Creating a Portuguese context sensitive lexicon for sentiment analysis

Mateus Tarcinalli Machado[1][0000−0003−1848−3082], Thiago A. S. Pardo[2][0000−0003−2111−1319], and Evandro Eduardo Seron Ruiz[1][0000−0002−7434−897X]

[1] Department of Computing and Mathematics – FFCLRP
University of São Paulo – Ribeirão Preto, Brazil
{mateusmachado, evandro}@usp.br
[2] Interinstitutional Center for Computational Linguistics (NILC)
University of São Paulo – São Carlos, Brazil
taspardo@icmc.usp.br

Abstract. This work presents and evaluates a new Portuguese context sensitive lexicon for sentiment analysis. Distinctive composition approaches to produce lexicons from established ones were also tested. The experiments were carried out with the corpus ReLi, composed of opinative texts about books, and with the following sentiment lexicons: Brazilian Portuguese LIWC, Opinion Lexicon, and SentiLex.

Keywords: Polarity Detection · Lexicon Creation.

1 Introduction

Sentiment Analysis uses Natural Language Processing techniques to extract and classify opinions, emotions, evaluations, and attitudes related to products, services, organizations, people, events and subjects expressed in free text. A Sentiment Analysis (SA) first application area was product evaluation, an area that had a big boost with the advent of Web 2.0, following the growth of electronic commerce and the more active participation of consumers and users on the web. Since the seminal paper by Pang Lee [8], literature has presented SA as a rather complex task. The major obstacles may be divided into: a) analyzing the meaning of sentiments and b) detecting the suitable sentiment polarity. Some of these tasks are further discussed by Hussein [6].

SA can become more demanding if, for instance, the text mentions various characteristics about one assessed entity. In this case, the evaluator may qualify this single item with different feelings. This way, it may also occur that certain ratings can be positive or negative, depending on the analyzed *aspect* of the entity. Aspect-Based Sentiment Analysis (ABSA) is a fine grained form of SA aiming to identify the aspects of given entities and the related sentiments [16]. Pavlopoulos and Androutsopoulos [9] suggest that most ABSA systems subdivide this deeper SA processing into three subtasks, which are: (i) aspect term extraction (detection of product's characteristics described in textual format),

(ii) aspect term sentiment estimation (identification of the sentiment polarity – usually positive, negative or neutral – related to each aspect), and (iii) aspect aggregation (not always present, performs the grouping of identified aspects that are synonymous or near-synonyms).

Specifically for the aspect term sentiment estimation subtask (ii), there are two main approaches: a) the lexicon-based approaches, and; b) machine learning ones [15]. In the first, the polarity of a text is identified by analyzing the semantic orientation of the words and phrases composing the text. This orientation is obtained through dictionaries containing word lists and their polarities. For machine learning approaches it is necessary to build a classifier. This is generally accomplished by using examples of manually classified texts, which labels these methods as supervised classification task.

Although one may find plenty of research related to lexicon-based sentiment analysis for the English language, this paper focuses in the creation of a Portuguese context sensitive lexicon. The suggested approach is tested on opinative review texts about books, found in the only known Portuguese annotated corpus for aspects, the ReLi corpus [4]. The results of this approach are also compared to other methods of lexicon creation based on existing sentiment lexicons.

The rest of this paper is organized as follows: Section 2 discusses lexicon-based SA and some Portuguese lexicon creation processes. Section 3 describes the method used to create the proposed context sensitive lexicon and the dataset used to test it. Section 4 compares different SA lexicons performance under lexicon-based approaches. Finally, Section 5 draws some conclusions from previous results and outlines directions for future work.

2 Related Work

The Bing Liu book [7] encapsulates a great part of the research on concepts related to the Sentiment Analysis area on its origins. In his book, he presents examples of applications, related problems, methods for sentiment analysis, aspect extraction, generation of lexicons, and opinion summarization, among others.

Souza *et al.* [14] created a new lexicon (OpLexicon) combining lexicons using three different techniques. A first lexicon was created using the analysis of an annotated corpus. Another one was created by searching for synonyms and antonyms in a thesaurus. A final lexicon was obtained through the automatic translation of *Liu's English Opinion Lexicon* [5].

Another important work was the one that created the lexicon Sentilex [12]. The development of the lexicon occurred in two stages: firstly, a dictionary of adjectives was created with their respective polarities. From these adjectives, a set of lexical-syntactic rules was manually created. These rules were applied to a large collection of n-grams. The frequencies of the adjectives and found rules were used as inputs for a statistical classifier. In a second step, polarities were assigned to the new found adjectives and this list was expanded through the exploration of synonym graphs.

Lexicons play a major role in SA. An important related work is the paper by Taboada *et al.* [15], which presents the Semantic Orientation CALculator (SO-CAL) technique. SO-CAL calculates the polarity of a text using word lexicons marked with their semantic orientation (polarity and force). This system also identifies words that intensify or reverse (negation words) these polarities. Balage *et al.* [1] used SO-CAL to evaluate the Brazilian Portuguese LIWC dictionary for SA, comparing this lexicon with SentiLex and OpLexicon.

The International Workshop on Semantic Evaluation (SemEval) [10] is an important venue for researchers in SA. This workshop always presents new datasets for analysis, providing the basis for comparing results of different methods.

3 Material and methods

3.1 Methods

This work is focused in sentiment polarity identification using sentiment lexicons. In this next subsection, **LexReLi**, we introduce our methodology, constructing the proposed new context sensitive lexicon. Then, at **Lexicon combination** subsection, we investigate lexicons produced by different combination techniques that involve three well-known sentiment lexicons: Brazilian Portuguese LIWC, Opinion Lexicon and SentiLex.

LexReLi Here we describe the construction of the Lexicon ReLi (LexReli), specialized in identifying the polarity of aspects in opinion texts about books. For this task, we extended the dataset, collecting, from the skoob website, more reviews about the books mentioned in ReLi (see more about ReLi in Section 3.2). This corpus is composed of 6,698 reviews, 51,148 phrases, and 980,640 words. For the construction of this lexicon, we combined some strategies of aspect identification with polarity detection and applied it to this corpus. The objective was to create a lexicon composed only by adjectives, where their polarities are identified through the context they belong to, *i.e.*, , the sentence polarity.

Both corpora, the original ReLi and its extended version, have been submitted to a preprocessing phase. Two tokenizers [2] were applied to the extended ReLi, one to split the reviews into sentences and another to divide the sentences into tokens. After that, both corpora were submitted to a part-of-speech tagger [3]. Here are the tasks pursued to construct the LeXReLi lexicon:

Constructing the LeXReLi

1. **Aspect identification.** The first step was the identification of phrases that have nouns close to adjectives, as this combination tends to indicate an aspect (noun) close to its characteristic (adjective);
2. **Polarity detection.** The selected sentences went through a process of polarity detection. We then applied the Adjectives Preference method (as explained in Section 3.3) with a lexicon obtained from the combination of

LIWC, SentiLex, and OpLexicon (see Section 4 for results that justify this preference);
3. The frequency of adjective occurrence in positive and negative sentences was computed;
4. If one adjective appears more often in positive phrases, this polarity was assigned to it; otherwise, a negative polarity was assigned. If the difference between the number of times it appears in positive and in negative sentences is less than two, the adjective was not included in our lexicon.

Now, in order to verify if the combination of well-known lexicons can enhance the performance of SA, we investigated the construction of a new lexicon from a combination of three representative ones.

Lexicon combination. Three frequently used lexicons in similar researches have been combined to form a new SA lexicon. They are: Opinion Lexicon [14, 13], SentiLex [12] and the Brazilian Portuguese Linguistic Inquiry and Word Count 2007 (LIWC) [1]. We have created combined lexicons from the three aforementioned dictionaries, using two combination methods.

Combined Lexicons. In a first approach, we only combine the dictionaries, one after the other, ignoring possible disagreements between them, prevailing, in these cases, the classification adopted by the first dictionary added to the set. This way, word order brought by the dictionary combination will interfere in the final result. This fact led us to create six possible combinations, altering the order of inclusion of the three dictionaries. In total, 6 lexicons were generated with equal amounts of words, but slight differences in polarities.

Conciliated Lexicon. In a second approach, we constructed a new lexicon based on the previous three aforementioned. It was established that a word is added to the new dictionary if it appears in only one of them, or if the word has the same polarity in at least two dictionaries. This way, eventual polarity disagreements among dictionaries are solved. Table 1 presents the final composition of the sentiment lexicons.

Table 1. Polarity distribution found on lexicons.

Lexicon	Positive	Neutral	Negative	Total
OpLexicon	8,595	8,974	14,550	32,119
SentiLex	20,478	7,600	51,112	79,190
LIWC	12,376	0	14,612	26,988
LexReLi	1,091	0	452	1,543
Conciliated	32,543	11,155	62,676	106,374
Combined *	34,433	12,636	64,638	111,707
LexReLi + Combined *	34,974	12,244	64,685	111,903

* average of the six obtained lexicons.

3.2 The Dataset

For our work on aspect-based polarity detection experiments, we worked with the ReLi corpus [4]. The ReLi is a Portuguese book review corpus composed by 1,600 reviews of 14 books collected from 'skoob' website[3]. Skoob is a collaborative social network for book readers. The reviews on skoob were manually annotated for opinion presence, identified aspects, and their respective polarities.

From the ReLi corpus, we selected only phrases that contained at least one aspect and their respective polarity. That way, we worked with a set of 2,675 aspects and respective polarities (2,089 positives and 586 negatives), showing an unbalanced sentiment polarity distribution.

3.3 Polarity Detection

We implemented four approaches for polarity detection in a two-step method. We combined the Aspect Based method with three phrasal level methods: Words Polarities, Adjectives Polarities and Adjectives Preference. As a first step we apply the Aspect Based method; if it can not identify the polarity of the sentence the algorithm follows a second step and uses one of the three phrasal level methods, which are: words polarities, adjectives polarities and adjectives preference. We applied our approach using the mentioned sentiment lexicons in Section 3.1.

Aspect Based. Demonstrated in Algorithm 1, this method tries to find the polarity related to every aspect in the sentences. The algorithm locates in each sentence every aspect marked in the ReLi corpus (line 5). Given the aspect, the algorithm searches for the nearest adjective and verifies its polarity in the lexicon (lines: 13, 14). To deal with negation, a list of negation terms was obtained from the Brazilian Portuguese 2007 LIWC lexicon (line: 3). If there is a negation word between the aspect and the adjective, the polarity of the adjective is reversed (lines: 15, 16). The polarities of adjectives found close to aspects are then summed up (line: 18). The final result of this sum indicates the polarity of the sentence (line: 21).

Words Polarities. This is a simple method of polarity identification using a lexicon. Each word of the phrase is sought in a sentiment lexicon. The polarity value found for each word in a sentence is then summed. If the value obtained from this sum is greater than zero, it indicates a positive polarity; if it is equal to zero, we have a neutral one; and, if less than zero, we have a negative polarity. As in TABOADA *et al.* [15] SO-CAL method, negation words modify polarity of nearest words. When the algorithm finds a word from LIWC negation word list, the polarity of the next found word is reversed.

Adjectives Polarities. Similar to the previous method, but, in this case, only the adjective's polarities found in a sentence were added. Negation was also taken into account, reversing the polarity of the adjective closest to the negative term.

[3] https://www.skoob.com.br/

Algorithm 1 Aspect Based Polarity Detection algorithm

```
1: function aspectBased(sentence)
2:     adjectives ← POSTagger(sentence).getAdjectives()
3:     negationWords ← LIWC.getNegationWords()
4:     polaritySum ← 0
5:     for aspect in locateAspects(sentence) do
6:         wordPosition ← 0
7:         aspectLastWordPosition ← getAspectLastWordPosition(aspect)
8:         for word in sentence do
9:             if wordPosition <= aspectLastWordPosition then
10:                continue
11:            if word in negationWords then
12:                negationFlag ← True
13:            if word in adjectives and word in lexicon then
14:                adjectivePolarity ← lexicon.getWordPolarity(word)
15:                if negationFlag then
16:                    adjectivePolarity ← adjectivePolarity * −1
17:                    negationFlag ← False
18:                polaritySum ← polaritySum + adjectivePolarity
19:                break // inside for command
20:            wordPosition ← wordPosition + 1
21:    return polaritySum
```

Adjectives Preference. This method is based on both previous methods. Initially, similar to the second method, only the polarities of the adjectives were taken into account. The difference in this method is that, when no adjective is found in a lexicon, the algorithm analyzes the polarities of every word, just as in the first method. Negation was treated in the same way as in the Adjective Polarities method.

4 Results

In our experiments, we combine the method of detecting aspect polarity with the three above methods for detecting polarity. We apply these methods initially using the lexicons individually and then using the combined ones. Our intention is to analyze both, the performance of the algorithms and, the approaches used to create a lexicon either as in LexReLi or as a combination of other lexicons.

To evaluate the results of the experiments, we compared the polarities found by the implemented methods with those indicated in the ReLi corpus. This was accomplished using an evaluation methodology similar to the one used in the SemEval workshops [11]. We calculated the accuracy for each experiment, defined as the number of correctly predicted polarities divided by the total number of polarity aspects found on ReLi corpus.

4.1 Individual Lexicons

The results obtained with each method, individually using the lexicons, are shown in Fig. 1. These experiments evaluate the methods of identifying polarity and our strategy for the creation of LexReLi lexicon.

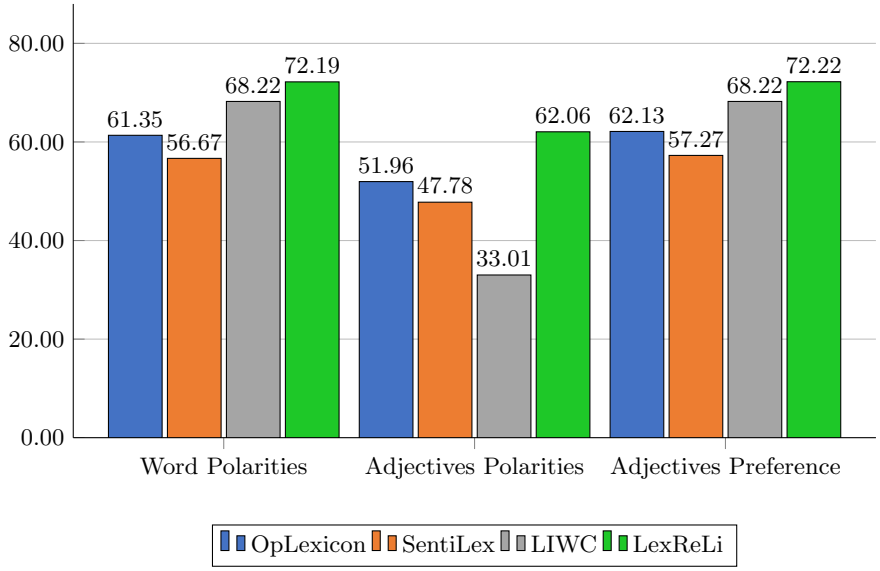

Fig. 1. Accuracy (%) for experiments with individual lexicons.

In the first experiment, we combined the Aspect Based polarity analysis with the phrase polarity analysis identified by the polarity of the words, through the method Words Polarities. Among the three standard lexicons we tested, the LIWC obtained the best result with an accuracy of 68.22%, showing that the size of the lexicon does not define its quality. The largest lexicon, the SentiLex with 79,190 words, obtained the worst result, 56.67% of accuracy. The intermediary size lexicon, OpLexicon, with 32,119 words, obtained 61.35% of accuracy. The proposed LexReLi lexicon, specialized in the literary context, despite the 1,543 words, obtained the best result in this experiment, with 72.19% of accuracy.

In our second experiment, we evaluated, on phrase level, the detection of polarity analyzing only the adjectives, using the Adjectives Polarities method. In this experiment, the LexReLi obtained the best result, reaching 62.06% of accuracy. The Lexicon OpLexicon obtained an accuracy of 51.96%, followed by SentiLex with 47.78%. Finally, came the LIWC with 33.01%.

As a final experiment, we evaluated the Adjectives Preference method that gives priority to adjectives in the sentence-level analysis. The results were slightly higher than the first experiment, with LexReLi obtaining the best result with

an accuracy of 72.22%. The LIWC obtained 68.22%, OpLexicon 62.13% and SentiLex 52.27%.

4.2 Combined Lexicons

As explained, we used various approaches to combine lexicons. We have a first lexicon that has undergone a conciliation process where possible conflicts between combined lexicons have been solved. We also have six lexicons formed by the combination of LIWC, OpLexicon, and SentiLex, where what has changed between one and other was the order the words were included in the lexicons. Finally, we presented LexReLi combined with these same six last lexicon combinations. In these experiments, we tried to evaluate these lexicon construction techniques. Fig. 2 shows the obtained results.

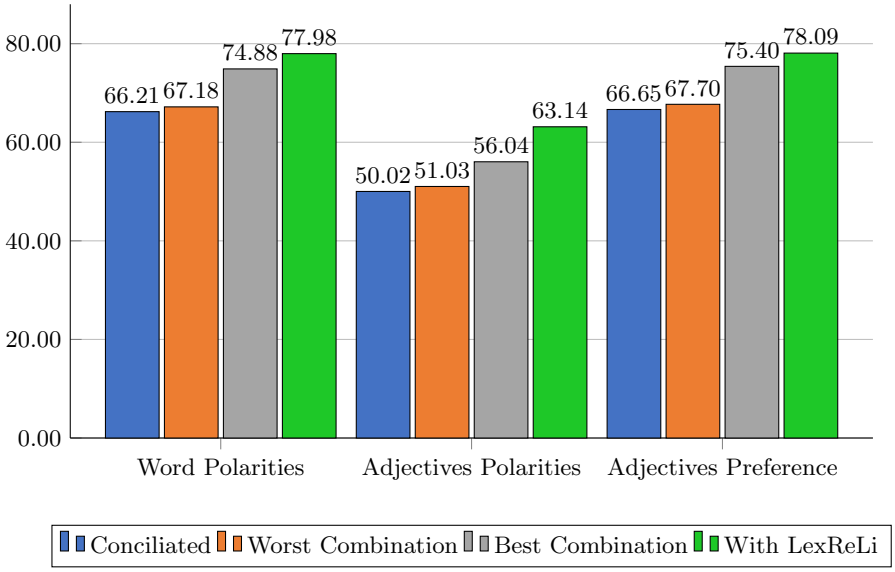

Fig. 2. Accuracy (%) for experiments with the combined lexicons.

Regarding the methods of analysis, the results were very similar to the experiments with the individual lexicons. The adjectives-preferred approach yielded the best results, but these were little better than the approach that analyzes the polarity of words. The method that analyzes only the adjectives obtained the lowest results.

The lexicon obtained in the conciliation process, although this seems to be the correct way to combine lexicons, was the one that obtained the lowest results. It obtained 66.21% of accuracy when analyzed through word polarities, 50.02% only by adjectives, and 66.65% when using the method that gives preference to adjectives.

In the combination of lexicons, we obtained, for each method of analysis, six different results. We present here only the smallest and highest obtained results. For the lexicon combinations with LexReLi, we only present the best result, since the difference between the worst and the best results was not as expressive as the previous combinations.

For word polarities method, the combination made in the order LIWC + SentiLex + OpLexicon obtained the best accuracy with 74.88%. The lowest result was obtained with SentiLex + OpLexicon + LIWC with 67.18% of accuracy. The LexReLi + LIWC + SentiLex + OpLexicon combination achieved the best result of the experiments with the method, reaching an accuracy of 77.98%.

The method that uses only adjectives for analysis, as well as in the experiment carried out with the individual lexicons, obtained the lowest results. The best result, 63.14% of accuracy, was obtained for the combination LexReLi + LIWC + OpLexicon + SentiLex. Without LexReLi, the LIWC + OpLexicon + SentiLex combination achieved 56.04% and the SentiLex + OpLexicon + LIWC combination had the lowest result with 51.03% of accuracy.

In our last experiment, we tested again the preference for adjectives method and obtained the highest results among all the performed experiments, although they were little superior to the method that analyzes only the polarity of words. The LexReLi + LIWC + OpLexicon + SentiLex combination obtained the highest result with 78.09% of accuracy. Without LexReLi, the highest result was obtained by LIWC + SentiLex + OpLexicon with 75.40%, and the lowest by SentiLex + OpLexicon + LIWC with an accuracy of 67.70%.

5 Conclusion

We may conclude that the approaches used to create and combine lexicons were effective, given the evident improvement in the results of the application of the methods using these lexicons. The method used to create LexReLi may be easily used for the creation of lexicons for different contexts. Improvements in the methods of polarity analysis used in its creation would also imply improvements in the resulted lexicon.

The lexicon combination method that used a conciliation approach proved to be less effective because, in addition to being more laborious in creation, it obtained inferior results in relation to the methods that used a simple combination of lexicons.

In the combination approaches, we obtained indications that it is a good practice to give priority to the lexicon that obtained the best individual result, since in the experiments we observed that the combinations initiated by the LIWC lexicon achieved better results.

Although slightly superior, the results of the preference to the adjectives method are promising. Improvements may be implemented in the two steps of the approach. In the first stage, one may work with advanced methods for detecting words that qualify aspects. In the second step, it is possible to improve or apply other methods for detecting polarity at the phrase level.

References

1. Balage Filho, P.P.P., Aluísio, S., Pardo, T.T.: An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. 9th Brazilian Symposium in Information and Human Language Technology STIL pp. 215–219 (2013)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc. (2009)
3. Fonseca, E.R., Rosa, J.L.G.: Mac-morpho revisited: Towards robust part-of-speech tagging. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. pp. 98–107 (2013)
4. Freitas, C., Motta, E., Milidiú, R.L., César, J.: Vampiro Que Brilha... Rá! Desafios Na Anotação De Opinião Em Um Corpus De Resenhas De Livros. In: XI Encontro de Linguística de Corpus (ELC 2012). São Paulo (2012)
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 168–177. ACM (2004)
6. Hussein, D.M.E.D.M.: A survey on sentiment analysis challenges. Journal of King Saud University - Engineering Sciences (2016)
7. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1), 1–167 (2012)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2), 1–135 (2008)
9. Pavlopoulos, J., Androutsopoulos, I.: Aspect Term Extraction for Sentiment Analysis: New Datasets, New Evaluation Measures and an Improved Unsupervised Method. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL. pp. 44–52 (2014)
10. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., , Androutsopoulos, I.: SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 486–495. Denver, Colorado (2015)
11. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: Semeval-2015 task 12: Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 486–495. Association for Computational Linguistics, Denver, Colorado (June 2015)
12. Silva, M.J., Carvalho, P., Costa, C., Sarmento, L.: Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis. Technical Report TR 1008 University of Lisbon Faculty of Sciences LASIGE **6694**(December) (2010)
13. Souza, M., Vieira, R.: Sentiment Analysis on Twitter Data for Portuguese Language. Proceedings of the 10th international conference on Computational Processing of the Portuguese Language pp. 241–247 (2012)
14. Souza, M., Vieira, R., Chishman, R., Alves, I.M.: Construction of a Portuguese Opinion Lexicon from multiple resources. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology pp. 59–66 (2011)
15. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics **37**(2), 267–307 (2011)
16. Thet, T.T., Na, J.C., Khoo, C.S.: Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science **36**(6), 823–848 (2010)