information

MDPI

Article

Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese

Luana Balador Belisário, Luiz Gabriel Ferreira and Thiago Alexandre Salgueiro Pardo *

Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematics and Computer Science, University of São Paulo, São Carlos/SP 13566-590, Brazil; luana.belisario@usp.br (L.B.B.); luizgferreira@usp.br (L.G.F.)

* Correspondence: taspardo@icmc.usp.br

Received: 31 July 2020; Accepted: 8 September 2020; Published: 11 September 2020

Abstract: Texts published on social media have been a valuable source of information for companies and users, as the analysis of this data helps improving/selecting products and services of interest. Due to the huge amount of data, techniques for automatically analyzing user opinions are necessary. The research field that investigates these techniques is called sentiment analysis. This paper focuses specifically on the task of subjectivity classification, which aims to predict whether a text passage conveys an opinion. We report the study and comparison of machine learning methods of different paradigms to perform subjectivity classification of book review sentences in Portuguese, which have shown to be a challenging domain in the area. Specifically, we explore richer features for the task, using several lexical, centrality-based and discourse features. We show the contributions of the different feature sets and evidence that the combination of lexical, centrality-based and discourse features produce better results than any of the feature sets individually. Additionally, by analyzing the achieved results and the acquired knowledge by some symbolic machine learning methods, we show that some discourse relations may clearly signal subjectivity. Our corpus annotation also reveals some distinctive discourse structuring patterns for sentence subjectivity.

Keywords: subjectivity classification; feature sets; discourse structure; Portuguese language

1. Introduction

Social networks and the web environment, in general, have become a valuable source of information. By mining such data, companies and users may improve and/or select products and services of interest. Therefore, techniques for automatically analyzing user opinions (also referenced by "user generated data") have been extensively investigated. The research field that investigates these techniques is called sentiment analysis, also known as opinion mining.

The subjectivity analysis is one of the first steps in opinion mining. In this task, the documents of interest, which may be complete texts, sentences or even shorter text passages, are classified as subjective or objective [1]: when classified as objective, they express facts (for example, "I bought a Philco netbook in Nov/2010"); otherwise, when said subjectively, they express opinions or sentiments ("This book is very good, it is incredibly deep!"). In contrast to the objective sentence, it is possible to notice in the subjective example that the author of the review liked the book, as s/he used expressions like "very good" and "incredibly". These words that evidently denote opinion and polarity are called sentiment words. Table 1 shows some other labeled sentences in Portuguese (and possible translations to English) from the computer-BR corpus [2]. One may see that the subjective sentences can be further divided into "positive" and "negative" polarities. The objective sentences, despite having adjectives that could indicate opinions, like *boas* ("good", in English) and *baum* (misspelled form of *bom*—"good"), do not express opinions about the products. Besides the orthographic errors, it is worth noting that the examples include abbreviations and vocabulary typical of the language use in the web that must be taken into consideration when developing methods for sentiment analysis (for dealing with such issues, several approaches make use of some text normalization strategy, as the one proposed in [3] for the Portuguese language).

Table 1. Examples of labeled sentences in the computer-BR corpus [2].

Sentence	Polarity
Alguem me indica marcas boas de notebook? (Can anyone suggest me good notebook brands?)	Objective
Esse Not é baum? Alguem sabe? (Is this Note good? Anybody know?)	Objective
Sem notebook de novo . . . Parece brincadeira de mau gosto (No notebook again... It seems like a bad joke)	Subjective/Negative
Logo um precioso notebook Dell lindíssimo chega aqui em casa tô mt feliz (Soon a precious Dell notebook arrives here at home I'm very happy)	Subjective/Positive

The relevance of the subjectivity classification may be evidenced by its application in other tasks. For instance, in polarity classification, filtering out objective sentences is interesting (see, e.g., [4]); in opinion summarization, subjective sentences are much more important (see, e.g., [5]). Despite the relevance of the task, to the best of our knowledge, there is only one previous work specifically dedicated to subjectivity classification for the Portuguese language—the initiative of Moraes et al. [2]. In this paper, we focus our efforts in this task.

In our previous attempt [6], we started by reproducing the experiments of Moraes et al. and extended the evaluation of their methods to two other corpora, aiming to evaluate their robustness for other domains. We evaluated methods for subjectivity classification based on sentiment lexicons and machine learning (ML) techniques. We then explored other methods for the task: one that used word embeddings in a ML-based approach and another one based on graphs. We showed that the methods have varied performances for the different domains, but that some methods are more stable than others. We also showed that one of the new methods outperformed the previous results for Portuguese.

In this paper, we go on and focus on the machine learning approaches. We are mainly interested in exploring richer feature sets and evidencing their contribution to the task. Specifically, we bring the study of discourse analysis to the task, which is amongst the most complex and abstract linguistic knowledge levels in natural language processing. Following the widely known rhetorical structure theory (RST) [7], we additionally investigate the discourse structuring of subjective and objective sentences, looking for patterns of relations that are more frequent in user opinions.

We focus our study on a text domain that has shown to be a very challenging one in the area—book reviews. In [6], such reviews showed very irregular results across different techniques and, as pointed out in [8], book reviews show more writing style variance, including both formal and highly informal and orally-marked descriptions (depending on the characteristics of the reader and the book intended audience) and include more comments about general aspects of the book (differently from what happens in other domains, as electronic products, where users tend to comment about technical aspects). We could also notice that users are more engaged, showing more passionate behavior when producing their texts.

In the next section, we briefly present the relevant related work. In Section 3, we introduce the corpus used in this study. Section 4 details the investigated methods, while the results are reported in Section 5. Some final remarks are made in Section 6.

2. Related Work

2.1. Previous Attempts on Subjectivity Classification for Portuguese

Moraes et al. [2] are the only known authors to specifically investigate the theme of subjectivity classification for the Portuguese language. They created a corpus of tweets on the area of technology. This corpus—called computer-BR—was manually labeled and preprocessed to increase the efficiency of the applied methods. Inspired by the approaches for English, the tested methods were based on the use of sentiment lexicons and ML approaches. The best result with lexicon-based methods was 64% in f-measure, while ML-based methods reached 75%.

In a related task—polarity classification—Vilarinho and Ruiz [9] propose a classification method based on word graphs (named SentiElection) to predict if sentences show positive or negative polarities. They use a training set to create "positive" and "negative" graphs and then each test sentence is added to both graphs, being classified according to the graph that produces better centrality measurements. The best achieved classification result was 82% in f-measure in an airline dataset.

In our previous attempt to the task [6], we have extended the work of Moraes et al. We have included other corpora in the evaluation and also tested the above approach of Vilarinho and Ruiz. We improved the results and, in particular, achieved 83.2% of overall accuracy with a ML approach for book reviews, which have showed to be one of the most challenging domains for the methods of subjectivity classification, as we commented before.

For comparison purposes only, we briefly comment about some of the main results achieved for the English language, which have a longer research tradition in the area. Ref. [10], probably the best result so far, reports a performance of above 95% using sophisticated neural networks. Ref. [11] is one of the most relevant initiatives in the area. It achieved 91% accuracy and explored the contribution of similarity and Bayesian classification approaches. Ref. [12] is also worthy citing, as the authors are amongst the first ones to deal with the subjectivity classification task. They have developed a reference dataset and achieved over 81% accuracy with a Bayesian classifier. One may notice that the results for Portuguese are worse than the best ones achieved for English. However, to the best of our knowledge, none of the previous initiatives for any of the languages has investigated the issue of the impact of discourse in the task. We introduce the discourse model that we explore in what follows.

2.2. Rhetorical Structure Theory

In the discourse analysis task, Mann and Thompson [7] created the well-known rhetorical structure theory (RST), which is "a linguistically useful method for describing natural texts, characterizing their structure primarily in terms of relations that hold between parts of the text". Therefore, RST represents a text by a discourse tree, where the propositions corresponding to the parts of the text (referenced by the term "span") are leaves connected by discourse relations. Propositions, or discourse segments, must convey full ideas. The relations indicate how the segments are related to one another to form a coherent discourse. Figure 1 shows an example of a RST-annotated short text (using RSTTool annotation software [13]).

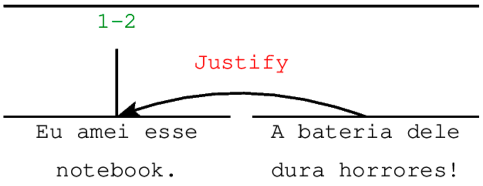


Figure 1. Rhetorical structure (RST) annotation for the short text *Eu amei esse notebook. A bateria dele dura horrores!* ("I loved this notebook. Its battery lasts a long time!").

In this analysis, one may see that the second segment "justify" why the author of the text says that s/he loved the notebook. The arrow leaves from the less important segment (called "satellite") and points to the most important one (the "nucleus") of the relation. The nuclear segment is also visually signaled by the vertical axis in the representation. The "justify" relation is a nucleus-satellite relation, but RST also includes multinuclear relations, without satellites. In these cases, the relation does not have a particular span which is more central in the related text (e.g., "list", "contrast" and "same-unit" relations, etc.).

Although not directly used for subjectivity classification, RST has already been used for sentiment analysis. Here, we highlight the work of Chenlo et al. [14], which investigated the use of RST for various sentiment analysis tasks, using textual and discourse characteristics of the sentences as features for the classification of sentence polarity. Our experiments in this paper were largely inspired by this work, using some of the proposed features, as we describe latter. We detail the investigated methods in Section 4, but, before, we introduce the corpus that we used.

3. The Corpus

In our previous work [6], we used three corpora to evaluate the explored methods. All of them were composed of material collected from the web. Two of them present sentences related to technology products, while the other presents sentences from book reviews. The corpus of book reviews have shown to be one of the most challenging ones for sentiment analysis tasks for Portuguese and, therefore, we have adopted it in this work in order to evaluate the potential of the methods and features that we test.

The book review corpus is composed by sentences taken from the ReLi corpus [15], Amazon website and the Skoob social network, with 270 sentences equally divided between subjective and objective sentences, which were also manually labeled according to their subjectivity. For this current work, we expanded it to 350 reviews proportionally divided between objective and subjective sentences. The subjective sentences are, in turn, proportionally divided into positive and negative sentences. On average, each sentence in the corpus has 22 words.

All the sentences were manually annotated according to RST, using the classical RSTTool annotation software [13]. We have used a traditional relation set composed of 37 relations, which are listed in Table 2.

Table 2. Relation set used in the manual annotation of the book reviews corpus.

Nucleus-Satellite Relations				Multinuclear Relations
antithesis	elaboration	motivation	summary	conjunction
attribution	enablement	non-volitional cause	unconditional	contrast
background	evaluation	non-volitional result	unless	disjunction
circumstance	evidence	otherwise	volitional cause	joint
comparison	explanation	parenthetical	volitional result	list
concession	interpretation	preparation		restatement
conclusion	justify	purpose		same-unit
condition	means	solutionhood		sequence

Table 3 synthesizes the main characteristics of the annotated corpus, as the number of sentences of each class and the average number of segments by sentence.

Table 3. Number of sentences and average number of segments per class (subjective/objective).

Characteristic	Class	Value
Number of sentences	Objective	175
	Subjective	175
Average number of segments per sentence	Objective	2.91
	Subjective	2.83
	Both	2.87

Table 4 shows the frequency of each relation for both objective and subjective sentences. By these numbers, one may already see that the types of relations that mostly occur with each type of sentence are different.

Table 4. Discourse relations present in the sentences per class, presented in decreasing order of the number of sentences in which they were found, with this information in parentheses.

Class	Relations
Objective	elaboration (105), list (74), same-unit (58), circumstance (37), sequence (34), non-volitional cause (17), purpose (13), justify (13), contrast (10), parenthetical (10), joint (8), non-volitional result (7), explanation (7), concession (7), attribution (4), means (4), conclusion (2), preparation (2), motivation (1), interpretation (1), antithesis (1), evidence (1), volitional cause (1), condition (1), summary (1)
Subjective	list (65), elaboration (59), justify (51), concession (35), same-unit (27), contrast (26), attribution (16), circumstance (14), evaluation (12), non-volitional result (9), joint (8), sequence (7), means (7), non-volitional cause (6), explanation (5), purpose (4), comparison (4), otherwise (4), conclusion (3), condition (3), preparation (3), parenthetical (3), antithesis (2), restatement (2), motivation (1), interpretation (1), evidence (1), background (1), summary (1)

Besides the more usual relations that happen for both sentence types (as elaboration and list), the most frequent relations in objective sentences are the ones aligned with descriptive, narrative, time and space content, as sequence, circumstance and cause relations. In the subjective sentences, on the other hand, we find more criticism, argumentation and personal view regarding characteristics of the story, expressed via relations as justify, concession and contrast.

We have also looked for patterns of co-occurrence of relations for each class in the annotated data. We could find that, for the objective sentences, the most common pairs of relations that happen together in the annotated sentences are elaboration and same-unit (with a proportion of 11.1% in relation to all possible pairs of relations), elaboration and circumstance (11.1%), elaboration and list (5.9%), and circumstance and same-unit (5.2%). For the subjective sentences, the pairs justify and elaboration (7.4%) and justify and concession (7.4%) were the most common ones. We have also checked the occurrences of frequent groups of 3 relations, but no distinctive pattern could be found.

The above results, although obtained for a single annotated corpus for a specific domain, are already an interesting contribution, as future initiatives may look for similar discourse patterns in other datasets.

4. The Methods

4.1. Lexicon-Based Method

The first method we tested was the traditional lexicon-based one, following the best configuration proposed by Moraes et al. [2]. It uses a heuristic that relies on a sentiment lexicon with pre-classified words. The words are associated with a value 1 if they are positive, −1 if negative and 0 if neutral. The subjectivity of a sentence is computed by simply summing the polarities of the words that compose it. If the result assumes a value different from 0, the sentence is considered “subjective”; if the value is zero, the sentence is considered “objective”. Due to the simplicity of the method, there is no treatment of negation, irony and adverbs, whose functions would be to intensify, neutralize or even change the orientation of the sentiment words. More than this, one may realize that this method is quite naive as it would classify a sentence with equal number of positive and negative words as “objective”, as the positive values would cancel the negative ones, resulting in a 0 value.

Additionally, following Moraes et al., we have used Sentilex-PT [16] and WordnetAffectBR [17] sentiment lexicons to test the method.

4.2. Graph-Based Method

We also tested the best configuration of the graph-based classification method of Vilarinho and Ruiz [9], which evaluates each sentence of interest in relation to reference objective and subjective graphs, adopting the class that produces the best centrality measurement in the graphs.

The objective and subjective graphs are built from a training corpus in a way that the nodes are the words and the links represent the sequence they appear. For example, considering a *word frame* (window) of size 3, the phrase "I love pizza" would produce (i) nodes labeled as "I", "love" and "pizza" and (ii) directed edges from "I" to "love", from "I" to "pizza" and from "love" to "pizza". Following this strategy, we built the objective graph from the objective sentences and the subjective graph from the subjective sentences in the training set. To classify a new sentence (in the test set), the method incorporates the sentence in each of the graphs and computes three global centrality measures—Eigenvector Centrality, Katz Index, and PageRank. For each measure, the graph with the highest value scores 1 point. In the end of this process, the sentence is classified according to the graph that scored best.

4.3. Machine Learning-Based Methods

Three ML-based methods were evaluated. Two are replications of the proposals of Moraes et al. [2] and the other is a different proposal using word embeddings.

Following Moraes et al., two classification techniques (Naive-Bayes and SVM) were tested, using the traditional bag of words representation, which considers the words in a sentence as distinct features. To perform feature selection, we quantify the relevance of the words in each class to select those that will form the bag of words. Two metrics were used for this: the first is simply the frequency of the word in the class; the second metric is the Comprehensive Measurement Feature Selection (CMFS) proposed in [18], which aims to calculate the relevance of the words in each class considering their occurrences in other classes.

We used the *scikit-learn* package to run the methods. For Naive-Bayes, the "ComplementNB" implementation was used with the standard library hyper-parameters and the 99 most relevant words (excluding those that were in both objective and subjective sets—which is the "exclusion" configuration proposed by Moraes et al.). For SVM, the "SVC" implementation was used and gamma was changed from standard to "auto", with penalty parameter C set to 30, with 60 words from each class (also using the "exclusion" configuration). These configurations in both methods were the ones that produced the best results for each case.

In our ML variant method, we used word embeddings, i.e., the vector representation of words learned from their contexts of occurrence. As described in [19], with the vector representation, it is possible to obtain values of semantic similarity between terms. Thus, we used the word embeddings to assist in two fronts of the task: (i) to give representativeness to subjective terms that were not present in the classification models (as consequence of not being present in the training data or, otherwise, simply being filtered out by the feature selection techniques of Moraes et al.) and (ii) to obtain a way to represent the whole sentence for classification, without the limitations of the bag of words representation.

The word embeddings were trained with the use of the well-known *gensim* library, with approximately 86-thousand sentences taken from the Buscapé corpus [20]. We used a dimension of 600, a window size of 4 words and only words that had more than 4 occurrences. The used model was the Continuous Bag of Words (CBOW), trained with 100 epochs. Such a configuration was the one that produced the best results for this approach. To represent the whole sentence, we explored methods that combine the vectors of the words of a sentence and generate a final vector that can represent its semantics. The best result was produced by the usual strategy of simply adding the word vectors.

The vector of each sentence was used to train a multi-layered neural network with input layer of size 600, one hidden layer (with 100 neurons), and an output layer with 2 neurons. The network receives the sentence vector and returns the class ("objective" or "subjective"). The tested activation functions were Rectified Linear Unit (ReLU) and Softmax for hidden and output layers, respectively. We used 20 epochs. The neural network was implemented using the *keras* library.

4.4. Enriched Machine Learning-Based Methods

The idea here was to define and test richer features with machine learning methods. In order to test the contribution of the features, we grouped them into three categories: lexical, graph centrality-based and discourse features.

Lexical features include grammatical characteristics of the sentence. We used the proportion of negation and intensity adverbs present in the sentence; the presence of exclamation ('!', '?!') and interrogation ('?') points (as Boolean features—1 for presence and 0 for absence); and the proportion of subjective words in nuclei and satellites of the RST trees.

Graph centrality-based features are the results of the three global centrality measures—Eigenvector Centrality, Katz Index, and PageRank—used in the graph-based method (Section 4.2). To compute the measures for our dataset, we used 120 validation sentences (out of the 350 sentences), being 60 subjective (30 positive and 30 negative) and 60 objective sentences.

Following the ideas in [14], most of the discourse features include Boolean indications of the presence or absence of the different relations of the RST model. If a certain relation is present in the proposition, the value of that feature is 1, otherwise it is set to 0. Another discourse feature is computed with respect to the RST tree structure: the distance between a leaf node and the tree root (similar to the height of the tree).

Here, to run the tests, we used the WEKA data mining system [21], trying several machine learning techniques of different paradigms in order to evaluate their performances and additional knowledge they might offer to the task.

We report the main achieved results and obtained conclusions in what follows.

5. Results and Discussion

To evaluate the effectiveness of the methods, tests were performed on the sentences of our book review corpus. The ML and graph-based methods were evaluated with 3-fold cross-validation (in order to get close to the 70/30 training-testing corpus division used for the graph-based method). For the lexicon-based method, as a training set was not necessary, the test was performed on the full corpora (but taking the average of the results for the corresponding three folds, in order to have a fair comparison of results).

We report our tests and results in what follows. We divided the tests according to the contributions that we envision for each situation. To start, we reproduce the best results of Belisário et al. [6] in Section 5.1.

5.1. The Methods of Belisário et al. (2020)

Table 5 synthesizes the best results achieved by Belisário et al. [6] for the book review corpus. We show the traditional precision, recall and f-measure values for each class, as well as the overall accuracy results.

Table 5. Results obtained for the corpus of book reviews in Belisário et al. [6].

Measures	**Lexicon-Based Methods**		**Graph-Based Method**	**Machine-Learning-Based Methods**		
	Sentilex-PT	**WordnetAffectBR**		**NB**	**SVM**	**Neural Network**
Precision (objective)	0.490	0.518	0.545	0.759	0.782	0.806
Recall (objective)	0.600	0.931	0.723	0.736	0.83	0.863
F-measure (objective)	0.539	0.665	0.652	0.747	0.805	0.831
Precision (subjective)	0.524	0.730	0.674	0.763	0.831	0.865
Recall (subjective)	0.413	0.181	0.536	0.782	0.783	0.804
F-measure (subjective)	0.461	0.288	0.596	0.772	0.806	0.832
Overall accuracy	0.504	0.545	0.627	0.761	0.806	**0.832**

One may notice that machine learning (neural networks, in particular) produced the best results, achieving an overall accuracy of 83.2% (indicated in bold in the table). The lexicon-based methods were the worst ones. This previous study, however, did not look in depth to the used feature sets. In fact, the results produced by the neural network were simply based on the word embeddings, as we commented before.

A detailed error analysis showed several interesting linguistic issues. For the machine learning methods, many errors happened due to the lack of information about some words. This fact was due to the informality of the language, with low frequency terms and several term variants with similar meanings. Regarding the lexicon-based method, as it is based on searching and counting sentiment words, it is not possible to effectively deal with figurative language (such as sarcasm and irony), adverbs of negation and intensity, disambiguation of words and the occurrence of implicit opinions in objective sentences. Some of these issues motivated the creation of some lexical features, which were cited before.

In this paper, we are more interested in understanding the potentiality of different feature sets used with varied machine learning techniques, which is the focus of this paper, whose results we start reporting in the next section.

5.2. Evaluating the Feature Sets

At first, we tested each feature category individually and then incrementally combined them, in order to evaluate the contribution of each category. For each test cycle, we present a table with the best achieved results for each machine learning paradigm (according to WEKA classification).

Tables 6–8 show the results for the individual use of lexical, centrality-based and discourse features, respectively. For the lexical features, one may see that SMO produced the best overall accuracy, achieving 67.6%. The centrality-based features significantly improved the results, achieving 75.1% of overall accuracy (with the Multilayer Perceptron), being 11% better than the results produced by the lexical features. For the discourse features, the results were intermediate, being better than the ones produced by lexical features, but worse than the ones of the centrality-based features.

Table 6. Results for the lexical features.

Category	Method	Class	Precision	Recall	F-Measure	Accuracy	Accuracy
Rules	OneR	Subjective	0.656	0.457	0.539	0.457	0.609
		Objective	0.583	0.760	0.660	0.760	
	PART	Subjective	0.658	0.726	0.690	0.726	0.674
		Objective	0.694	0.623	0.657	0.623	
Trees	J48	Subjective	0.658	0.726	0.690	0.726	0.674
		Objective	0.694	0.623	0.657	0.623	
Bayes	NaiveBayes	Subjective	0.669	0.680	0.674	0.680	0.671
		Objective	0.674	0.663	0.669	0.663	
Functions	SMO	Subjective	0.660	0.727	0.691	0.728	**0.676**
		Objective	0.694	0.623	0.656	0.623	
	MultilayerPerceptron	Subjective	0.651	0.640	0.646	0.640	0.649
		Objective	0.646	0.657	0.652	0.657	

Table 7. Results for the centrality-based features.

Category	Method	Class	Precision	Recall	F-Measure	Accuracy	Accuracy
Rules	OneR	Subjective	0.531	0.537	0.534	0.537	0.531
		Objective	0.532	0.526	0.529	0.526	
	PART	Subjective	0.631	0.566	0.596	0.566	0.617
		Objective	0.606	0.669	0.636	0.669	
Trees	RandomForest	Subjective	0.747	0.691	0.718	0.691	0.729
		Objective	0.713	0.766	0.738	0.766	
Bayes	NaiveBayes	Subjective	0.626	0.440	0.517	0.440	0.589
		Objective	0.568	0.737	0.642	0.737	
Functions	LibLINEAR	Subjective	0.735	0.697	0.716	0.697	0.723
		Objective	0.712	0.749	0.730	0.749	
	MultilayerPerceptron	Subjective	0.744	0.766	0.755	0.766	**0.751**
		Objective	0.759	0.737	0.748	0.737	

Table 8. Results for the discourse features.

Category	Method	Class	Precision	Recall	F-Measure	Accuracy	Accuracy
Rules	OneR	Subjective	0.794	0.286	0.420	0.286	0.806
		Objective	0.564	0.926	0.701	0.926	
	PART	Subjective	0.658	0.760	0.706	0.760	0.683
		Objective	0.716	0.606	0.656	0.606	
Trees	J48	Subjective	0.701	0.709	0.705	0.709	**0.703**
		Objective	0.705	0.697	0.701	0.697	
Bayes	NaiveBayes	Subjective	0.707	0.634	0.669	0.634	0.686
		Objective	0.668	0.737	0.701	0.737	
Functions	SMO	Subjective	0.723	0.566	0.635	0.566	0.674
		Objective	0.643	0.783	0.706	0.783	
	MultilayerPerceptron	Subjective	0.692	0.669	0.680	0.669	0.686
		Objective	0.680	0.703	0.691	0.703	

We also incrementally combined the feature sets. Joining lexical and centrality-based features produced better results than lexical features, but worse than the centrality-based features alone. The best result was produced by Multiplayer Perceptron, which achieved 73.1% overall accuracy. Combining all feature sets produced the best results, reaching 77.1% of overall accuracy (with LibLinear), demonstrating the relevance of using features of varied types in the process. Table 9 shows such results.

Table 9. Results for combined features.

Category	Method	Class	Precision	Recall	F-Measure	Accuracy	Accuracy
Rules	OneR	Subjective	0.531	0.537	0.534	0.537	0.531
		Objective	0.532	0.526	0.529	0.526	
	PART	Subjective	0.712	0.663	0.686	0.663	0.697
		Objective	0.684	0.731	0.707	0.731	
Trees	RandomForest	Subjective	0.797	0.720	0.757	0.720	0.769
		Objective	0.745	0.817	0.779	0.817	
Bayes	NaiveBayes	Subjective	0.755	0.669	0.709	0.669	0.726
		Objective	0.703	0.783	0.741	0.783	
Functions	LibLINEAR	Subjective	0.781	0.754	0.767	0.754	**0.771**
		Objective	0.762	0.789	0.775	0.789	
	MultilayerPerceptron	Subjective	0.744	0.714	0.729	0.714	0.734
		Objective	0.725	0.754	0.739	0.754	

It is widely known that using all the features is not necessarily the most intelligent strategy for classification; in fact, irrelevant features can decrease the performance of some algorithms. Therefore, in a new classification attempt, we have performed feature selection over all the features. We used the Best-First search algorithm and the CfsSubsetEval evaluator, which assesses the worth of a subset of features by considering the individual ability of each feature along with the degree of redundancy among them.

As a result of the feature selection process, 13 (thirteen) features were indicated as being more relevant:

1. From the lexical features: the proportion of adverbs of negation and intensity, and the presence of exclamation point in the sentence;
2. From the centrality-based features: the Eigenvector centrality of the subjective graph;
3. From the discourse features: the presence of some specific discourse relations (antithesis, cause, circumstance, background, comparison, contrast, condition, restatement and disjunction).

Other feature selection techniques produced similar results. Unfortunately, the results were worse than the ones produced by using all the features, reaching 73.7% of overall accuracy (for Naïve Bayes). Maybe the most interesting learned lesson in this feature selection effort comes from observing the discourse features that were indicated as more relevant to the task, which we will address again in the next section.

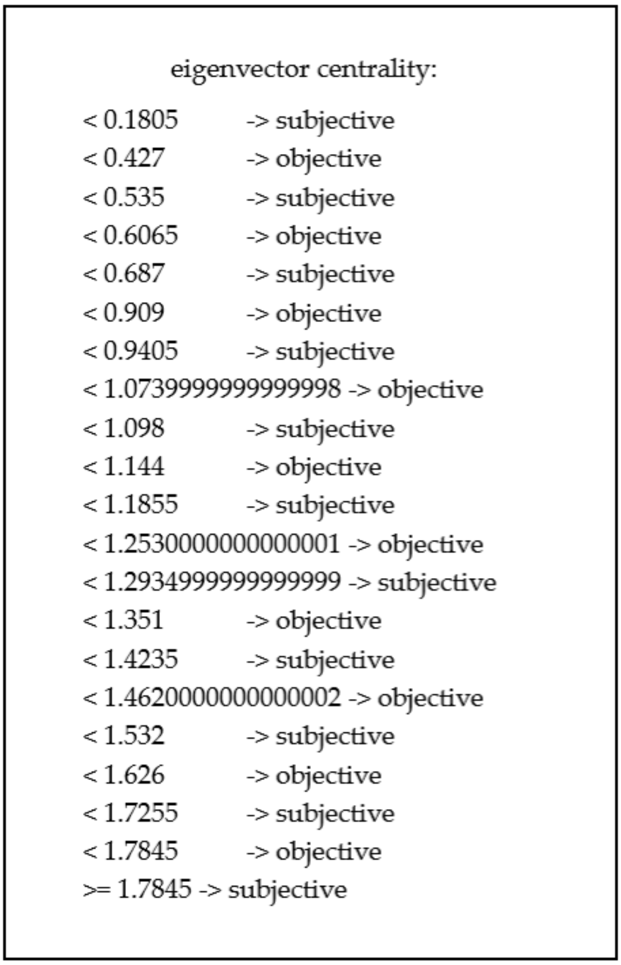
Overall, to the attentive reader, another interesting learned lesson comes from noticing that no single machine learning technique showed predominant distinctive behavior. Different techniques were useful for different feature set configurations.

5.3. Analysis of Acquired Knowledge

Finally, independently of the variation in the results achieved for each experiment, we look for relevant new knowledge that may emerge from the experiments. For this purpose, symbolic techniques are more appropriate and are the ones whose results we analyze here.

As specialized knowledge is relevant for assessing the meaning of some of the features (e.g., the discourse features demand knowing the RST model and what each discourse relation indicates), the corpus annotation expert has led this effort.

Figures 2 and 3 show the rules learned by the OneR technique and part of the decision tree built by J48, respectively. Both schemes were set up by running the machine learning techniques with all the features for the full dataset. The most relevant information we can extract from these two categories of methods is which features were considered as the most relevant to the task, so we have another way to evaluate the contribution of each feature or category of features.



```
eigenvector centrality:
< 0.1805           -> subjective
< 0.427            -> objective
< 0.535            -> subjective
< 0.6065           -> objective
< 0.687            -> subjective
< 0.909            -> objective
< 0.9405           -> subjective
< 1.0739999999999998 -> objective
< 1.098            -> subjective
< 1.144            -> objective
< 1.1855           -> subjective
< 1.2530000000000001 -> objective
< 1.2934999999999999 -> subjective
< 1.351            -> objective
< 1.4235           -> subjective
< 1.4620000000000002 -> objective
< 1.532            -> subjective
< 1.626            -> objective
< 1.7255           -> subjective
< 1.7845           -> objective
>= 1.7845 -> subjective
```

Figure 2. Rule learned by the OneR technique using all the features.

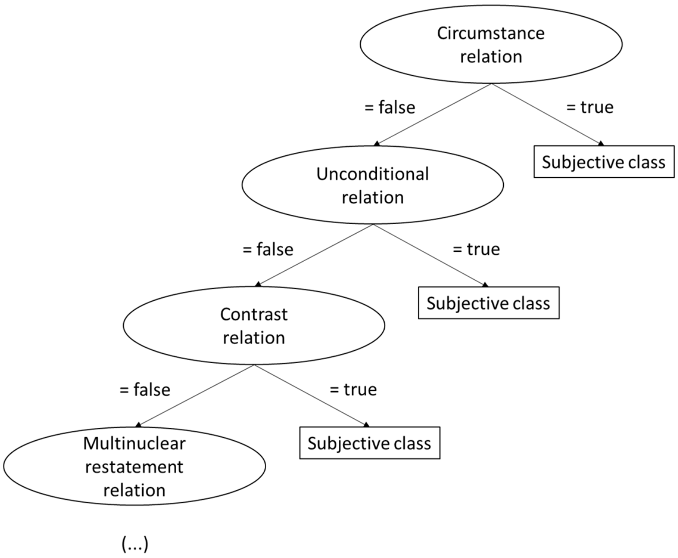
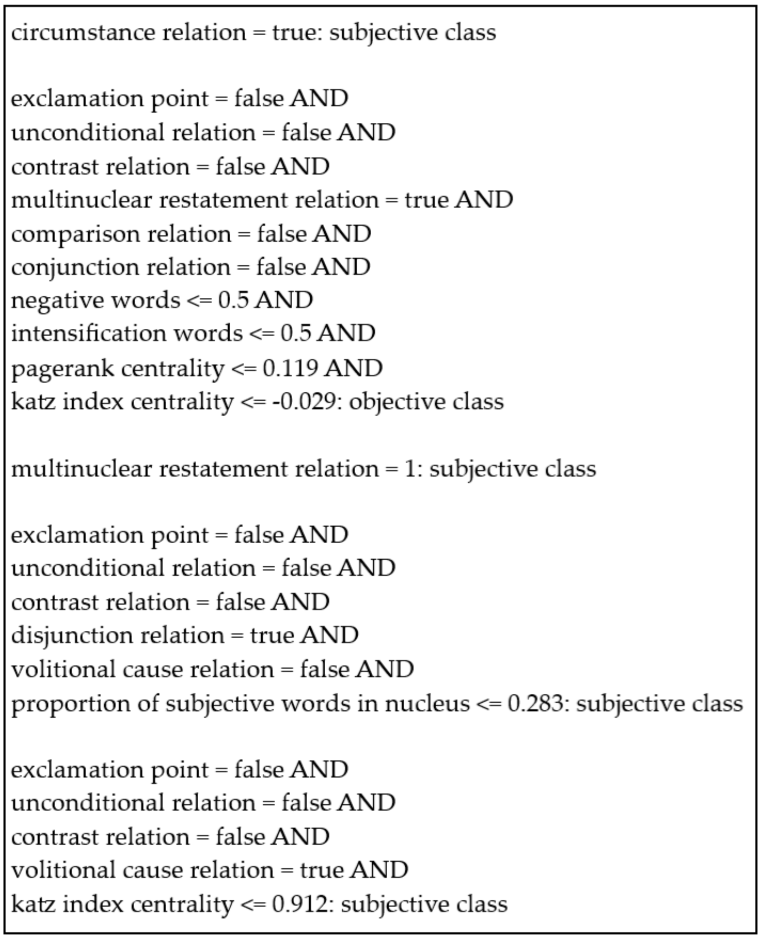


Figure 3. Part of the decision tree learned by J48 using all the features.

OneR indicates that the EigenVector feature is the most relevant one in the dataset (in fact, for the centrality-based features, the feature selection process in the previous section also indicated it as the most relevant one). However, it is not possible to establish a clear separation between value ranges and classes. For instance, one may see that if the EngenVector value is below 0.1805, the subjective class is assigned to the instance, but, if it is above this value but below 0.427, the objective class is indicated, and the classes keep changing places for each of the following value ranges.

Taking J48 tree as another information source, there are 13 levels and only 3 of them do not show discourse relations. This indicates how important such discourse information is (and, again, confirming the results of the feature selection process). The root of the tree already tests a discourse feature, indicating that it is the most important feature. As one may see, if a circumstance relation happens, the sentence is directly classified as belonging to the subjective class; otherwise, the unconditional relation is tested, and so on.

By analyzing the learned rules by the PART technique in Figure 4, it is possible to confirm again the relevance of the discourse relations. For instance, the first learned rule also checks the occurrence of the circumstance relation: if it happens, then the subjective class is assigned to the instance; the second rule tests 10 features, where 7 of them are discourse features.



circumstance relation = true: subjective class

exclamation point = false AND
unconditional relation = false AND
contrast relation = false AND
multinuclear restatement relation = true AND
comparison relation = false AND
conjunction relation = false AND
negative words <= 0.5 AND
intensification words <= 0.5 AND
pagerank centrality <= 0.119 AND
katz index centrality <= -0.029: objective class

multinuclear restatement relation = 1: subjective class

exclamation point = false AND
unconditional relation = false AND
contrast relation = false AND
disjunction relation = true AND
volitional cause relation = false AND
proportion of subjective words in nucleus <= 0.283: subjective class

exclamation point = false AND
unconditional relation = false AND
contrast relation = false AND
volitional cause relation = true AND
katz index centrality <= 0.912: subjective class

Figure 4. Some of the rules learned by the PART method tested with all the features.

Looking at (i) such observations, (ii) the best results that were achieved with the machine learning techniques (using all the features) and (iii) the occurrence patterns of discourse relations that were observed during the corpus annotation step, we may conclude with some confidence that discourse does help signaling subjectivity and that this is a linguistic level that is valuable to explore. In what follows, we comment about future work in such line and present some final remarks.

6. Final Remarks

This paper presented the investigation of machine learning methods of different paradigms and richer feature sets for performing subjectivity classification for Portuguese language. Our results show that the combination of lexical, centrality-based and discourse features produce better results than any of the feature sets individually, considering a very challenging dataset. We also show that subjectivity may be clearly signaled by some discourse relations and, as our corpus annotation reveals, there are some distinctive discourse structuring patterns for sentence subjectivity.

Dealing with discourse, however, is not straightforward. In contrast to other linguistic levels, there are limited parsing tools for producing discourse information for the classification step. In this paper, we have manually annotated our corpus according to RST, but it remains for future work to measure the impact of using fully automatic discourse parsing for the task (e.g., the ones of [22,23]). This will probably affect the achieved results.

We also believe that explicit discourse relations are not the only kind of discourse pattern that may be learned for subjectivity classification. Other models may also help in the task. For instance, the entity grids of [24] may be useful for finding distinctive entity distributions that may differentiate subjective from objective content in longer texts. The appraisal theory [25] may also reveal different types of subjective content and help distinguishing it from objective texts. More than practical results, studies like these might bring relevant theoretical contributions to the area.

To the interested reader, more information about this work may be found at the OPINANDO project webpage (https://sites.google.com/icmc.usp.br/opinando/).

Author Contributions: L.B.B. and L.G.F. have conducted the experiments and written the first version of this paper. T.A.S.P. have supervised the work and also helped reviewing and editing this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FAPESP (2018/11479-9) and USP Research Office (PRP 668).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [CrossRef]
2. Moraes, S.M.W.; Santos, A.L.L.; Redecker, M.; Machado, R.M.; Meneguzzi, F.R. Comparing Approaches to Subjectivity Classification: A Study on Portuguese Tweets. In Proceedings of the International Conference on the Computational Processing of the Portuguese Language (PROPOR), Tomar, Portugal, 13–15 July 2016; pp. 86–94.
3. Bertaglia, T.F.P.; Nunes, M.G.V. Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), Osaka, Japan, 11 December 2016; pp. 112–120.
4. Avanco, L.V.; Nunes, M.D.G.V. Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese. In Proceedings of the 2014 Brazilian Conference on Intelligent Systems (BRACIS), São Carlos, Brazil, 18–23 October 2014; pp. 277–281.
5. Condori, R.E.L.; Pardo, T.A.S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Syst. Appl.* **2017**, *78*, 124–134. [CrossRef]
6. Belisário, L.B.; Ferreira, L.G.; Pardo, T.A.S. Evaluating Methods of Different Paradigms for Subjectivity Classification in Portuguese. In Proceedings of the 14th International Conference on the Computational Processing of Portuguese (PROPOR), Evora, Portugal, 2–4 March 2020; pp. 261–269.
7. Mann, W.; Thompson, S. *Rhetorical Structure Theory: A Theory of Text Organization*; Technical Report ISI/RS-87-190; Information Science Institute, University of Southern California: Los Angeles, CA, USA, 1987.
8. Vargas, F.A.; Pardo, T.A.S. Hierarchical clustering of aspects for opinion mining: A corpus study. In *Linguística de Corpus: Perspectivas*; Finatto, M.J.B., Rebechi, R.R., Sarmento, S., Bocorny, A.E.P., Eds.; Instituto de Letras da UFRGS: Porto Alegre, Brazil, 2018; pp. 69–91.

9. Vilarinho, G.N.; Ruiz, E.E.S. Global centrality measures in word graphs for Twitter sentiment analysis. In Proceedings of the 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, Brazil, 22–25 October 2018; pp. 55–60.
10. Zhao, H.; Lu, Z.; Poupart, P. Self-Adaptive Hierarchical Sentence Model. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina, 25–31 July 2015; pp. 4069–4076.
11. Yu, H.; Hatzivassiloglou, V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Sapporo, Japan, 11–12 July 2003; pp. 129–136.
12. Wiebe, J.; Bruce, R.; O'Hara, T. Development and use of a gold standard data set for subjectivity classifications. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), College Park, MD, USA, 22–26 June 1999; pp. 246–253.
13. O'Donnell, M. RSTTool 2.4—A Markup Tool for Rhetorical Structure Theory. In Proceedings of the International Natural Language Generation Conference (INLG), Mitzpe Ramon, Israel, 12–16 June 2000; pp. 253–256.
14. Chenlo, J.M.; Hogenboom, A.; Losada, D.E. Rhetorical Structure Theory for polarity estimation: An experimental study. *Data Knowl. Eng.* **2014**, *94*, 135–147. [CrossRef]
15. Freitas, C.; Motta, E.; Milidiú, R.; Cesar, J. Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. In Proceedings of the XI Encontro de Linguística de Corpus (ELC), São Carlos, Brazil, 13–15 September 2012; pp. 1–12.
16. Carvalho, P.; Silva, M.J. Sentilex-PT: Principais Características e Potencialidades. *Oslo Stud. Lang.* **2015**, *7*, 425–438. [CrossRef]
17. Pasqualotti, P.R.; Vieira, R. WordnetAffectBR: Uma base lexical de palavras de emoções para a língua portuguesa. *Rev. Novas Tecnol. Educ.* **2008**, *6*, 1–10. [CrossRef]
18. 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. Manag.* **2012**, *48*, 741–754. [CrossRef]
19. Mikolov, T.; Corrado, G.; Chen, K.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
20. Hartmann, N.S.; Avanço, L.; Balage, P.P.; Duran, M.S.; Nunes, M.G.V.; Pardo, T.; Aluísio, S. A Large Opinion Corpus in Portuguese—Tackling Out-Of-Vocabulary Words. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 26–31 May 2014; pp. 3865–3871.
21. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Cambridge, MA, USA, 2016.
22. Maziero, E.G.; Hirst, G.; Pardo, T.A.S. Adaptation of Discourse Parsing Models for Portuguese Language. In Proceedings of the 4th Brazilian Conference on Intelligent Systems (BRACIS), Natal, Brazil, 4–7 November 2015; pp. 140–145.
23. Maziero, E.G.; Hirst, G.; Pardo, T.A.S. Semi-Supervised Never-Ending Learning in Rhetorical Relation Identification. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria, 7–9 September 2015; pp. 436–442.
24. Barzilay, R.; Lapata, M. Modelling local coherence: An entity-based approach. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, MI, USA, 25–30 June 2005; pp. 141–148.
25. Martin, J.R.; White, P.R.R. *The Language of Evaluation: Appraisal in English*; AIAA: London, UK, 2005.