different categories published on Amazon.com.br and TripAdvisor.com.br, 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 four popular sentiment lexicons in Portuguese and also with SentiProdBR. 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 0.889 on average across all 12 domains.

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 111,724 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

Sentiment lexicons are expressions where each word is associated with an opinion polarity (negative or positive). Sentiment lexicon plays a crucial role in sentiment analysis, therefore, how to build a sentiment lexicon for different languages has attracted lots of attention, including Chinese [Zhang et al. 2018], Italian [Catelli et al. 2022], Korean [Park et al. 2018], and Swedish [Nusko et al. 2016]. The methodologies to create sentiment lexicons are classified into four main categories: manual, rule-based, dictionary-based and corpus-based. 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. The methods mentioned below are considered state-of-the-art for creating sentiment lexicons for Portuguese, and therefore, they were used as baselines against SentiLexBR.

Manual creation of a sentiment lexicon consists annotating a list of lexical units with their sentiment polarity. The major advantage of this methodology is the correctness in the annotation of terms because this is done by humans [Ahire 2014]. However, the main drawbacks of this methodology are time consuming and costly to perform. Another disvantage is need of domain experts to manually assign sentiment values to words. For these reasons, the size of the manual sentiment lexicon is typically low. 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).

Rule-based methodology identifies new sentiment lexicons by rules hand-written or methods that attempt to determine syntactic patterns in the language. Valence Aware Dictionary and sEntiment Reasoner (VADER) [Hutto and Gilbert 2014] is a well-known rule-based lexicon and sentiment analysis tool that is specifically adapted to sentiments expressed in social media. VADER has a set of sentiment lexicons whose semantic orientation is classified as positive or negative. The list of lexicons is only in English and the tool is based on using machine-learning web service to automatically translate the

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

text into English. As such, the results are reliant not only upon the accuracy of the sentiment tool but also upon the accuracy of which ever translation tool you use to create the English version of the input. This explains why we do not consider VADER as a baseline.

SentiStrength[2] is a sentiment lexicon that uses linguistic information and rules to detect sentiment strength in English text [Thelwall 2014]. The tool is quite extensible and flexible and this has motivated many researchers to adapt it to other languages including Portuguese. SentiStrength provides negative and positive sentiment scores for each term. Both scores are from 1 to 5, where 1 represents a weak sentiment and 5 represents a strong sentiment. The overall polarity is calculated by subtracting the negative sentiment score from the positive sentiment score.

Dictionary-based methodology identifies new sentiment lexicons by their relations, such as antonyms or synonyms, with a small set of seed lexicons. This set of seed consists of a small group of words for which the sentiment orientation is known in advance. The small set of seed words is expanded by looking up the seed words antonyms and synonyms in a dictionary. According to Deng et al. [Deng et al. 2017], the effectiveness of this methodology is highly dependent of the dictionary used. Vieira and Souza [Souza and Vieira 2011] proposed a method to create a lexicon called *OpLexicon*. 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. They generate a list composed of the adjective and verb's name and polarity, which assign ones of two values: 1 and -1. We used the list of adjectives and verbs in Portuguese. Although the relationships between entries are highly accurate, sentiment lexicons generated with dictionary-based methodologies do not contain any domain-specific information.

Corpus-based methodology identifies sentiment lexicons based on their relationships with each other from a corpus. This methodology could also use a list of seed words, but the list is expanded using corpora instead of a dictionary [Bos and Frasincar 2021]. 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.

Table I presents a comparative summary of aforementioned methods of building sentiment lexicons and SentiLexBR along with relevant characteristics.

Table I. A comparative summary of building sentiment lexicons.

	Methology	**Purpose**	**Original Language**
SentiLexBR	Rule	Domain-specific	Portuguese
SentiProdBR	Rule	Domain-specific	Portuguese
OpLexicon	Dictionary	Domain-specific	Portuguese
BabelSentic	Corpus	General	English
SentiStrength	Rule	General	English
ReLi	Manual	Domain-specific	Portuguese

3. MATERIALS AND METHODS

3.1 Data of Domain-Specific

To build SentiLexBR, we collected 342,815 user reviews from Amazon[3] for 10 different product categories submitted from 2012 through 2021 and 575,080 user reviews submitted from 2008 through 2021

[2]http://sentistrength.wlv.ac.uk
[3]https://www.amazon.com.br

in Tripadvisor[4] website. In the latter, we collected reviews regarding restaurants and tourist attractions named as Point of Interest (POI). 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, SentiProdBR uses only adjectives in the construction of its set of sentiment lexicons. Unlike SentiProdBR, we hypothesise that verbs can be used to denote sentiments. Following this, we carry out pos-tagging using spaCy[5] to identify adjectives and verbs, which we filter results by. For example, Fig. 1 shows a sentence along with POS tagging by using spaCy. In this example, only the verb "*loved*" would be extracted to compose the set of sentiment lexicons. Although the sentence has no adjectives, it is clearly positive and indicates the importance of considering verbs as sentiment lexicons.

I	loved	the	robustness	of	laptop
PROPN	**VERB**	DET	NOUN	DET	NOUN

Fig. 1. Example of sentence with POS tagging.

Table II presents a summary of statistics from the dataset. From Table II, it is possible to notice that the rating distribution is quite skewed, where the majority of ratings are 4-star and 5-star with 95.3% and the smallest proportion is 1-start and 2-star with only 4.7%. However, in terms of review text length, it is quite different. Users tend to write longer texts to justify a lower rating in all domains.

Table II. Statistics of user reviews for each domain.

Domain	**#Entities**	**#Reviews**		**Average Reviews per Entities**	**#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
POI	126	542,245	15,785	4,811	95.17%	4.83%	40.7	71.1
Restaurants	383	16,075	975	49.83	89.44%	10.56%	35.9	70.1
Toys	1,196	29,312	3,018	27.03	85.82%	14.18%	13.4	21.5

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 $p \in \{positive, negative\}$.

The algorithm iterates through the words $w_i \in \mathcal{W}$ (Loop 7- 16), where words that are not adjectives or verbs are discarded (Lines 8 and 9). In Line 10, the algorithm calculates the probability $p(+|w_i)$

[4]https://www.tripadvisor.com.br
[5]https://spacy.io

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 $w_i \in \mathcal{R}$;
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 or verb)* **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** $\mathcal{L}$

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 $\mathcal{R}$; $p(w_i|+)$ is the probability to observe the word w_i given the positive reviews $\mathcal{R}^+$; and $p(w_i)$ is the total number of occurrences of w_i 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 $\mathcal{R}$; $p(w_i|-)$ is the probability to observe the word w_i given the negative reviews $\mathcal{R}^-$; and $p(w_i)$ is the total number of occurrences of w_i in all reviews $\mathcal{R}$. Notice that $p(w_i)$ must be nonzero since the word $w_i \in \mathcal{R}$.

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 SentiLexBR, we compared with the previous version called SentiProdBR and also used four popular lexicons for the Portuguese language: a) OpLexicon; b) BabelSentic; c) ReLi; d) SentiStrength, 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 SentiLexBR 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.

Fig. 2 shows an example of user review regarding laptop 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 Fig. 2 would have a score of -0.729 ($\frac{-0.822-0.636}{2}$) and would be classified as negative.

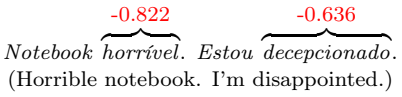

Fig. 2. Example of computing score.

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:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \times P \times R}{P + R}, \tag{1}$$

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

The goal of this first set of experiments is to evaluate the quality of SentiLexBR lexicons versus SentiProdBR and four baselines. Our results are reported in Table III, which shows the precision, recall, and F1-score averaged across all 12 domains. As shown, lexicons of SentiProdBR are more accurate than both generic lexicons. Our domain-specific lexicons achieve an F1-score of 0.889 on average, which is an improvement of 36.14% over OpLexicon and an improvement of 36.76% over BabelSentic and SentStrength. This validates our assumption that some words are associated with different sentiments and sentiment strengths depending on the domain. In addition, on average, SentiLexBR yielded gains of F1-score of approximately 6.08% when compared to SentiProdBR. This validates our assumption that verbs can also be used as sentiment lexicons. ReLi is a domain-specific lexicon; it 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.

Table III. Evaluation across all domains (average).

	Precision	**Recall**	**F1 Score**
SentiLexBR	0.909	**0.871**	**0.889**
SentiProdBR	0.901	0.785	0.838
OpLexicon	0.797	0.554	0.653
BabelSentic	0.678	0.624	0.650
SentiStrength	0.800	0.548	0.650
ReLi	**0.916**	0.405	0.560

We further evaluated the performances of each lexicon against each domain and reported the results in Fig. 3. ReLi, SentiProdBR, and SentiLexBR 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. SentiLexBR achieved better results than others in all domains for recall and F1-score metrics. To consider verbs as sentiment lexicons led to an improvement in the recall of SentiLexBR when compared to SentiProdBR which considered only adjectives. We observed that there are several sentences that do not have adjectives, but that are really opinionated. For example, the sentence "*Eu amei a pizza*" (I loved the pizza) has no adjectives and therefore would not be identified as subjective by SentiProdBR. However, the verb "*amei*" (loved) clearly indicates a positive sentiment.

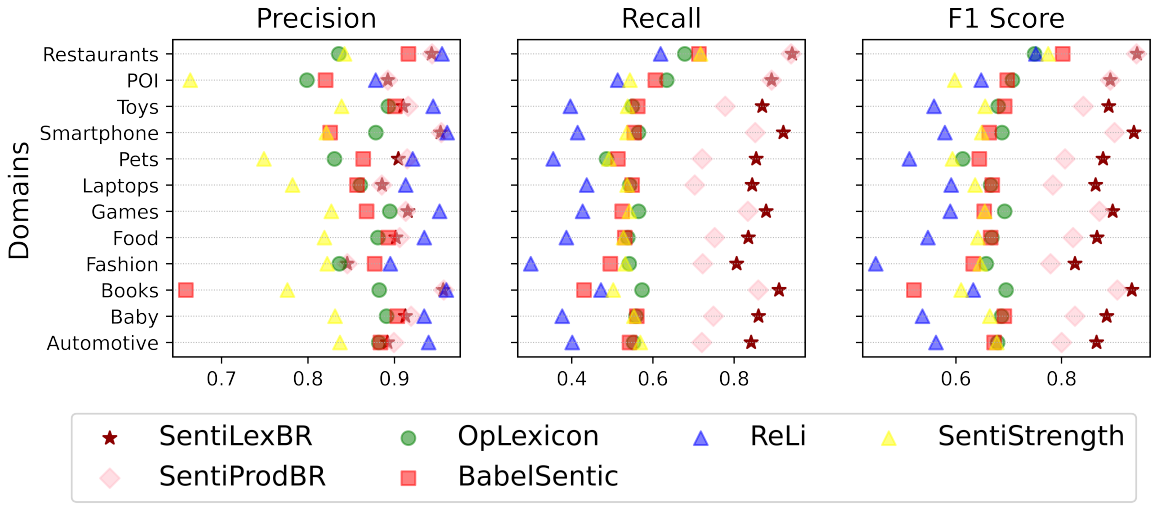

Fig. 3. Metrics of all approaches on all domains.

Our best domain-specific lexicons reached 0.94 for F1-score in the domain Restaurants. Conversely, our lowest domain-specific F1-score was achieved in the category of *Fashion* products with 0.83 versus the second best method, SentiProdBR, that achieved 0.78. 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.

4.3 Estimating Factor τ

The goal of this set of experiments is to estimate the best value of the weight 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 Fig. 4, where we plot F1-score averaged across all 12 domains when varying factor τ from 0 to 0.9. As shown, τ = 0.3 produces the best average across all domains, with approximate F1-score gains of 0.02, when compared to default $\tau = 0$.

4.4 Misspelled Words

In order to give an intuitive feel for the robustness of SentiLexBR, Table IV shows the 5 most misspelled words from our lexicons manually identified. The words correctly identified by the methods are represented by a ✓ symbol and, otherwise, the words are represented by a × symbol. OpLexicon, BabelSentic, SentiStrength and ReLi could not identify any of the top 5 misspelled words because all these lexicons only consider correctly spelled words. This is an important limitation of these methods, as although the words do not exist in an official dictionary, these terms are strong indications of

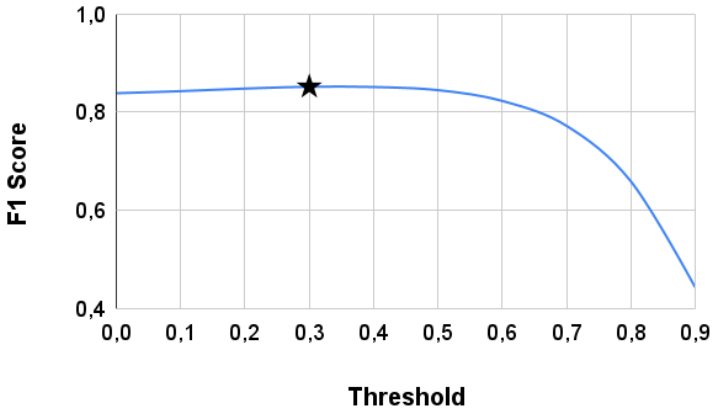

Fig. 4. Influence of the factor τ across all 10 domains.

sentiment in the texts. For example, the term *otimo* (good) is the 46th most frequent in the Cellphone domain. Only SentiProdBR could identify correctly the misspelled adjective words, but it could not identify the verbs misspelled words.

Table IV. Top 5 misspelled words.

	otimo (good)	*rapido* (fast)	*exelente* (great)	*estao* (are)	*facil* (easy)
SentiLexBR	✓	✓	✓	✓	✓
SentiProdBR	✓	✓	✓	×	✓
OpLexicon	×	×	×	×	×
BabelSentic	×	×	×	×	×
SentiStrength	×	×	×	×	×
ReLi	×	×	×	×	×

4.5 Part-of-Speech Analysis

The goal of this section is to analyze the relevance of adjectives and verbs to build sentiment lexicons in different domains. Part-of-speech (POS) is a category of words which have similar grammatical properties. SentiLexBR use only adjectives and verbs in the construction of the set of sentiment lexicons. The frequency distribution of part-of-speech tags are shown in Fig. 5. The horizontal axis denote the domains and the vertical axis denotes the corresponding frequency of each type of POS.

As shown in Fig. 5, more than 59.4% of part-of-speech is verb. Therefore, SentiLexBR has a greater representation of terms when compared to SentiProdBR, which considers only sentiment lexicons as adjectives. Interestingly, there are more verbs than adjectives in almost every domain. The only exception is the point-of-interest (POI) domain. Our intuition is that user reviews regarding tourist attractions are more descriptive than in the other categories and, consequently, the use of adjectives becomes more frequent.

4.6 Domain Specific Terms

Sentiment lexicons frequency analysis is one of the most fundamental analytic methods in semantic analysis [Yang et al. 2019]. The frequency distribution of terms is quite different for each domain. In order to clearly illustrate the difference, we count the frequency of each sentiment lexicon that appears

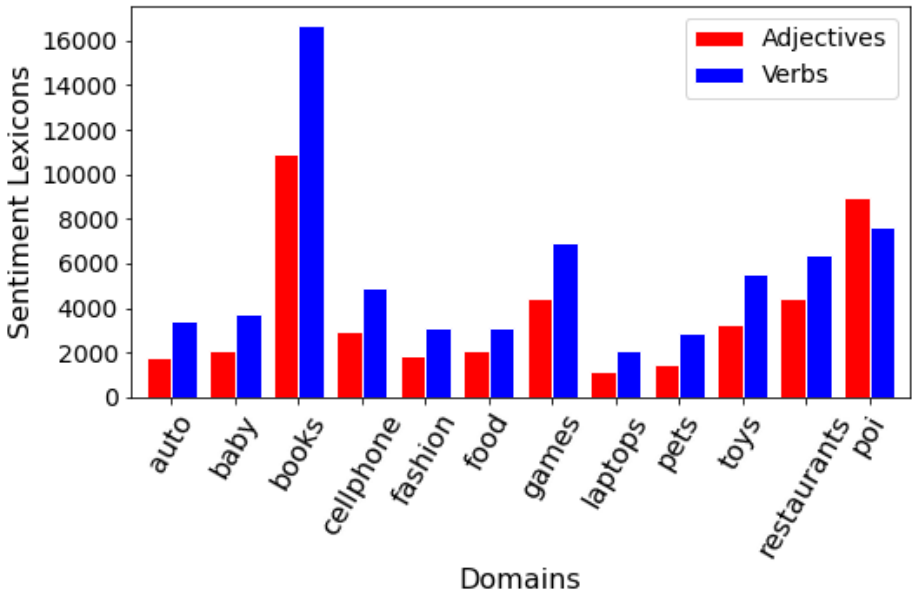

Fig. 5. Distribution of POS tag for each domain.

in a single domain. There are 48,080 unique sentiment lexicons and that corresponds to 43% of the total.

The distribution of unique sentiment lexicons are shown in Fig. 6. The horizontal axis denotes the number of domains and the vertical axis denotes the corresponding frequency of distinct sentiment lexicons. We can observe that there is a significant amount of sentiment lexicons that appear exclusively in a single domain and that the set of lexicons that appear in all 12 domains is low. These results support our argument of the importance of generating lexicons for specific domains.

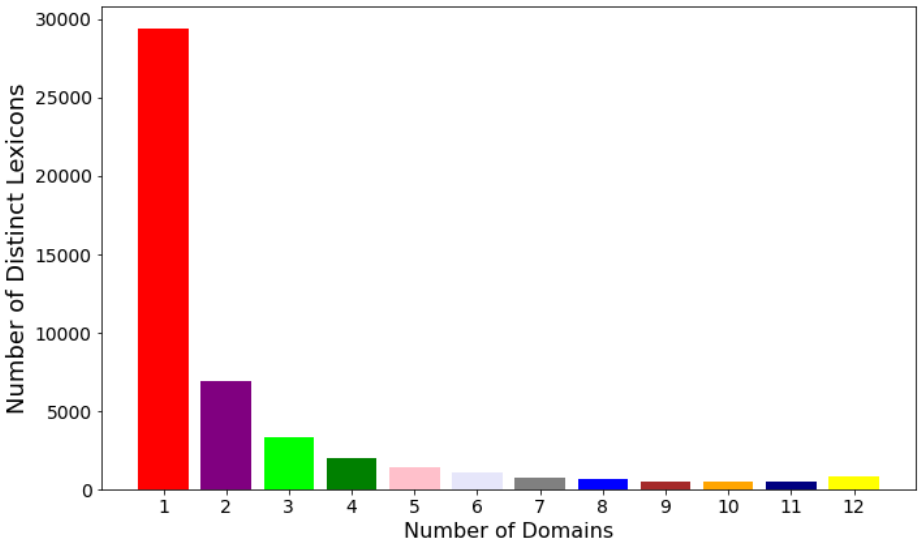

Fig. 6. Distribution of unique sentiment lexicons.

We evaluated the distribution of the set of lexicons in Fig. 6 that appear in a single domain (red color) and found that the Books domain has 44.3% of this set. The main reason for this is that vocabulary used in books in different categories can vary greatly. For example, the vocabulary used in science fiction books is very different from medical books.

From Fig. 6 we find that less than 1% of the lexicon set appears in all domains (yellow color). Interestingly, although this set appears in all domains, there is still a variation in the polarity of terms according to the domain. For example, the word "*fraco*" (weak) commonly appears with the negative polarity. In the sentence "*O ponto fraco do laptop é a duração da bateria*" (Laptop's weak point is the battery life) from Laptops domain, the term "*fraco*" (weak) indicates a negative opinion. However, in the sentence "*Achei o café um pouco fraco em sabor e aroma, ou seja, ideal para quem gosta de um café menos amargo*" (I found the coffee a little weak in flavor and aroma, that is, ideal for those who like a less bitter coffee), the term "*fraco*" (weak) indicates a positive opinion.

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 domain-specific lexicons. To achieve this goal, we employed probabilities to calculate the sentiment strength of each word. We evaluated our method with five 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 SentiLexBR as a large domain-specific sentiment lexicon for 12 different categories. In future work, we plan to experiment with using deep learning and word embeddings for sentiment lexicon creation.

REFERENCES

Ahire, S. A survey of sentiment lexicons. *Computer Science and Engineering IIT Bombay, Bombay*, 2014.

Almatarneh, S. and Gamallo, P. A lexicon based method to search for extreme opinions. *PLOS ONE* 13 (5): 1–19, 05, 2018.

Amora, P. R. P., Teixeira, E. M., Lima, M. I. V., Amaral, G. M., Cardozo, J. R. A., and de Castro Machado, J. An analysis of machine learning techniques to prioritize customer service through social networks. *Journal of Information and Data Management* 9 (2): 135–135, 2018.

Birjali, M., Kasri, M., and Beni-Hssane, A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 2021.

Bos, T. and Frasincar, F. Automatically building financial sentiment lexicons while accounting for negation. *Cognitive Computation*, 2021.

Catelli, R., Pelosi, S., and Esposito, M. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics* 11 (3): 374, 2022.

Chaturvedi, I., Cambria, E., Welsch, R. E., and Herrera, F. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion* vol. 44, pp. 65–77, 2018.

de Melo, T. Sentiprodbr: Building domain-specific sentiment lexicons for the portuguese language. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*. SBC, pp. 349–354, 2021.

de Melo, T., da Silva, A. S., de Moura, E. S., and Calado, P. Opinionlink: Leveraging user opinions for product catalog enrichment. *Information Processing & Management* 56 (3): 823–843, 2019.

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

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

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

Hutto, C. and Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8, 2014.

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

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

Nusko, B., Tahmasebi, N., and Mogren, O. Building a sentiment lexicon for swedish. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, July 11, 2016, Krakow, Poland*. Number 126. Linköping University Electronic Press, pp. 32–37, 2016.

Oliveira, M. d. and Melo, T. d. An empirical study of text features for identifying subjective sentences in portuguese. In *Brazilian Conference on Intelligent Systems*. Springer, pp. 374–388, 2021.

Park, S.-M., Na, C.-W., Choi, M.-S., Lee, D.-H., and On, B.-W. Knu korean sentiment lexicon: Bi-lstm-based method for building a korean sentiment lexicon. *Journal of Intelligence and Information Systems* 24 (4): 219–240, 2018.

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

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

Thelwall, M. Heart and soul: Sentiment strength detection in the social web with sentistrength, 2017. *Cyberemotions: Collective emotions in cyberspace*, 2014.

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

Xiang, R., Jiao, Y., and Lu, Q. 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)*, 2019.

Yang, L., Zhai, J., Liu, W., Ji, X., Bai, H., Liu, G., and Dai, Y. Detecting word-based algorithmically generated domains using semantic analysis. *Symmetry* 11 (2): 176, 2019.

Zhang, S., Wei, Z., Wang, Y., and Liao, T. Sentiment analysis of chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems* vol. 81, pp. 395–403, 2018.