Comparing Approaches to Subjectivity Classification: A Study on Portuguese Tweets

Silvia M.W. Moraes, André L.L. Santos(✉), Matheus Redecker, Rackel M. Machado, and Felipe R. Meneguzzi

Faculdade de Informática, PUCRS, Avenida Ipiranga, 6681, Prédio 32, 90619-900 Porto Alegre, RS, Brazil
{silvia.moraes, felipe.meneguzzi}@pucrs.br,
andre.leonhardt.santos@gmail.com,
{matheus.redecker, rackel.machado}@acad.pucrs.br

Abstract. In this paper, we compare lexicon-based and machine learning-based approaches to define the subjectivity of tweets in Portuguese. We tested SentiLex and WordAffectBR lexicons, and Sequential Machine Optimization and Naive Bayes algorithms for this task. In our study, we used the Computer-BR corpus that contains messages about the technology area. We obtained better results using the Comprehensive Measurement Feature Selection method and the Sequential Machine Optimization algorithm as the classifier. We achieved considerable accuracy when we included the polarities of words in the vector space model of tweets.

Keywords: Subjectivity classification · Sentiment analysis · Natural language processing

1 Introduction

In the past decade, people have used social web to express and share their 'sentiments' about products and services. Texts published in social media (e.g., Twitter, Facebook, forums, blogs, and user forums) have become important sources of information for organizations. The analysis of these snippets of text is a way of monitoring the opinion and response from the clients of these organizations [1]. The area of research that automatically performs this processing is known as Sentiment Analysis or Opinion Mining. In this area, textual information can be categorized into two main types: facts and opinions. Opinions, unlike facts, describe people's sentiments, appraisals, or feelings toward entities, events, and their properties. The task of defining whether a sentence expresses an opinion or a fact can be treated as a classification problem. This task is called subjectivity classification [2]. The subjectivity classification is a stage that precedes the Opinion Mining. When used, it improves Opinion Mining performance by preventing noisy and irrelevant extraction [3, 4]. In approaches that use machine learning algorithms for polarity classification, the improved results can be attributed to the balancing of training sets. Some authors mention that the imbalance of such approaches is caused by the class of objective sentences, which usually has a larger number of samples [5].

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 86–94, 2016.
DOI: 10.1007/978-3-319-41552-9_8

In this paper, we compare two traditional approaches to subjectivity classification: based on lexicon, and machine learning algorithms [10]. We tested SentiLex and WordAffectBR lexicons, and Sequential Machine Optimization (SMO) and Naive Bayes algorithms to determine the subjectivity of the tweets from Computer-BR corpus, we that built for this study. This corpus is composed of messages in Portuguese language about the technology area. We categorized the tweets according to their sentiment orientations (polarities). We considered as subjective those sentences with positive or negative polarity, and as objective sentences the remaining ones. In the approach using machine learning, we tested a new method for feature selection: the Comprehensive Measurement Feature Selection (CMFS), indicated for unbalanced corpora [15, 16]. Our best results were obtained using this method and the SMO algorithm as classifier. We achieved an accuracy on average of 78.51 % when we included the polarities of words in the vector space model of tweets. Besides the results obtained from the research described in this paper, we also consider Computer-BR corpus as one of our contributions to the sentiment analysis area.

This paper is organized into 7 sections. In Sect. 2, we present some works related to ours. In Sect. 3, we introduce the subjectivity classification task and approaches used to treat it. In Sect. 4, we describe the corpus we created and the pre-processing realized. In Sects. 5 and 6, we detail the approaches used to define the subjectivity of the sentences. In Sect. 7, we present our conclusions. Finally, the acknowledgment and references are presented.

2 Related Work

The opinion mining from web texts is a non-trivial Natural Language Processing task, for this reason it has received much attention. Most literature on sentiment analysis for Portuguese language addresses polarity classification at sentence and aspect (feature) level. In applications at the sentence-level, in which sentences are classified as positive, negative, or neutral, the accuracy ranges from 55 % to 71,79 % [6–8, 17]. In these applications, the best results were obtained from the Sequential Minimal Optimization (SMO) [6, 7] and Naive Bayes [3] algorithms. In lexicon-based approaches, the accuracy for Portuguese language is around 57.3 % [9]. It is worth mentioning that linguistic resources for sentiment analysis in Portuguese language are still developing [13, 14]. The lack of benchmarks corpora, for example, makes more challenging the comparison of results.

According to researchers in the area, the performance of these tasks improves when we perform the subjectivity classification in a previous step [3, 4]. Initially, Kamal [3] classifies sentences in English as subjective and objective, and later he performs features-based sentiment analysis for the sentences defined as subjective. In the subjectivity classification stage, the author has reached an accuracy of 91.6 % with the Naive Bayes algorithm. Lambov et al. [18] says that the classification of subjectivity is a specific domain problem, showing that the results fall around 20 % for across domains.

3 Subjectivity Classification

We understand the subjectivity classification as a task to define whether a sentence is objective or subjective. Objective sentences express facts, while subjective sentences express opinions. Opinions, unlike facts, describe people's sentiments, appraisals, or feelings toward entities, events, and their properties [2]. In our study, we determined the classification of sentences according to their polarities (positive, negative, or neutral), that indicated the sentiment orientation. The algorithm assigns the polarity by the presence of certain adjectives, verbs, and nouns in the sentences. For example, words such as *fast* (adjective), *to love* (verb) and *joy* (noun) expressed sentiments with positive polarity, whereas words such as *slow* (adjective), *to hate* (verb) and *sadness* (noun) indicated negative polarity to the sentences. The polarity was neutral when the word was neither positive nor negative. We considered as subjective sentences those with positive or negative polarity. The remaining sentences (with neutral polarity) were considered objective.

According to Madhat et al. [10], in the sentiment analysis area, the most common techniques use approaches based on lexicons or machine learning algorithms. We used the lexicons SentiLex-PT [11] and WordNet-Affect [12], and the SMO and Naive Bayes algorithms in our study. We chose these resources and techniques, because they are widely known [3, 6, 7] and achieves good results. Figure 1 shows the pipeline that we implemented to define the subjectivity of the sentences. In the following sections this pipeline is detailed.

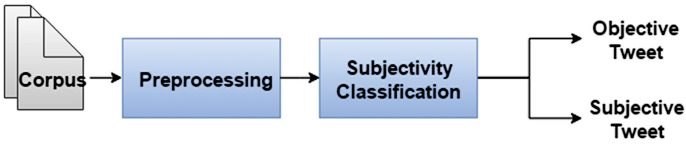

Fig. 1. Subjectivity classification

4 Computer-BR Corpus and Preprocessing

To perform this study, we built a corpus in Portuguese, the Computer-BR. This corpus consists of 2,317 tweets that was extracted in the period from January 1st to September 18th, 2015. The Corpus has 34,437 tokens and 4,653 types. To build it, we used keywords related to computers: notebook, analysis, testing, and so on. We relied on 4 human annotators who defined the polarity of the tweets, 3 of them participated in the whole process of annotation and the fourth decided the final polarity only in cases of disagreement. Although our study is only on subjectivity classification, the annotation considered 4 classes: irony (−2), negative (−1), neutral (0) and positive (1). We built the Computer-BR also for its application in future works. The final kappa index was 0.69. It is worth mentioning that three annotators were from the Computer Science area and one from the Linguistics area. The Table 1 shows the polarity distribution in the corpus.

As the treatment of irony is not included in this work, tweets classified into the irony class became negative. Thus, 443 was the total amount of tweets classified as

Table 1. Sentiments distribuition in the Computer-BR corpus.

Classes	#Tweets (%)
Irony	39 (1,7 %)
Negative	404 (17,4 %)
Neutral	1,677 (72,4 %)
Positive	197 (8,5 %)

Table 2. Examples of tweets from Computer-BR corpus.

Tweet	**Polarity**
TO IRADOO! (I'M ANGRYY!)	negative
Bateria do meu notebook já era... (Battery of my notebook is gone...)	neutral
Apaixonada pelo meu Notebook😍❤ (I am in love with my Notebook😍❤)	positive
Aiii que maravilha, meu notebook parou de ligar! (Ahhh, wonderful, my notebook no longer switches on!)	irony, negative

negative in our study. We intend to investigate the polarity classification including irony in a future work. Table 2 shows some examples of tweets from the corpus.

Web texts, especially those posted on microblogs, have a lot of noisy and uninformative pieces (HTML tags, scripts and advertisements). In these texts, it is common the repetition of vowels, punctuation problems, misspelling, emoticons, colloquialism, unconventional use of upper and lower cases, and out-of-vocabulary words (abbreviations, acronyms, and slang). Portuguese texts from the technology domain still use technical terms in English. All these factors reduce the efficiency of automatic classifiers. To minimize this problem, it was necessary to normalize the tweets. So, in the texts processing stage, we removed (or treated) special characters and hashtags, transformed emoticons and hyperlinks into text, and replaced abbreviations and slang with usual expressions, such as "*vc*" into "*você*" (you) and "*novis*" into "*novidade*" (news).

After this stage, we used the parser VISL[1] to supply the morphosyntactic annotation to the texts. Although this tool provides a richer annotation on linguistic information, in our study, to simplify, we used only lemmas and Part-of-Speech (PoS) tags of the words.

In the following sections, we describe studies on subjectivity classification.

5 Lexicon-Based Subjectivity Classification

The lexicon-based approach depends on finding the opinion lexicon that is used to analyze the text. This approach is divided into dictionary-based approach and corpus-based approach that use statistical or semantic methods to find sentiment polarity [10]. In our study, we used the dictionary-based approach. We did tests with

[1] http://beta.visl.sdu.dk/visl/pt/.

two lexicons: SentiLex-PT and WordNetAffect BR. SentiLex-PT [11] is a lexicon for Portuguese language that has 7,014 lemmas (4,779 adjectives, 1,081 nouns, 489 verbs, and 666 expressions). Each lemma shows the grammatical category, the target in a sentence; for each target, the polarity associate with it (positive, neutral, or negative), and the last information is about the method of assignment (if it was manual or with a tool named Judgment Analysis Lexicon Classifier - JALC). WordNetAffect BR [13] was built from WordNet terms translated from English into Portuguese, with terms that connote different emotions. The 289 words, including adjectives and nouns, were manually translated.

Initially, we did some tests with both lexicons separately, however, when we decided to use them together, we obtained a small improvement. Therefore, in our study, we are looking for the polarity of the words of tweets in both lexicons.

In the next section, we present the strategies we use to define the subjectivity of tweets.

5.1 Strategies to Identify Subjective Tweets

We tested three different heuristics to define the class of the tweets. All heuristics consider the polarity of the words (lemma) from the lexicon. If the word did not exist in the lexicon, its polarity was considered neutral.

- Heuristic 1 – The sum of polarities: This heuristic consists of adding the polarity of each tweet token [10]. If the result of this sum is non-zero, the tweet has some sentiment. The formula is represented in (1), where $|sent|$ is the number of tweet tokens and $term_i$ represents each token present in the sentence of tweet.

$$\text{subjectivity}_{\text{sent}} = \sum_{i=1}^{|\text{sent}|} \text{polarity}(\text{term}_\text{i}) \tag{1}$$

- Heuristic 2 – The number of words with polarity: This heuristic[2] assumes that if n tweet words have polarity, then the tweet is subjective. The equations for this heuristic are in (2) and (3), where $n \in [1;4]$.

$$subjectivity_{sent} = \sum_{i=1}^{|sent|} countPolarity(n, term_i) \tag{2}$$

$$countPolarity(n, term_i) = \begin{cases} 1, & \textit{if tweet has n terms with polarity} \\ 0, & \textit{otherwise} \end{cases} \tag{3}$$

- Heuristic 3 – The proportion of words with polarity: This heuristic is similar to the previous one, however, it considers the number of tokens with polarity in relation to the total of tweet tokens [8]. The equation for this heuristic is in (4). The decision about the subjectivity of the tweets was based on a threshold. We tested the threshold value in the range $[0.5; 0.35]$.

[2] This heuristic is a small variation of the strategy proposed in [9].

$$subjectivity_{sent} = \sum_{i=1}^{|sent|} \frac{countPolarity(term_i)}{|sent|} \quad (4)$$

5.2 Results

To evaluate the results we adopted the usual measures in information retrieval: precision, recall, F-measure, and accuracy. Table 3 shows the best results obtained with the studied strategies.

Table 3. Lexicon-based approach results.

Heuristic	Precision	Recall	F-measure	Accuracy
1	0.64	0.64	**0.64**	0.70
2 (n = 2)	0.57	0.63	0.62	**0.74**
3 (threshold >=0.25)	0.50	0.64	0.57	0.72

By employing the most basic strategy (heuristic 1), we obtained the best F-measure. However, the number of words with polarity (heuristic 2, for n = 2) achieved the best accuracy. It is important to mention that we used the polarity of nouns, verbs, and adjectives. We also checked if the use of only adjectives in the heuristic 2 would be enough, but the results were worse.

The lexicon-based approach had limitations. Sometimes the meaning of a word in the lexicon did not correspond to its meaning in the sentence. In our study, the high number of errors is also due to the number of advertisements in the corpus. The tweets with advertisements are objective, but if they have words that promote products and these words have polarities, the tweet was classified as subjective. This is the case of the tweet: "*VENDO NOTEBOOK MARCA DELL Excelente condição!*" (SELL NOTEBOOK BRAND DELL Excellent condition!).

6 Machine Learning-Based Subjectivity Classification

In this approach, we used the classification algorithms SMO and Naive Bayes, both from Weka[3] tool. We chose these algorithms for being frequently referenced in related works [3, 6, 7]. The feature selection is a fundamental stage of this approach. For this reason, we tested a new method at this stage: the CMFS [15, 16]. We chose this method because it is new, suitable for class imbalance, and showed good results in [16]. We compared this new method to the usual method based on frequency. These methods are described in the next section.

[3] http://www.cs.waikato.ac.nz/ml/weka/.

6.1 Methods for the Features Selection

Regardless the method used, initially we made a list of words for both tweet classes: objective and subjective. For each class, we used the features selection methods to define the relevance of the words. Based on this relevance, we ranked n most important words (for n ranging from 10 to 100). To generate the Bag-of-Words (BoW) we chose two strategies which we called "union" and "exclusion". The strategy "union" merges n most relevant words for each class. The strategy "exclusion" merges, but also to excludes the words common to both classes. We also tested two vector space model to the tweets: binary and based on polarity. In the latter, we replaced the binary values by the polarities of words. The polarity was defined from the lexicons used in Sect. 4. We used three methods to select the most relevant words:

- Absolute Frequency (*fa*): indicates the number of occurrences of a word w_k in a class c_j.
- Relative Frequency (*fr*): corresponds to the relation between the number of occurrences of a word and the total number of words of the class. The equation for this frequency is in (5).

$$fr(w_k, c_j) = \frac{fa(w_k, c_j)}{|W|} \tag{5}$$

- Comprehensive Measurement Feature Selection (CMFS): indicates the significance of a word w_k in one class c_j, against the occurrences of the same word in the corpus. According to Yang et al. [16], this significance can be reached by multiplying the probability that the word w_k occurs in the category $c_j, P(w_k|cj)$, and the probability that the word w_k belongs to the category c_j, when the word w_k occurs, $P(c_j|w_k)$. The equation for this frequency is in (6).

$$CMFS(w_k, c_j) = \frac{P(w_k|c_j)P(C_j|W_k)}{P(W_k)} \tag{6}$$

6.2 Results

As in Sect. 5.2, the usual measure were used to evaluate the results. Tables 4 and 5 show the best results obtained with the studied methods for the features selection applying SMO and Naive Bayes classification algorithms, respectively. The *BoW* column represents the number of attributes used, following the strategy of the most relevant words for each class.

When we used SMO, the best method for the feature selection was CMFS based on polarity. Combining this method with the strategy "exclusion", we had improvements in both F-measure and accuracy. Nevertheless, with the Naive Bayes approach, we had a good accuracy with the relative frequency method and the strategy "exclusion", but this combination had the worst results for F-measure.

Table 4. Approach results using SMO

Method	Strategy	BoW	Precision	Recall	F-measure	Accuracy
Fa	Union	133	75,30	77,00	74,90	77,04
Fr	Union	133	75,30	77,00	74,90	77,04
CMFS	Exclusion	125	76,80	78,00	75,30	78,03
CMFS + pol.	Exclusion	125	77,60	78,50	**75,70**	**78,51**

Table 5. Approach results using Naive Bayes

Method	Strategy	BoW	Precision	Recall	F-measure	Accuracy
Fa	Exclusion	21	71,80	73,20	63,70	73,20
Fr	Exclusion	21	71,80	73,20	63,70	**73,20**
CMFS	Union	171	71,40	71,30	**71,30**	71,26
CMFS + pol.	Union	16	70,20	71,80	70,80	71,77

7 Conclusion

This paper describes two different approaches to subjective classification, one uses machine learning and the other lexicon based. We propose different heuristics to highlight the differences between them and compare the results. The evaluation of the results showed that machine learning obtain better results than lexicon based. For the lexicon based approach the best result was 74 % of accuracy, while in machine learning approach the best result was 78 %. For future studies we want to classify the subjective tweets into positive or negative polarities.

Acknowledgments. Our thanks to Dell for the financial support of this work.

References

1. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013)
2. Dale, R., Moisl, H., Somers, H. (eds.): Handbook of Natural Language Processing. CRC Press, Boca Raton (2000)
3. Kamal, A.: Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources (2013). arXiv preprint arXiv:1312.6962
4. Fersini, E., Messina, E., Pozzi, F.A.: Subjectivity, polarity and irony detection: a multi-layer approach. In: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA (2014)
5. Drury, B., de Andrade Lopes, A.: A comparison of the effect of feature selection and balancing strategies upon the sentiment classification of Portuguese news stories. In: Proceedings of ENIAC (2014)
6. Santos, A.P., Ramos, C., Marques, N.C.: Sentiment classification of Portuguese news headlines. Int. J. Softw. Eng. Appl. **9**(9), 9–18 (2015)

7. Rosa, R.L., Rodríguez, D.Z., Bressan, G.: SentiMeter-Br: a social web analysis tool to discover consumers' sentiment. In: 2013 IEEE 14th International Conference on Mobile Data Management (MDM), vol. 2, pp. 122–124. IEEE (2013)
8. Morgado, I.C.: Classification of sentiment polarity of Portuguese on-line news. In: Proceedings of the 7th Doctoral Symposium in Informatics Engineering, pp. 139–150 (2012)
9. Filho, P.P.B., Pardo, T.A., Aluısio, S.M.: An evaluation of the Brazilian Portuguese liwc dictionary for sentiment analysis. In: 9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, Ceara (2013)
10. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
11. Carvalho, P., Silva, M.J.: SentiLex-PT: principais características e potencialidades. Oslo Stud. Lang. **7**(1), 425–438 (2015)
12. Pasqualotti, P.R., Vieira, R.: WordnetAffectBR: uma base lexical de palavras de emoções para a língua Portuguesa. RENOTE **6**, 1–10 (2008)
13. Généreux, M., Martinez, W.: Contrasting objective and subjective Portuguese texts from heterogeneous sources. In: Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, pp. 46–51. Association for Computational Linguistics (2012)
14. Moraes, S., Silveira, M., Manssour, I.: 7x1-PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa. BRACIS, STIL (2015)
15. Yang, J., Liu, Y., Zhu, X., Liu, Z., Zhang, X.: A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Inf. Process. Manage. **48**(4), 741–754 (2012)
16. Yang, J., Qu, Z., Liu, Z.: Improved feature-selection method considering the imbalance problem in text categorization. Sci. World J. (2014)
17. Souza, M., Vieira, R.: Sentiment analysis on twitter data for Portuguese language. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS, vol. 7243, pp. 241–247. Springer, Heidelberg (2012)
18. Lambov, D., Dias, G., Noncheva, V.: High-level features for learning subjective language across domains. In: Proceedings of International AAAI Conference on Weblogs and Social Media ICWSM (2009)