Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3272–3280
Marseille, 11–16 May 2020

Computing with Subjectivity Lexicons

Caio L. M. Jeronimo, Claudio E. C. Campelo, Leandro Balby Marinho, Allan Sales
Adriano Veloso, Roberta Viola
UFCG - Brazil, UFMG - Brazil
Campina Grande - Brazil, Belo Horizonte - Brazil
{caiolibanio, allanmelo}@copin.ufcg.edu.br, {campelo, lbmarinho}@computacao.ufcg.edu.br
adrianov@dcc.ufmg.br, robertaviola@ufmg.com

Abstract

In this paper, we introduce a new set of lexicons for expressing subjectivity in text documents written in Brazilian Portuguese. Besides the non-English idiom, in contrast to other subjectivity lexicons available, these lexicons represent different subjectivity dimensions (other than sentiment) and are more compact in number of terms. This last feature was designed intentionally to leverage the power of word embedding techniques, i.e., with the words mapped to an embedding space and the appropriate distance measures, we can easily capture semantically related words to the ones in the lexicons. Thus, we do not need to build comprehensive vocabularies and can focus on the most representative words for each lexicon dimension. We showcase the use of these lexicons in three highly non-trivial tasks: (1) Automated Essay Scoring in the Presence of Biased Ratings, (2) Subjectivity Bias in Brazilian Presidential Elections and (3) Fake News Classification Based on Text Subjectivity. All these tasks involve text documents written in Portuguese.

Keywords: subjectivity, lexicon, word embedding

1. Introduction

Many everyday tasks in Natural Language Processing (NLP) rely on lexicons, which are vocabularies that represent different branches of knowledge. Lexicons are useful resources for identifying semantics relevant to sentiment, emotion, personality, language bias, mood, and attitude. Using lexicons in computational tasks has enabled several research works and exciting new applications to arise, such as detecting sentiment towards politicians and products, frustration of callers in call centers, stress in drivers or pilots, depression, media bias, biased ratings in essay scoring, and fake news classification.

Choi and Wiebe (2014), for example, develop a sense-level lexicon to support opinion mining. The authors combine a graph-based method, based on WordNet relations, with a standard classifier, trained on gloss information, and show that the model can be useful to guide manual annotations to find positive/negative senses. Recasens et al. (2013) identify familiar linguistic cues related to framing and epistemological bias, which include factive verbs, implicative, hedges, and subjective intensifiers. From these cues, the authors developed a bias lexicon for extracting features for a model to solve the problem of bias-inducing word identification.

The recent advances in Deep Learning have also brought relevant changes in the way lexicons are constructed and exploited. Zeng et al. (2018), for example, expand the well known LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001) resource for the Chinese language, by the usage of Sequence-to-Sequence models and word embeddings to classify words in the lexicons. The authors show that their model outperforms state-of-the-art approaches with a better comprehension of the words meanings.

While most of the lexical resources available are still aimed for English, the construction of linguistic resources for non-English documents represent an essential path to democratize the usage of NLP tools across multiple languages and ethnicities. MorphoBr (de Alencar et al., 2018) is a project that aims at building a full-form lexicon (i.e. a lexicon containing all inflected, declined and conjugated forms of a language) to perform Portuguese morphological analysis. The project provides resources that allow a straightforward compilation of finite-state morphological analyzers. Filho et al. (2013) evaluate LIWC[1] for sentiment classification in documents written in Portuguese. The authors find that LIWC, at least for Portuguese, performs better at detecting positive sentiment in contrast to negative.

In the revised literature, sentiment is usually linked to subjectivity, i.e., the more sentiment a text has, the more subjective it is. Sentiment is, however, a limited notion of subjectivity that fails to capture all its nuances. For example, an argumentative text, devoid of emotion, might still be subjective as one might be trying to convince someone else of a specific point of view using subjective arguments. Moreover, most of the sentiment/subjectivity resources are available for the English language. In this paper, we describe a new set of lexicons for the Brazilian Portuguese language that aims at expressing different levels of textual subjectivity in documents.

We propose a set of lexicons for representing five different subjectivity dimensions, that are: *argumentation*, *presupposition*, *sentiment*, *valuation*, and *modalization*. We have used these lexicons in a variety of different tasks such as (1) Automated Essay Scoring in the Presence of Biased Ratings, (2) Subjectivity Bias in Brazilian Presidential Elections and (3) Fake News Classification Based on Text Subjectivity. The first task uses the proposed lexicons to represent comments of human raters of the standardized Brazilian national exam, to examine how rater bias can affect models that learn how to score these essays automatically. In the second task, we have leveraged the subjectivity lexicons to characterize media bias across distinct news media outlets covering the Brazilian presidential elections and how these biases evolve over subsequent elections. Finally,

[1] `http://liwc.wpengine.com/`

in the third task, our lexicons are compared to existing ones in the context of fake news classification based on the news subjectivity levels. In all these tasks, the proposed lexicons proved to be useful in identifying subjectivity nuances in text documents written in Portuguese. In the rest of the paper, we will provide more details on the lexicons construction (Section 2) and recap the methodology and results for the three tasks mentioned above (Section 3). We conclude the paper in Section 4 with a summary of this work and directions for future work.

2. Subjectivity Lexicons Construction

The choice of certain words in discourse defines a subject's identity, thereby the bias and subjectivity can be tracked by lexicons, as stated by Anscombre and Ducrot (1976). The literature about bias and subjectivity in the discourse is vast in linguistics. Along with that, according to Verhagen (2005), the theory of argumentation and pragmatics assume implicit markers of positioning in language as clues to the speakers subjectivity. As indicated by Amorim et al. (2018), the marks that display subjectivity and bias are features of elements such as argumentative markers, specific kind of verbs, modal verbs, among other linguistic expressions. Based on the pragmatics theory and (Recasens et al., 2013), we have developed in the context of Page (1967) a handcraft list of Portuguese lexicons that indicate subjectivity in texts from a multitude of domains. These lexicons comprise five possible sources of subjectivity, as follows[2]:

- Argumentation: markers of argumentative discourse, including lexical expressions and connectives, such as: "even" (até), "as a consequence" (como consequência), "or else" (ou então), "as if" (como se), "rather than" (em vez de), "somehow" (de certa forma), "despite" (apesar de), among others.

- Presupposition: markers that suggest the rater assumes something is true. Some examples: "nowadays" (hoje em dia), "to keep on doing" (continuar a), and factive verbs.

- Modalization: expressions that indicates that the writer exhibits a stance towards its statement. Such markers are adverbs, auxiliary verbs, modality clauses, and some type of verbs.

- Sentiment: includes markers that indicate a state of mind or a sentiment or evaluation. In expressions such as: "with regret" (infelizmente), "fortunately" (felizmente), and "it is preferable" (preferencialmente).

- Valuation: This lexicon assigns a value to the facts. Usually, adjectives are employed as valuation, but as adjectives are context-dependent we use only in this class the markers that are related to intensification, such as: "absolutely" (absolutamente), "highly" (altamente), and "approximately" (aproximadamente).

While more traditional lexicons try to be as comprehensive as possible, LIWC, for example has more than six thousand words[3], we intentionally made ours more compact. Our lexicons have little more than 450 words. The choice of compactness was made intentionally to leverage the power of word embedding techniques, i.e., with the words mapped to an embedding space and the appropriate distance measures, we can easily capture semantically related words to the ones in the lexicons. Thus, we do not need to build comprehensive vocabularies and can focus on the most representative words for each lexicon dimension. A nice side effect is that this makes the computation with our lexicons lightweight.

3. Use Cases

In this section, we present three use cases that involve the identification of different levels of subjectivity in text written in Portuguese, in order to showcase the use of our lexicons.

3.1. Automated Essay Scoring in the Presence of Biased Ratings

Automated Essay Scoring (AES) aims at developing models that can grade essays automatically or with reduced involvement of human raters (Page, 1967). When given the same set of essays to evaluate and enough graded samples, AES systems tend to achieve high agreement levels with trained human raters (Taghipour and Ng, 2016).
However, there is a sizeable literature in cognitive science, psychology and other social studies offering evidence that biases can create situations that lead us to make decisions that project our experiences and values onto others (Baron, 2007). Since the research in AES has focused on designing scoring models that maximize the agreement with human raters(Chen and He, 2013; Alikaniotis et al., 2016), there is a lack of discussion on how biased human ratings affect AES performance. Thus, our objective here is to examine the extent to which rater bias affects the effectiveness of different AES models (Amorim et al., 2018).

3.1.1. Dataset

In order to study the effects of rater bias in essay scoring, we created an annotated corpus containing essays written by high school students as part of a standardized Brazilian national exam. Our corpus contains a number of essays, written in Portuguese, along with their respective scores. Further, raters must also provide a comment for each essay in order to ground their scores.
The dataset comprises 1,840 essays that were written by high-school students as part of a standardized Brazilian national exam. The final score is given as the sum of the scores associated with different aspects. Raters are supposed to perform impartial and objective evaluations, and they must enter specific comments in order to ground their scores. Thus, for each essay in our dataset we also have the corresponding rater comments.

[2]The lexicons with the complete list of associated terms are available at: :`https://github.com/caiolibanio/subjectivity_lexicons_PTBR`

[3]`https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf`

3.1.2. Method

AES models are built on the basis of predefined features (e.g. number of words, average word length, and number of spelling errors) that are given to a machine learning algorithm. The features used to build our AES models are discussed and evaluated in Amorim and Veloso (2017). Our method aims at analysing the subjectivity associated with rater comments, so that a better training set is obtained after filtering out essays associated with ratings that look biased. More specifically, our de-biasing method starts by finding the norm (in terms of the subjectivity within rater comments) for each score value (i.e., intervals between 0 and 10). Intuitively, the amount of subjectivity within a comment should be similar to the amount of subjectivity within another comment, given that the scores associated with the corresponding essays are close to each other. So, we should not expect to find essays having discrepant scores, but for which the corresponding comments show a similar amount of subjectivity. Our method is divided into three steps:

1. Rater comments are represented according to the amount of subjectivity cues. In order to represent a comment, we calculate the distance between it and each of the five subjectivity lexicons. More specifically, we learn word embeddings (Mikolov et al., 2013c) for the Portuguese language, and then we employed the Word Mover's Distance function (Kusner et al., 2015) between a comment and the five subjectivity lexicons. As a result, each comment is finally represented by a five-dimensional subjectivity vector, where each dimension corresponds to the amount of a specific type of subjectivity. This results in a subjectivity space, where comments are placed according to their amount of subjectivity.

2. We group subjectivity vectors according to the score misalignment (AES model vs. human rater) associated with the corresponding essay. Then, we calculate centroids for each group in order to find the prototypical subjectivity vector for each group (or misalignment level).

3. The distance to the prototypical subjectivity vector is used as a measure of deviation from the norm. Specifically, we sort essays according to the distance between the subjectivity vector and the corresponding centroid. Then, we define a number of essays to be removed from the training set. The relative number of essays to be removed from the training set is controlled by hyper-parameter α.

3.1.3. Validation

We implemented AES models using different machine learning algorithms. Specifically, we learn AES models using Support Vector Regression (SVR), Random Forests (RF), Logistic Regression (LR), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP). All models are based on the same set of features. The measure used to evaluate the effectiveness of the different models is the quadratic weighted kappa (κ) which measures the inter-agreement between human raters and AES models (Cohen, 1960). We conducted five-fold cross validation, where the dataset is arranged into five folds with approximately the same number of examples. At each run, four folds are used as training set, and the remaining fold is used as test set. We also kept a separate validation set. The training set is used to learn the models, the validation set is used to tune hyper-parameters and the test set is used to estimate κ numbers for the different the models. The results reported are the average of the five runs, and are used to assess the overall effectiveness of each AES model. To ensure the relevance of the results, we assess the statistical significance of our measurements by comparing each pair of models using a Welch's t-test with $p-$value ≤ 0.01.

In order to properly evaluate our de-biasing method, we employ a set of 50 separate essays with bias-free ratings as our test set (Amorim et al., 2018). In this case, biased ratings are manually removed from the training set, and the test set is composed by unbiased ratings. Table 2 shows κ numbers for different α values. As expected, the inter-agreement increases significantly with α, until a point in which keeping removing essays from the training set becomes detrimental. This happens either because we start to remove unbiased ratings, or the training set becomes too small. In all cases, the MLP model showed to be statistically superior than the other models.

	κ				
α	SVR	RF	LR	GB	MLP
–	.404	.410	.408	.432	**.446**
0.1	.390	.339	.364	.378	**.393**
0.2	.365	.331	.344	.370	**.393**
0.3	.345	.326	.338	.365	**.386**
0.4	.340	.324	.333	.361	**.384**
0.5	.307	.317	.328	.358	**.382**

Table 1: κ numbers for different models with varying α values. There are potentially biased ratings in the test set.

	κ				
α	SVR	RF	LR	GB	MLP
–	.451	.472	.466	.491	**.521**
0.1	.467	.491	.481	.505	**.544**
0.2	.481	.511	.490	.521	**.562**
0.3	.488	.526	.497	.542	**.571**
0.4	.491	.523	.499	.547	**.569**
0.5	.481	.518	.494	.545	**.560**

Table 2: κ numbers for different models with varying α values. Ratings in the the test set are likely to be unbiased.

3.2. Subjectivity Bias in Brazilian Presidential Elections

It is not unusual to see different news outlets framing the same events differently in order to make one specific perspective looks better than another and, possibly, influence people's judgment about the event. This behavior is indicative of a Media Bias type known as Framing Bias, which aims to influence the reader's opinion towards a particular

interpretation of the event (Entman, 2007).
When it comes to news articles, it is possible to avoid the Framing Bias with the adoption of objective reporting, which requires the journalist to remain impartial and describe facts with objectivity and neutrality, despite his/her opinion about the topic (Clark, 2014). However, more often than desired, a more subjective reporting is observed in news articles, indicating an attempt to influence the reader's judgment (Mihalcea et al., 2007; Wiebe et al., 2004) on some subject.
With that in mind, we have proposed a methodology based on word embeddings, WMD, and the our Subjectivity Lexicons for measuring the subjectivity levels of news articles (Sales et al., 2019). We have used news articles reporting the last three Brazilian presidential elections (e.g., 2018, 2014, 2010) as input for our methodology. Presidential elections are convenient events since media bias tends to appear more prominently (Hamborg et al., 2018). We used our methodology to spot differences in subjectivity language across different news outlets covering the same candidates and parties. Next, we describe our method and show how we employed the Subjectivity Lexicons.

3.2.1. Method

As input, the method requires a word embedding model for representing the language vocabulary, a news article set, and the Subjectivity Lexicons. As output, it returns a 5-dimensional vector representing the articles associated with subjectivity values.
We train a word2vec model (Mikolov et al., 2013b; Mikolov et al., 2013a) based on the Portuguese Wikipedia dump and, then, calculate the WMD between each article and lexicon to represent the subjectivity values. The WMD takes two documents as input (e.g., the news article and the lexicon) and compute the distance between them as the sum of the Euclidean distance between the set of words of the two documents. Note that since WMD is a distance metric, the lower its value, the higher the subjectivity associated with the article. Intuitively, each value quantifies the amount of subjectivity related to a target article in terms of a subjectivity dimension.
The obtained values are used, later, for estimating the presence of Framing Bias by comparing, for each election and news outlet, whether the subjectivity of the news outlet's reports about two opposing parties/candidates are significantly different or not.

3.2.2. Dataset

We crawled news articles, summarized in Table 3, from the politics section of four distinct news outlets in Brazil: *Folha de Sao Paulo (FolhaSP)*[4], *Estadao*[5], *Carta Capital*[6] and *Veja*[7]. *FolhaSP* and *Estadao* are two of the most popular news outlets in Brazil, accessed by people from all political ideology-wings and, therefore, representatives of the Brazilian mainstream media. On the other hand, *Carta Capital* and *Veja* are biased representatives of the media. *Carta Capital* is a popular self-declared left-wing news outlet, while *Veja* is used here as a right-wing representative since the news outlet declares itself as a watchdog (against the government despite the government ideology orientation). In practice, it is observed a strong opposing message in *Carta Capital's* and *Veja's* discourses after the Brazilian 2014 presidential election, where *Carta Capital* is presented in favor of the government and *Veja* against it (de Matos and Formentin, 2016).

News Outlet	Year	# News	Total
Carta Capital	2010	1,264	4,651
	2014	997	
	2018	2,390	
Veja	2010	2,778	16,902
	2014	6,226	
	2018	7,898	
Estadao	2010	10,072	52,238
	2014	8,717	
	2018	33,449	
FolhaSP	2010	11,895	41,339
	2014	11,816	
	2018	17,628	

Table 3: Number of news by source and year.

3.2.3. Brazilian Elections

The Brazilian presidential election occurs in a two-round system, in which the elected candidate (often affiliated with a party) is the one who obtains more than 50% of the valid votes. In case neither candidate receives a majority, it is decided in a second round between the two most voted candidates. Coincidentally, all the last three elections had a second round composed by one left- and one right-wing party.

3.2.4. Validation

For validating the methodology and checking its reliability, we run an experiment comparing the subjectivity associated with each news outlet and a sample of 29,000 randomly picked Wikipedia articles. Since Wikipedia has policies that enforce a less biased point-of-view in its articles[8], we would expect Wikipedia to present a lower level of subjectivity than the news outlets. Also, since our data consist of news of two biased and two mainstream news outlets, it would be expected to find more subjectivity associated with the two biased sources.
Figure 1 presents the subjectivity confidence interval of the mean for the news outlets and Wikipedia per subjectivity dimension. Results show lower levels of subjectivity in Wikipedia's articles for all dimensions as expected. Also, the subjectivity of *Veja* and *Carta Capital*, the two biased sources, are higher than the other news outlets for all dimensions except Valuation, also as expected.

3.2.5. Application

This experiment aims to show the difference of subjectivity associated with news articles reporting the two parties that

[4] https://www.folha.uol.com.br/
[5] https://www.estadao.com.br/
[6] https://www.cartacapital.com.br/
[7] https://veja.abril.com.br/
[8] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

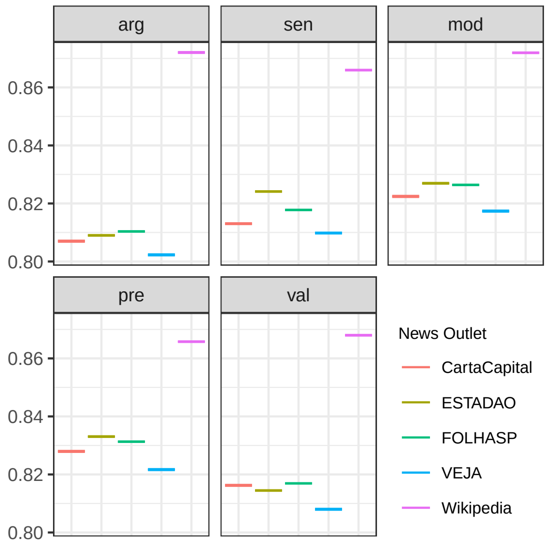

Figure 1: Subjectivity confidence intervals for the mean of news outlets and Wikipedia's articles across the subjectivity dimensions argumentation (arg), sentiment (sen), valuation (val), modalization (mod), and presupposition (pre).

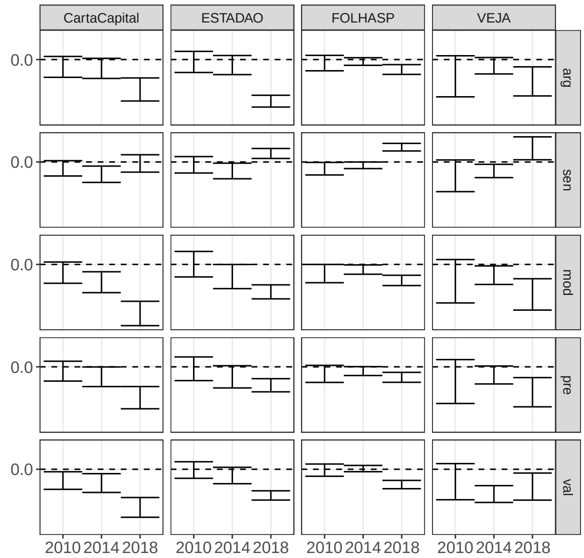

Figure 2: Confidence intervals for the mean of the difference of subjectivity in news related to parties that made the second round. Intervals below or above zero mean subjectivity bias towards the left-wing or right-wing party, respectively.

made the second round in each election. Since we always had one left- and one right-wing party in the second round, we are, in some sense, computing the bias trends regarding opposite ideology-wings throughout the elections.
The confidence intervals were computed with 95% of confidence level, and for determining which news was reporting about one party, we select those news that referenced the party in its headline. Confidence intervals containing the zero value indicates no subjectivity differences in the reports about the left and right-wing parties. Confidence intervals entirely negative or positive indicate significant more subjectivity in the news about the left or right-wing parties, respectively.
The results, depicted in Figure 2, show a rising trend of subjectivity throughout the elections. In 2010, all but *Carta Capital's* Valuation confidence intervals contained the zero value. In 2014, the majority of *Carta Capital's* and *Veja's* confidence intervals were indicating more subjectivity in the news reporting the left-wing party. Finally, in 2018, the majority of all news outlets confidence intervals are showing more subjectivity in the news about the left-wing party. More details of this study can be found at (Sales et al., 2019).

3.3. Fake News Classification Based on Text Subjectivity

Many studies use subjectivity lexicons to perform fake news classification (Volkova et al., 2017; Wang et al., 2018; Zhou et al., 2019). The usage of lexicons to extract subjectivity aspects from documents can bring relevant insights from fake and legitimate news, exposing the usage of biased language and emotional polarities of documents, for example. The main assumption regarding the usage of subjectivity to identify fake news is that fake documents may rely on a more subjective writing style, for example appealing to persuasive arguments and emotion (Potthast et al., 2018).
In order to evaluate the usefulness of our lexicons for fake news classification, we compare the performance of classifiers that exploit them for fake news detection. We consider the performance of such classifications in Portuguese, using our subjectivity lexicons, against the performance of classifiers using other existing lexicons, but in English language. The methodology used for such evaluation will be described in the following sections.

3.3.1. Dataset

The dataset of legitimate news in Portuguese were collected from two of the biggest news sites in Brazil, that are *Estadao* and *Folha de Sao Paulo*. As these two news sites are mainstream, we consider that all their news are legitimate. We collected a total of 207,914 legitimate news, from 2014 to 2017. We divided the news into different domains: Politics, Sports, Economy, and Culture. Table 4 shows the distribution of the news in these topics.
The fake news dataset in Portuguese is composed of fact-checked fake news that were strongly disseminated in Brazil, from 2010 to 2017. We collected these news from two popular fact-checking services, that are *e-Farsas*[9] and *Boatos*[10]. These two services keep track of the most "viral" fake news, providing the link to the fake article and the evidence that the news is fake. We collected a total of 121 fake news from more than 40 news sources. This dataset, although small, has the properties of being highly disseminated in Brazilian's social networks and web, meaning that, in fact, they deceived people. These fake news also have the advantage that they come from a wide variety of different

[9] http://www.e-farsas.com/
[10] http://www.boatos.org/

Domain	Estadao	FolhaSP	Total	%
Politics	24,638	30,765	55,403	26.6
Sports	31,692	31,908	63,600	30.5
Economy	20,512	30,412	50,924	24.4
Culture	15,456	22,531	37,987	18.2

Table 4: Distribution of the Portuguese legitimate news by domains.

sources, which may reflect different writing styles.

The English legitimate news were collected from a popular Kaggle dataset called "All the News"[11] between the years of 2016 and 2017. From this dataset, we collected news from *The Guardian* (1798 articles), from *New York Times* (1598 articles) and 2598 articles from *CNN*. The English fake news dataset was compiled by (Torabi Asr and Taboada, 2019) and correspond to political fake articles from *Snopes*[12] (103 articles from fake and mostly-fake categories), political fake news from (Horne and Adali, 2017) (75 articles) and the top fake news collected from Buzzfeed[13] (41 articles).

3.3.2. Related Lexicons

In order to compare our lexicon's usefulness, we consider three different set of lexicons used to access different aspects of subjectivity in English, here called as Related Lexicons (RL). The first one has been compiled by (Recasens et al., 2013). This lexicon comprises six different aspects/dimensions of bias-inducing terms, that are:

1. Factive Verbs: presuppose the truth of a complement clause, e.g., realize, forget, exciting. (27 terms)

2. Implicative Verbs: imply the truth or untruth of their complement clause, e.g., succeed, fail, neglect. (32 terms)

3. Assertive verbs: are those verbs that their complement clauses assert a proposition, e.g., believe, figure, affirm. (66 terms)

4. Hedges: used to reduce commitment to the truth of a proposition, thus avoiding direct statements, e.g., apparently, could, estimate. (100 terms)

5. Reporting Verbs: usually used to report other person's activities or actions, e.g., accuse, assure, claim. (181 terms)

6. Bias-inducing lemmas, e.g., advocate, amazing, barbarian. (654 terms)

The second RL we use has been presented by (Wilson et al., 2005). This lexicon is part of the Multi-Perpective Question Answering (MPQA) Subjectivity Lexicons[14] project, and is divided in sentiment polarities (positive and negative) classified by strong subjectivity and weak subjectivity. From this lexicon, we extracted the terms from the strong subjectivity category for both polarities. After filtering, we obtained 3,078 lexicons with negative polarity and 1,482 with positive polarity. The third RL is the lexicon proposed by (Deng et al., 2013), and also represents sentiment polarities (positive and negative) extracted from subjective documents (i.e., editorials and blogs). This lexicon contains 1,003 terms for negative and 493 for positive sentiments.

Train	Test
LN (340) × FN (85)	LN (144) × FN (36)

Table 5: Table of evaluation scenario. The abbreviations are LN (Legitimate News), FN (Fake News). The legitimate and fake news respect a distribution of 4:1, where 70% of fake news are used in training and 30% in test set.

3.3.3. Method

To perform the comparison between our new lexicons and the ones described in previous section, we use the method adopted by (Jeronimo et al., 2019), based on WMD distances between the news documents and each lexicon dimension. The documents are divided into sentences, and the distance of each sentence to a lexicon dimension is calculated, generating an average as a global distance from the document to each lexicon dimension. Thus, the output is a n-dimensional vector, where n is the number of different dimensions in the lexicon set.

These output vectors are then used to classify fake and legitimate news, using our lexicons for the news written in Portuguese (Brazilian news dataset) and using the RL for the other datasets containing news written in English.

This experimental approach aims at validating the hypothesis that our lexicons can generate results that are compatible with other well-established subjectivity lexicons present in literature, in terms of fake news classification. Such an analysis can reveal the level of usefulness of our new resources in detecting subjectivity nuances present in Portuguese fake and legitimate documents.

3.3.4. Validation

As classification models, we use both XGBoost and Random Forests, which are known to have a strong predictive power, achieving state-of-the-art result in complex classification problems (Olson et al., 2017).

Following (Silverman, 2016), we define the proportion of four legitimate news to one fake news (4:1), replicating the proportion found in an analysis of the 2016 US presidential election, where the authors found this proportion when analyzing Facebook news profiles. As the dataset of legitimate news is far larger than the fake one, for both Portuguese and English documents, we randomize the train/test executions by varying the legitimate news documents 500 times. With this method, we can generate the average results of the executions, and we can exploit a higher number of legitimate

[11] https://www.kaggle.com/snapcrack/all-the-news/version/4

[12] https://github.com/sfu-discourse-lab/Misinformation_detection

[13] https://github.com/BuzzFeedNews/2017-12-fake-news-top-50

[14] https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

news.
To calculate the semantic distances with WMD, we train a word embedding model from a large Wikipedia dump in Portuguese. For the English dataset, we use the pre-trained Google News embeddings[15].
Table 5 shows the dataset sizes used for classification of the documents. The sizes respect the (4:1) proportion between legitimate and fake news. We also consider a split of 70%-30% of fake news in the training and test sets, respectively. Since the Portuguese fake news dataset have less samples than the English one, we use the Portuguese fake news size (121 documents) as a base for the fake news distribution between train/test samples. To evaluate the models, we use the Area Under Precision-Recall Curve (PR-AUC). This metric reports the trade-off between the precision and recall values when varying the classifier's thresholds. In order to report a broader evaluation of the models, we also consider the classical ROC-AUC, that considers the trade-off between the recall and false positive rate, giving a better understanding of the model's performance in both classes of a binary classification.
Table 6 shows the results of classification using the three RL lexicons and our proposed lexicons, in terms of PR-AUC, for both XGboost and Random Forest. From such results, we can notice that the classifications using our new lexicons were quite comparable with the results presented by the other three lexicons for the English fake news classification. A similar result can be seen in Table 7, that shows the results in terms of ROC-AUC. As the ROC-AUC considers the trade-off between Recall and False Positive Rate, it provides a more global evaluation, since a low FPR will imply in better classification of the legitimate news (class 0). Again, as we perform an indirect evaluation, we avoid to say that our lexicons is better or worst then the RL ones, but even so, we can see a level of similarity between the results, without hugely differences for both metrics.
An important finding is that our new set of lexicons also showed a comparable performance with RL lexicons that have a greater number of terms. This is the case for (Wilson et al., 2005) and (Deng et al., 2013) lexicons. Such results can demonstrate that our lexicons may express different levels of subjectivity, but relying on less terms.
It is also important to highlight that while many studies use subjectivity in conjunction with other features for classifying fake news, our experiments do not include other features to make such a classification. This leads to inferior overall classification results, as expected, but which are useful for our purposes in this paper. In addition, other strategies to capture the subjectivity of the documents (rather than the average subjectivity of documents' sentences) can bring additional improvements to the models.

4. Conclusion and Future Work

We have presented a set of lexicons specifically designed to the extraction of subjectivity in text documents written in Brazilian Portuguese. The lexicons are composed by five subjectivity dimensions, covering different aspects of subjectivity, providing a better understanding of the language

[15] https://code.google.com/archive/p/word2vec/

Lexicons	XGBoost	RF
(Recasens et al., 2013)	0.24 ± 0.04	0.28 ± 0.05
(Wilson et al., 2005)	0.22 ± 0.04	0.22 ± 0.04
(Deng et al., 2013)	0.30 ± 0.05	0.31 ± 0.06
Proposed lexicons	0.25 ± 0.05	0.25 ± 0.05

Table 6: Average PR-AUC results for classifications using the RL lexicons and our proposed lexicons.

Lexicons	XGBoost	RF
(Recasens et al., 2013)	0.56 ± 0.05	0.61 ± 0.05
(Wilson et al., 2005)	0.53 ± 0.05	0.53 ± 0.05
(Deng et al., 2013)	0.59 ± 0.05	0.60 ± 0.05
Proposed lexicons	0.59 ± 0.04	0.61 ± 0.04

Table 7: Average ROC-AUC results for classifications using the RL lexicons and our proposed lexicons.

in three kinds of data: essay raters comments, legitimate journalistic news, and fake news. In the first scenario, the lexicons helped in creating a subjectivity space, where biased rating/scores can be perceived, improving AES models. In the second scenario, the proposed linguistic resources allowed the identification of variations in the levels of subjectivity of distinct news media outlets covering the same subjects. We showed that these variations tended to increase in subsequent elections. The third scenario shows an application of the lexicons in the fake news identification problem, demonstrating that our lexicons can be compared to other subjectivity lexicons present in literature. As future directions, we intend to expand the lexicons to cover other aspects of subjectivity, and the application of them in new tasks. We also plan to evaluate whether these lexicons can be used for other idioms, for example, through translation.

5. Acknowledgements

This work was supported in part by CAPES, MASWeb (Grant FAPEMIG/PRONEX APQ-01400-14), Kunumi and Fapemig.

6. Bibliographical References

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 715–725.
Amorim, E. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for brazilian portuguese. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Student Research Workshop*, pages 94–102.
Amorim, E., Cançado, M., and Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237.
Anscombre, J. and Ducrot, O. (1976). L'argumentation dans la langue. in: Langages. *Paris: Didier-Larousse*, (42):5–27.

Baron, J. (2007). *Thinking and Deciding*, volume 4. Cambridge University Press.

Chen, H. and He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752.

Choi, Y. and Wiebe, J. (2014). +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar, October. Association for Computational Linguistics.

Clark, R. (2014). The pyramid of journalism competence: what journalists need to know. *Poynter Media News*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

de Alencar, L. F., Cuconato, B., and Rademaker, A. (2018). Morphobr: an open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

de Matos, O. V. H. and Formentin, C. N. (2016). Veja e cartacapital: a polarização política das páginas às ruas. *Em Debate*, (16):15–40.

Deng, L., Choi, Y., and Wiebe, J. (2013). Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Sofia, Bulgaria, August. Association for Computational Linguistics.

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Filho, P. P. B., Pardo, T. A. S., and Aluísio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Hamborg, F., Donnay, K., and Gipp, B. (2018). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, pages 1–25.

Horne, B. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news.

Jeronimo, C., Marinho, L., Campelo, C., Veloso, A., and Sales, A. (2019). Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications and Services - iiWas)*.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., and Moore, J. H. (2017). Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*.

Page, E. (1967). Grading essays by computer: progress report. In *Proceedings of the Invitational Conference on Testing Problems*, pages 87–100.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia, July. Association for Computational Linguistics.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, August. Association for Computational Linguistics.

Sales, A., Balby, L., and Veloso, A. (2019). Media bias characterization in brazilian presidential elections. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, pages 231–240, New York, NY, USA. ACM.

Silverman, C. (2016). Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. buzzfeed, nov. 16.

Taghipour, K. and Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

Torabi Asr, F. and Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310.

Verhagen, A. (2005). *Constructions of intersubjectivity: Discourse, syntax, and cognition*. Oxford University Press on Demand.

Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Pro-*

ceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 647–653, Vancouver, Canada, July. Association for Computational Linguistics.

Wang, L., Wang, Y., de Melo, G., and Weikum, G. (2018). Five shades of untruth: Finer-grained classification of fake news. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 593–594, Aug.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 347–354, USA. Association for Computational Linguistics.

Zeng, X., Yang, C., Tu, C., Liu, Z., and Sun, M. (2018). Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. (2019). Fake news early detection: A theory-driven model. *arXiv preprint arXiv:1904.11679*.