Exploring Resources for Sentiment Analysis in Portuguese Language

Larissa A. de Freitas
PUCRS - PPGCC
Email: larissa.freitas@acad.pucrs.br

Renata Vieira
PUCRS - PPGCC
Email: renata.vieira@pucrs.br

***Abstract*—Sentiment Analysis is the field of study that analyzes people's opinions in texts. In the last decade, humans have come to share their opinions in social media on the Web (e.g., forum discussions and posts in social network sites). Opinions are important because whenever we need to take a decision, we want to know others' points of view. The interest of industry and academia in this field of study is partly due to its potential applications, such as: marketing, public relations and political campaign. Research in this field often considers English data, while data from other languages are less explored. In this work we evaluate the available resources to assist Portuguese language sentiment analysis. For doing this, we perform sentiment analysis in a data set of the accommodation sector. We compare different pos-taggers and sentiment lexicons. We also evaluate the impact of some linguistic rules regarding negation and the position of adjectives.**

I. Introduction

According to Liu [1] the availability of the large volumes of opinion data encouraged a new research area in computer science called sentiment analysis, which became increasingly prominent in the 2000s.

Due to numerous practical applications, both, industry and academy have interests in the area of sentiment analysis. Opinions are available on the Web in an unstructured or in a semi-structured way, so it is very difficult to automatically process it [1]. Moreover, it is time consuming and expensive. Therefore creating tools to automate tasks related to sentiment analysis becomes increasingly important.

Customers identify online reviews as having a significant influence on their purchase in various economic sectors (e.g., hotel 87%, travel 84%, restaurant 79%, legal 79%, automotive 78%, medical 76%, and home 73%)[1].

Research on feature-level sentiment analysis often deals with data in English, while data from other languages are less explored. In Brazil, research on this area is still in its beginning, but an effort is being made to create resources and techniques to be used in this task. For instance, the external resources needed for the process, such as sentiment lexicons, only started to be developed for Brazilian Portuguese in 2011.

On the other hand, Brazil is the fourth country with more Internet users in the world, according to BBC [3]. Moreover, according to Semiocast [2], Portuguese is the third most used language on Twitter, after English and Japanese.

Therefore, in this work we evaluate the resources to assist Portuguese language sentiment analysis. We run some experiments comparing pos-taggers and sentiment lexicons. We also study the impact of a few linguistic rules for the process of polarity assignment.

The remainder of this paper is organized as follows. Section II presents the tools and options evaluated in our work (pos-taggers, sentiment lexicons, and linguistic rules). Section III presents the method we used for polarity assignment. Section IV shows the evaluation. Section V presents the discussion. Section VI shows our final remarks.

II. Resources

A. Part-Of-Speech Tagger

Pos-taggers are systems that read texts and assign morphological classes to each word in a sentence, such as nouns, verbs, adjectives, adverbs, *etc.*

These systems have been used effectively in sentiment analysis [7]. According to Chesley *et al.* [8] documents with higher number of adjectives and adverbs or with higher amounts of first-person subject and object pronouns it is more likely to be subjective.

TreeTagger, FreeLing, and CitiusTagger are the morphosyntactic tagger systems for the Portuguese language explored in this work. Because, all of them use the same set of tags, based on recommendations by the Expert Advisory Group on Language Engineering Standard (EAGLES) [6], to label grammatical categories.

TreeTagger uses Decision Trees to classify grammatical categories, while CitiusTagger uses Naïve Bayes from bigrams and FreeLing uses the Hidden Markov Model from trigrams to accomplish the same task.

A previous comparison between TreeTagger and FreeLing is made by Gamallo and Garcia [4]. The authors stated that FreeLing has a higher performance level than TreeTagger. The results obtained in the experiments, training and testing corpora in European Portuguese (Miscelâneo corpus), were 98% and 96% precision.

In Gamallo *et al.* [5], the three systems were compared (TreeTagger, FreeLing, and CitiusTagger). The experiments were made in Spanish (Ancora corpus). FreeLing obtained the highest precision (96.85%), followed by CitiusTagger (96.45%) and TreeTagger (95.52%).

[1] http://www.comscore.com

978-1-5090-0016-6/15 $31.00 © 2015 IEEE
DOI 10.1109/BRACIS.2015.52

These three taggers have not been evaluated before extrinsically, here we present such an evaluation in the context of sentiment analysis.

B. Sentiment Lexicon

Sentiment lexicons are the main linguistic resources employed in the task of Sentiment Analysis, they serve as polarity dictionary which is consulted in the process of sentiment polarity assignment (positive, negative or neutral).

In Portuguese, as far as we know, there are just four sentiment lexicons: OpLexicon [9], Brazilian Portuguese LIWC Dictionary[2], SentiLex [10], and synsets with polarity of Onto.PT [11].

OpLexicon [9] has 30,322 words (23,433 adjectives and 6,889 verbs) and was built based on a Brazilian Portuguese corpus (composed of 346 movie reviews and 970 journalistic texts), a thesaurus TEP [12] and the translated Liu's English Opinion Lexicon [13]. The results of each of these techniques are combined to create a large lexicon for Brazilian Portuguese.

Brazilian Portuguese LIWC Dictionary was built from the original English LIWC Dictionary [14] and has 127,149 entries, where each entry can be assigned to one or more categories.

SentiLex [10] finds which adjectives can be used as human modifiers and then assigns them a polarity attribute. The resource is available in two files, one where the word entries are inflected and other where all entries are lemmatized. The first file covers 82,347 lemmas, of which 16,863 are adjectives, 1,280 are nouns, 29,504 are verbs and 34,700 are idiomatic expressions. The second file covers 7,014 lemmas (5,473 manual and 1,541 automatic; 4,596 negative, 1,548 positive and 860 neutral), of which 4,779 are adjectives, 1,081 are nouns, 489 are verbs, and 666 are idiomatic expressions.

Onto.PT [11] contains 10,318 synsets with assigned polarity and tries to cover the entire language and not just a specific domain. The resource was constructed in two steps. Initially, the polarity of Onto.PT synsets was assigned using SentiLex as the polarity reference. After, the polarity was propagated through semantic relations.

C. Linguistic Rules

Negation is a very common linguistic construction that affects polarity. In the literature, we found some surveys about negation in Sentiment Analysis, such as Shah and Rekh [15], Wiegand *et al.* [16], and the precursors of this area: Polanyi and Zaenen [17], Kennedy and Inkpen [18], and Wilson *et al.* [7].

The complex scope of the negation model in Sentiment Analysis depends on the language. In Brazilian Portuguese, there are at least three ways of verbal negation. They are: one standard preverbal form, in which the negative particle appears before the verb and two nonstandard forms, one post verbal, in which the particle appears after the verb, and one in which there is double negation. In this case the verb is surrounded by two negation particles, one before and one after the verb [20]. Note that, differently from logical rules, double negation in natural language does not make the sentence positive.

In the experiments conducted for this work, all three forms of negation found in Portuguese were considered. We chose to apply a simple rule that consists of inverting the polarity of the opinion word if the negation particle ('não' ['no'], 'nunca' ['never'], 'nada' ['nothing'], 'nem' ['neither'], 'nenhum' ['none'], 'ninguém' ['nobody']) appears in any of the three verbal negation forms.

In this work we are also considering some variants of linguistic rules, regarding the position of adjectives as described below.

In Portuguese, the position the adjective occupies within the noun phrase has a relevant role. Neves [19] affirms that, in Portuguese, adjectives that are to the right of the noun are in their usual position, which means that this is the less marked position. Classifier adjectives are mostly found in this position. A qualifier adjective, on the other hand, can also be found on the left side of the noun, and this position implies a more subjective interpretation. The author points out that this position is associated to a more specific and restrictive interpretation. The left-head position is more stressed, that is, it is not so usual, and that is why it might mean an emphasis and trigger some special meaning effects.

III. Polarity assignment

Our polarity assignment process consists of receive as input a set of reviews, which are preprocessed. After, features are identified in the preprocessed reviews. The polarity is identified in the preprocessed reviews containing features using sentiment lexicons and linguistic rules.

There are many different approaches in the literature regarding polarity assignment, however we are not discuss these question here, since our main goal is to provide an evaluation of currently available tools in performing this kind of task. For other approaches of polarity assignment, see Popescu and Etzioni [21], Gamon et al. [22], and Peñalver-Martínez et al. [23].

A. Preprocessing

The main objective of this step is to obtain the grammatical categories and lemmas using pos-taggers.

Pos-taggers classify words into grammatical categories, based on the role they play in the context in which they appear. Most tools make use of the same basic grammatical categories (noun, verb, adjective, adverb, *etc.*) [24]. Some systems contain a much more elaborate set of tags, such as Portuguese TreeTagger[3]. For instance, adjectives may be marked with the following classification: AQ0, AQA, AQC and AQS (Adjective; Qualifier; and their degree ['0' for default form; 'A' for Augmentative; 'C' for Diminutive; 'S' for Superlative]), nouns may be marked with the following classification: NCCP, NCCS and NCCI (Noun; Common; Common; and their number ['P' for Plural; 'S' for Singular; 'I' for Invariable]).

Figure 1 shows the sentence "Os quartos são bons e baratos." ["The rooms are good and cheap."] marked with

[2] http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

[3] http://gramatica.usc.es/˜gamallo/tagger.htm

Portuguese TreeTagger, where: DA0 is a determinant article, NCCP is a plural common noun, VMI is an indicative main verb, AQ0 is a qualifier adjective, CC is a coordinated conjunction, Fp is a punctuation.

Token	Os	quartos	são	bons	e	baratos	.
Category	DA0	NCCP	VMI	AQ0	CC	AQ0	Fp
Lemma	o	quarto	ser	bom	e	barato	.

Fig. 1. Example of Preprocessing.

B. Feature Identification

Polarity is usually assigned to a group of pre-defined features, that is, there are some aspects into investigation to which a system must apply a polarity assessment.

In the method, these features are identified in the preprocessed reviews. Pos-taggers are efficient for explicit feature extraction in terms of accuracy [13]. We use features of TripAdvisor in order to identify explicit features (e.g.; 'Os **quartos** são bons e baratos.' ["The **rooms** are good and cheap."], 'quartos' ['rooms'] is a TripAdvisor feature).

C. Polarity Identification

The next step is to identify a polarity to the features mentioned in the reviews. For that we use sentiment lexicons (adjectives and their polarities) and linguistic rules.

The linguist rules used are: (1) **baseline**, we identify the presence of adjectives three positions before feature and three positions after feature, apply the negation rules, and attribute as feature polarity the sum of polarities of all the adjectives found; (2) **adjectives position**, we identify the presence of adjectives immediately before feature, if this is not found, then we look for adjectives after feature (this operation is performed until another feature is found or until the end of the sentence), apply the negation rules, and attribute as feature polarity the polarity of just one adjective found.

IV. Evaluation

A. Data Set

We collected data from TripAdvisor through a Web crawler. TripAdvisor is the world's largest travel site, reaching more than 190 million reviews and opinions covering more than 4.4 million accommodations, restaurants and attractions. The data collection contains 194 Brazilian Portuguese reviews published from March, 2010 to May, 2014.

The manual annotation of those reviews was conducted by two annotators, both native speakers of Portuguese, one linguist and one computer scientist.

The agreement between annotators was measured with Kappa Statistics [25]. The Kappa Statistics is a metric that evaluates concordance level classification tasks. The annotators agreement about sentiment analysis of the 5 features from TripAdvisor using Kappa was 0.67, which is considered a substantial agreement (in a scale consisting of 'poor', 'fair', 'moderate', 'substantial', and 'almost perfect'). We believe that the annotation has an acceptable Kappa value.

It is also important to note that only in a few cases the annotators disagreed between negative and positive polarities, the majority of disagreements was about positive and neutral polarities, or negative and neutral polarities.

B. Evaluation Configurations

We present a comparison of F-measure scores, obtained for positive and negative polarities for three Portuguese postaggers, four sentiment lexicons and two linguistic rules.

The first three evaluation configurations refer to pos-tagger alternatives, as describe below:

- configuration #1: this configuration uses TreeTagger, a union of Portuguese sentiment lexicons, and baseline;
- configuration #2: this configuration uses FreeLing, a union of Portuguese sentiment lexicons, and baseline;
- configuration #3: this configuration uses CitiusTagger, a Union of Portuguese sentiment lexicons, and baseline.

Based on the best tagger resulting from this first evaluation, next we present a comparison of F-measure scores for the four different sentiment lexicons, all combined with TreeTagger and baseline. The next four configurations are describe below:

- configuration #4: TreeTagger, a OpLexicon (Brazilian Portuguese sentiment lexicon), and baseline;
- configuration #5: TreeTagger, a SentiLex (European Portuguese sentiment lexicon), and baseline;
- configuration #6: TreeTagger, a LIWC-PT (Brazilian Portuguese sentiment lexicon), and baseline;
- configuration #7: TreeTagger, a synsets with polarities from Onto.PT (European Portuguese sentiment lexicon), and baseline.

Finally we present a comparison of F-measure scores for the four different sentiment lexicons, now combined with TreeTagger and adding rules for adjectives position. The last four configurations are describe below:

- configuration #8: TreeTagger, a OpLexicon, and adjectives position;
- configuration #9: TreeTagger, a SentiLex, and adjectives position;
- configuration #10: TreeTagger, a LIWC-PT, and adjectives position;
- configuration #11: TreeTagger, a synsets with polarities from Onto.PT, and adjectives position.

C. Results

We present a comparison of F-measure scores, obtained for positive and negative polarities of TripAdvisor's features, for three different pos-taggers (Table I).

Table I shows that using Portuguese TreeTagger we found the best positive F-measure for the following features: 'Localização' ['Location'] and 'Custo-benefício' ['Value'].

And, when using FreeLing, we found the best negative F-measure for the following features: ‘Quarto’ [‘Rooms’] and ‘Atendimento’ [’Service’]. For ‘Limpeza’ [‘Cleanliness’] and ‘Custo-benefício’ [‘Value’] the negative F-measure has the same value for all pos-taggers. Similarly, for the feature ‘Atendimento’ [’Service’] the positive F-measure has the same value for all tried pos-taggers.

As shown in Table I we obtained 0.539 as the best average between positive and negative F-measure, this result was obtained for configuration #1.

Based on results presented in Table I, we selected the Portuguese TreeTagger as the pos-tagger to be applied in next analysis, presented in Tables II and III.

Table II shows that using synsets with polarities from Onto.PT and baseline rule we found the best positive and negative F-measures for most features.

As shown in Table II we obtained 0.626 as the best average between positive and negative F-measure, this result was obtained for configuration #7.

Table III shows that using synsets with polarities from Onto.PT and adjectives position rule we found the best negative F-measure for most features, except for the ‘Localização’ [‘Location’] feature.

As shown in Table III we obtained 0.632 as the best average between positive and negative F-measure, this result was obtained for configuration #11.

If we compare Table II and III, we note that the best averages between positive and negative F-measure were 0.532, 0.346 and 0.632 using configurations #8, #10 and #11. In these configurations, we use the adjectives position rule and Brazilian or European Portuguese sentiment lexicon.

As we can see in Tables I, II and III, in general the F-measures for sentiment orientation recognition were better for positive than for negative cases. This may be explained because the reviews in the website were mostly marked as positive, against a low number of negative reviews.

TABLE I. POLARITY RECOGNITION OF TRIPADVISOR FEATURES USING DIFFERENT PORTUGUESE POS-TAGGERS.

Features	#1		#2		#3	
	Pos	Neg	Pos	Neg	Pos	Neg
Rooms	0.68	0.29	**0.69**	**0.31**	0.67	0.24
Location	**0.79**	0.17	0.71	0.33	0.71	0.33
Service	0.90	0.36	0.90	**0.43**	0.90	0.31
Cleanliness	0.80	0.40	0.80	0.40	0.67	0.40
Value	**1.00**	0.00	0.67	0.00	0.80	0.00
Avg.	0.834	0.244	0.754	0.294	0.750	0.256
	0.539		0.524		0.503	

V. DISCUSSION

In this work we evaluated the impact of three different pos-taggers (TreeTagger, FreeLing, and CitiusTagger) on the feature-level sentiment analysis method. According to our experiments (considering our accommodation reviews data set), in general the best results were obtained when using TreeTagger.

TABLE II. POLARITY RECOGNITION OF TRIPADVISOR FEATURES USING DIFFERENT PORTUGUESE SENTIMENT LEXICONS WITH BASELINE RULE.

Features	#4		#5		#6		#7	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Rooms	0.69	0.29	0.57	0.44	0.46	0.18	**0.70**	**0.61**
Location	**0.79**	0.17	0.77	0.31	0.50	0.00	0.77	0.31
Service	0.90	0.36	0.90	0.36	0.76	0.20	0.90	**0.50**
Cleanliness	0.67	0.40	0.80	0.67	0.80	0.40	0.80	0.67
Value	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Avg.	0.810	0.244	0.608	0.356	0.504	0.156	0.834	0.418
	0.527		0.482		0.330		**0.626**	

TABLE III. POLARITY RECOGNITION OF TRIPADVISOR FEATURES USING DIFFERENT PORTUGUESE SENTIMENT LEXICONS WITH ADJECTIVE POSITION RULE.

Features	#8		#9		#10		#11	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Rooms	**0.69**	0.24	0.46	0.40	0.41	0.12	0.66	**0.67**
Location	**0.79**	0.29	0.75	**0.37**	0.49	0.17	0.78	0.35
Service	0.84	0.50	0.84	0.36	0.71	0.36	0.84	**0.62**
Cleanliness	0.57	0.40	0.67	0.67	**0.80**	0.40	0.67	0.67
Value	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Avg.	0.778	0.286	0.544	0.360	0.482	0.210	0.790	0.474
	0.532		0.456		0.346		**0.632**	

Besides that, we also evaluated four different Portuguese sentiment lexicons: OpLexicon, SentiLex, LIWC-PT, and synsets with polarities from Onto.PT. From the experiments realized the best results, for our data set and our method, were obtained using synsets with polarities from Onto.PT.

We also tried different linguist rules, considering adjectives and negation words (baseline and adjectives position). In our experiments, the adjective position rule produced the best F-measure results for negative polarity identification for most features.

VI. FINAL REMARKS

Finally, this work is the first approach that evaluate many types of configurations, vary pos-taggers, sentiment lexicons and linguistic rules. Although, we identify the configuration #11 as the best configuration for our data set and our method, all resources (TreeTagger, FreeLing, CitiusTagger, OpLexicon, SentiLex, LIWC-PT, synsets with polarities Onto.PT, baseline and adjectives position) proved suitable.

REFERENCES

[1] Liu, B. “Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)”. California: Morgan & Claypool Publishers, 2012, 180p.

[2] Semiocast. “Half of Messages on Twitter are not in English Japanese is the Second Most Used Language”. Available from: https://semiocast.com/downloads/Semiocast_Half_of_messages_on _Twitter_are_not_in_ English_20100224.pdf, February 2010.

[3] BBC. “Brasil deve fechar 2014 como quarto país com mais acesso à internet, diz consultoria”. Available from: http://www.bbc.co.uk/portuguese/noticias/2014/11/141124_brasil _internet_pai, November 2014.

[4] Gamallo, P. and Garcia, M. “FreeLing e TreeTagger: um estudo comparativo no âmbito do Português”. Technical Report, ProLNat, 2013, 5p.

[5] Gamallo, P. and Pixel, J. C. and Garcia, M. and Abuín, J. M. and Fernández-Pena, T. "PoS tagging and Named Entity Recognition in a Big Data environment". *Procesamiento del Lenguaje Natural*, vol. 53, 2014, pp. 17–24.

[6] Leach, G. and Wilson, A. "Recommendations for the Morphosyntactic Annotation of Corpora". Technical Report, Expert Advisory Group on Language Engineering Standard (EAGLES), 1996, 28p.

[7] Wilson, T. and Wiebe, J. and Hoffmann, P. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis". In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347–354.

[8] Chesley, P.. and Vicenti, B. and Xu, L. and Srihari, R. "Using Verbs and Adjectives to Automatically Classify Blog Sentiment". In: Spring Symposium on Computational Approaches to Analyzing Weblogs, 2005, pp. 27–29.

[9] Souza, M. and Vieira, R. and Chishman, R. and Alves, I. M. "Construction of a Portuguese Opinion Lexicon from Multiple Resources". In: 8th Brazilian Symposium in Information and Human Language Technology, 2011, pp. 59–66.

[10] Silva, M. J. and Carvalho, P. and Sarmento, L. "Building a Sentiment Lexicon for Social Judgement Mining". In: 10th International Conference Computational Processing of the Portuguese Language, 2012, pp. 218–228.

[11] Oliveira, H. G. and Santos, A. P. and Gomes, P. "Assigning Polarity Automatically to the Synsets of Wordnet-Like Resource". In: 3rd Symposium on Languages, Applications and Technologies, 2014, pp. 169–184.

[12] Dias, B. C. and Moraes, H. R. "A Construção de um Thesaurus Eletrônico para o Português do Brasil". *Alfa*, vol. 47, 2003, pp. 101–115.

[13] Hu, M. and Liu, B. "Mining Opinion Features in Customer Reviews". In: 19th National Conference on Artificial Intelligence, 2004, pp. 755–760.

[14] Pennerbaker, J. W. and Francis, M. E. and Booth, R. J. "Linguistic Inquiry and Word Count". Mahwah, NJ: Erlbaum Publishers, 2001, pp. 1–13.

[15] Shah, V. and Rekh, P. "A Survey: Importance of Negation in Sentiment Analysis". *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, 2014, pp. 70–73.

[16] Wiegand, M. and Balahur, A. and Roth, B. and Klakow, D. and Montoyo, A. "A Survey on the Role of Negation in Sentiment Analysis". In: Workshop on Negation and Speculation in Natural Language Processing, 2010, pp. 60–68.

[17] Polanyi, L. and Zaenen, A. "Contextual Valence Shifters". In: Computing Attitude and Affect in Text: Theory and Applications, 2006, pp. 1–10.

[18] Kennedy, A. and Inkpen, D. "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters". *Computational Intelligence*, vol. 22, 2006, pp. 110–125.

[19] Neves, M. H. "Gramática de usos do português". São Paulo: Editora Unesp, 2011, 1005p.

[20] Schwenter S. A. "The Pragmatics of Negation in Brazilian Portuguese". *Lingua*, vol. 115, 2005, pp. 1427–1455.

[21] Popescu, A. M. and Etzioni, O. "Extracting Product Features and Opinions from Reviews". In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 9–28.

[22] Gamon, M. and Aue, A. and Corston-Oliver, S. and Ringger, E. "Pulse: Mining Customer Opinions from Free Text". In: 6th International Symposium on Intelligent Data Analysis, 2005, pp. 121–132.

[23] Peñalver-Martínez, I. and Valencia-García, R. and García-Sánchez, F. and Rodríguez-García, M. and Moreno, V. and Fraga, A. and Sánchez-Cervantes, J. "Feature-Based Opinion Mining Through Ontologies". *Expert Systems with Applications*, vol. 41, 2014, pp. 5995–6008.

[24] Feldman, R. and Sange, J. "The Text Mining Handbook, Advanced Approach in Analyzing Unstructured Data". New York: Cambridge University Press, 2007, 424p.

[25] Landis, J. and Koch, G. "The Measurement of Observer Agreement for Categorical Data". *Biometrics*, vol. 33, 1977, pp. 159–174.