Journal of Data Intelligence, Vol. 1, No. 4 (2020) 419–441
© Rinton Press

CHARACTERIZATION OF FAKE NEWS BASED ON SUBJECTIVITY LEXICONS

CAIO L. M. JERONIMO
Systems and Computing Department, Federal University of Campina Grande
Campina Grande, 58428-830, Brazil
caiolibanio@copin.ufcg.edu.br

LEANDRO BALBY MARINHO
Systems and Computing Department, Federal University of Campina Grande
Campina Grande, 58428-830, Brazil
lbmarinho@computacao.ufcg.edu.br

CLAUDIO E. C. CAMPELO
Systems and Computing Department, Federal University of Campina Grande
Campina Grande, 58428-830, Brazil
campelo@computacao.ufcg.edu.br

ADRIANO VELOSO
Department of Computer Science - Federal University of Minas Gerais
Belo Horizonte, 31270-901, Brazil
adrianov@dcc.ufmg.br

ALLAN SALES DA COSTA MELO
Systems and Computing Department, Federal University of Campina Grande
Campina Grande, 58428-830, Brazil
allanmelo@copin.ufcg.edu.br

While many works investigate spread patterns of fake news in social networks, we focus on the textual content. Instead of relying on syntactic representations of documents (aka Bag of Words) as many works do, we seek more robust representations that may better differentiate fake from legitimate news. We propose to consider the subjectivity of news under the assumption that the subjectivity levels of legitimate and fake news are significantly different. For computing the subjectivity level of news, we rely on a set subjectivity lexicons for both Brazilian Portuguese and English languages. We then build subjectivity feature vectors for each news article by calculating the Word Mover's Distance (WMD) between the news and these lexicons considering the embedding the news words lie in, in order to analyze and classify the documents. The results demonstrate that our method is robust, especially in scenarios where training and test domains are different.

Keywords: Fake news, Subjective Language, Misleading content detection

1. Introduction

The need for real-time fake news detection is clear and present given the ubiquitous reach of social media sites like Facebook and Twitter. Recognizing "made-up news" in real-time and enabling counter-measures on-the-fly have the potential to be a breakthrough technology.

Early examples of the immediate need for such technology were demonstrated by the arise of several accusations regarding the influence of fake news and hoaxes in the outcome of the 2016 United States presidential election. Despite the efforts to improve people awareness about fake news that circulate on the Web, there is still a lack of efficient automatized solutions for quickly and accurately detecting them.

If on one hand social networks enable political participation of users, on the other hand it helps pushing them toward ideological poles [1]. This environment of highly polarized opinions ends up facilitating the dissemination of fake news [2, 3, 4] or made-up stories with an intention to deceive. For example, during the week of Dilma Rouseff's[a] impeachment voting, three out of the five most shared news stories on Facebook were posteriorly found to be false.[b] Another recent example, is the 2018 Brazilian presidential elections where much evidence arouse about a massive dissemination of misleading content by the illegal usage of messaging applications.[c] The Lupa[d] fact-checking agency showed that from August to October of 2018, in the first round of Brazilian elections, ten of the most popular fake news had around 865,000 shares only in Facebook. Scenarios like these reinforce the need for new technologies that are able to early detect deceptive content.

Instead of relying on content representations of documents, as most of the related literature does (e.g. [5, 6, 7]), we seek more robust representations that may better differentiate hoaxes from real news. In this paper, we propose to build such representations considering the subjectivity level of news as abstract features, under the assumption that the subjectivity levels of legitimate and fake news are considerably different. Typically, documents that aim at sharing factual and impartial information, such as trustful journalistic articles and scientific papers, tend to use a more objective language that does not rely much on presuppositions or sentimental and argumentative expressions. By contrast, documents aiming at convincing or persuading tend to use a more subjective language [8, 9].

In this study, we conduct experiments in both Brazilian Portuguese and English languages. We collected a large scale dataset of legitimate news from two major mainstream media platforms in Brazil, namely *Folha de São Paulo*[e] and *Estadão*[f]. To collect examples of fake news, we relied on two Brazilian fact-checking services, *E-farsas*[g] and *Boatos*[h]. For the English experiments, we collected the legitimate news from popular sources like *The Guardian*, *New York Times* and *CNN*. The English fake news dataset was compiled from popular sources that tracks such documents, like *Snopes* and *BuzzFeed*. For computing the subjectivity level of news, we relied on subjectivity lexicons built by specialized Brazilian Portuguese linguists [10]. The English subjectivity lexicons were collected from resources already present in literature. The idea is that the more similar a news article is to these lexicons, the more subjective is the article. We calculate abstract features based on the semantic distance between the news

[a] Former Brazil's President.
[b] http://www.businessinsider.com/brazil-is-more-worried-about-fake-news-than-any-other-country-chart-2017-9
[c] https://www.dw.com/en/brazil-police-to-probe-allegations-of-election-disinformation-on-whatsapp/a-45965369
[d] https://piaui.folha.uol.com.br/lupa/2018/10/07/artigo-epoca-noticias-falsas-1-turno/
[e] https://www.folha.uol.com.br/
[f] https://www.estadao.com.br/
[g] http://www.e-farsas.com/
[h] http://www.boatos.org/

and the subjectivity lexicons in order to capture subjectivity features. We conduct a set of experiments driven by the following research questions:

Q1. Can we use the semantic distances provided by the use of subjectivity lexicons to find significant differences between legitimate and fake news?
Q2. Is it possible to determine, significantly, that fake news are more subjective than legitimate news?
Q3. Can the classification models based on the proposed subjectivity features outperform classical models based on BoW/TFIDF for fake news classification?
Q4. Can classification models based on subjectivity generalize better than models based on BoW/TFIDF (classical models)?

In Q1, we evaluate if the proposed approach, based on semantic distances between news documents and subjectivity lexicons can achieve significant differences in terms of subjectivity, between legitimate and fake news. We perform this verification by executing statistical tests over the semantic distances reported by the WMD. In Q2 we try to evaluate if the fake news are more subjective than the legitimate news. We perform this evaluation also by the use of statistical tests considering that the fake news distances in relation to the subjectivity lexicons should be smaller than the legitimate news distances. The Q3 evaluates the performance of classification models based on the proposed features, comparing them to classical models based on BoW/TFIDF for fake news classification. In Q4 we evaluate if the models based on subjectivity features can generalize better than models based on BoW/TFIDF. In this evaluation, we perform cross-domain classifications varying the topic of the news.

This article is an extension of the previous research presented in [11], providing the following increments:

- Inclusion of the English language in the experiments and evaluations;
- Usage of SHAP[i]as a robust technique for accessing models explanations, providing better insights regarding the fake news classifications;
- Remodeled analysis in order to give a broader view about the distribution of the semantic distances that we use;
- Evaluation of the Vector Subjectivity per Sentence (VSS), a new pre-processing method over the generated features.

2. Related Work

Studies involving the spread of deceptive content are vast. However, only recently, with the remarkable advances in Natural Language Processing, Social Media Mining, and Machine Learning it was possible to understand in more depth the characteristics of such content as well as how users interact with it. In what follows, we discuss previous work related to ours.

A recent survey on fake news detection in social media is presented in [12]. The authors describe the main aspects of fake news detection and provide characterizations on psychology and social theories, existing algorithms, evaluation metrics and data sets. They also point out

[i] https://github.com/slundberg/shap

that although some publicly available data sets exist [13, 14, 15], there is as yet no agreed-upon benchmark for this problem.

Besides text, auxiliary information such as user social engagement in the social media is typically used. For example, the works of [16, 17] consider the number of "likes" and the comments related to posts containing misleading information. One drawback of such approaches is that they are only able to detect fake news after they already gained traction in the dissemination chain.

Fake news detection based solely on textual content is even more challenging since fake news are intentionally written in order to mislead readers to believe false information. Most of the approaches in this direction consider lexical features of the text. Approaches using Bag of Words (BoW) and other simple features like the size of the documents are common [5, 6, 7]. Perez-Rosas et al. [18] introduce two novel datasets for fake news detection. These datasets consider seven different domains (sports, business, entertainment, politics, technology, education, and celebrities). Based on a set of linguistic features such as n-grams, punctuation, psycho-linguistic, readability and syntax (context free grammars) they report accuracies of up to 76%. It is noteworthy that the worst results were observed in cross-domain experiments where the models were trained on some domains and evaluated on others not present in training. This result, in particular, implies that more robust news representations are needed.

The only research work we have found addressing fake news classification in Portuguese is the one presented in [19]. The authors try out similar features as Perez-Rosas et al. [18] and find out that the best results come from BoW-based document representation.

In [5] it is pointed out that fake news are more similar to news satire than to real news. The authors extract lexical features from the documents and train a classifier considering the classes *real*, *fake* and *satire*. They show that it is harder to correctly discriminate between satire and fake news than between any of these and real news. This observation presents an interesting opportunity for using satirical texts (usually more abundant in the Web than fake news) in a transfer learning setting, i.e,. train on satires to predict fake news.

The authors of [20] go beyond lexical features searching for stylistic cues that may help determine the truthfulness of news. They compare texts from real news with three categories of news: *propaganda*, *hoaxes*, and *satire* with the aim of understanding the main characteristics of unreliable texts. They investigate the frequency of words present in some types of lexicons, represented as sets of tokens, in each of the aforementioned news categories. As lexicon types, they used *lying*, *subjective* or *sentimental*, *hedging* and *intensity*. The lexicons were retrieved from available linguistic resources in the Web. Among the main findings are that first and second person pronouns, as well as superlatives and modal adverbs, are used more often by fake than real news. The authors also try to predict news into *trusted*, *satire*, *hoax*, or *propaganda*, using a BoW representation of documents based on the used lexicons and report a F1 score of 65%.

A recent survey presented in [21] goes beyond textual content and provides a broader characterization of the classification problem by considering different kinds of news data. The authors argue that the main challenges are related to: the use of multi-modal datasets, covering all forms of fake news data; the development of multimodal detection approaches for considering, besides text, audio, multimedia and embedded content; the development of

methods for checking the quality of the news source; and the development of methods for checking the credibility of the news article authors.

These works reveal several solutions for tackling the fake news detection problem. We bring novel contributions in comparison to them. Similarly to [5, 6, 7, 20], we also focus on textual content, however, instead of relying on lexical features, we rely on semantic representations of words based on a small set of subjectivity dimensions. We evaluate the effectiveness of such semantic representation in research questions Q1, Q2 and Q3. Similarly to [18], we consider different news domains; however, we are able to show that, in our case, the model performance does not degrade with out of domain test examples. We evaluate this scenario in Q4, by swapping news domains and sources across train and test sets. Although the great majority of the existing works address fake news classification in English, we concentrate our efforts on both English and Portuguese.

3. Subjectivity-based Representation

Fake news is defined by The Ethical Journalism Network[j]as "Information deliberately fabricated and published with the intention to deceive and mislead others into believing falsehoods or doubting verifiable facts". In accordance with this definition, we observed that fake news are usually written using language that is deliberately inflammatory, and frequently present only one viewpoint. Often, the articles are designed to provoke an emotional response in order to entice readers into sharing them widely. Thus, instead of using lexical or morphological features (e.g. n-grams, number of nouns, word counts) for fake news classification, we developed features related to the subjectivity level of the articles. Specifically, we consider the semantic distances between subjectivity lexicons and articles as our unique set of features used to differentiate trustful from fake news.

3.1. *Subjectivity Lexicons*

We employ five subjectivity lexicons [10] for Portuguese, which are described and translated to English:

- The **argumentation** dimension represents words and expressions that are related to a more argumentative discourse. Such discourse is often used when someone is trying to convince another person of a specific point of view, e.g., at least, for this reason. (116 terms)
- The **presupposition** dimension encompasses terms that are related to a previous assumption of something. This kind of discourse is mainly used in situations where the interlocutor assumes something as true, even when this is not the case, e.g., to demonstrate, find out (54 terms)
- The **sentiment** lexicon contains words and terms related to emotional discourse. Such terms are also used in the context of fake news when the writer of the article tries to emotionally engage the reader, e.g., love, terrorize (151 terms).
- The **valuation** dimension expresses words related to the amount or intensification of something, e.g., completely, hugely (81 terms).

[j] https://ethicaljournalismnetwork.org/

- The **modalization** discourse is used when the interlocutor has an established stance about something or someone, e.g., advice, believe (55 terms).

In order to perform the experiments with the English data, we consider three different set of lexicons used to access different aspects of subjectivity in English. The first one is used by [22] and comprises six different dimensions:

- **Factive Verbs**: presuppose the truth of a complement clause, e.g., know, notice, remember. (27 terms)
- **Implicative Verbs**: imply the truth or untruth of their complement clause, e.g., condescend, happen, can. (32 terms)
- **Assertive verbs**: are those verbs that their complement clauses assert a proposition, e.g., think, acknowledge, affirm. (66 terms)
- **Hedges**: used to reduce commitment to the truth of a proposition, thus avoiding direct statements, e.g., approximately, apparently, almost. (100 terms)
- **Reporting Verbs**: usually used to report other person's activities or actions, e.g., announce, advise, argue. (181 terms)
- **Bias-inducing lemmas**: denotes a previously established or biased stance, e.g., apologetic, advocate, agree. (654 terms)

Another set of lexicons we use is part of the Multi-Perpective Question Answering (MPQA) Subjectivity Lexicons[k]project [23], 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, obtaining a total of 3,078 lexicons with negative polarity and 1,482 with positive polarity. The third set of lexicons that we use for the English language is the one used in [24]. This set of lexicons was extracted from subjective documents (e.g., editorials and blogs), and also represents sentiment polarities. We use the portion called "gold standard", which is a set of manually annotated lexicons, containing 1,003 terms for negative and 493 for positive sentiments.

3.2. *Semantic Distances as Features*

In order to build our features, we first trained a word embedding language model using a large Wikipedia dump [25, 26] for the Portuguese language experiments. For the English language experiments, we use the Google News embeddings[l]. To calculate the semantic distances between the news and the subjectivity lexicons, related to an embedding space, we use the Word Mover's Distance (WMD) [27].

Shortly, the WMD distance metric [27, 28] computes the minimum distance that a word from a document needs to "travel" to reach a word in another document in the embedding space. WMD assumes an embedding matrix $X \in R^{d \times n}$ for n words in the vocabulary where $x_i \in R^d$ is the embedding representation of the i^{th} word in a d-dimensional space. The model

[k]https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[l] https://code.google.com/archive/p/ word2vec/

also assumes two documents d and d' represented as normalized BoW. The WMD uses a "flow" matrix $\mathbf{T}$ to denote how much a word i in document d travels to word j in document d'. The distance between word i and word j becomes $\|x_i - x_j\|_2$. So, the method learns the flow matrix $\mathbf{T}$ to minimize

$$\begin{aligned} \min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|x_i - x_j\|_2 \quad subject\ to, \\ \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i, \sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \ \ \forall i, j \end{aligned} \tag{1}$$

Basically, the WMD returns a distance measure between 0 and 1, where smaller values mean more similar documents, considering the embedding space used. In order to perform the classifications, we execute two different feature engineering over the WMD distances. The first one we call Average Subjectivity per Document (ASD). In this feature engineering process, we calculate the semantic distance between each document sentence and a subjectivity lexicon, generating an average distance that represents the semantic distance between the document and the lexicon. For example, in the Portuguese dataset, since we have five subjectivity lexicons, each document will be represented as a 5-dimensional vector, where each value of the vector represents the average distance of the document sentences to one of the five lexicons. So, each distance between a document and one lexicon is calculated as:

$$\frac{1}{n} \sum_{i=1}^{n} WMD(x_i, l) \tag{2}$$

Where for a document x, we calculate the semantic distances using the WMD function for the n document sentences, generating an average that represents the distance of the document x to the lexicon l.

Another feature engineering process that we experiment is what we call Vector Subjectivity per Sentence (VSS). In this scenario, the features are treated considering the distance of each sentence of the document in relation to a given lexicon l, however, here there is no extraction of averages, but the values of the distances are used directly for each sentence. Thus, a document containing n sentences will be represented by a vector that contains all of the distances for each sentence in relation to a subjectivity lexicon. Given the variability in the size of the documents in terms of sentences, and the need to standardize the feature vectors in terms of their dimensions, we perform a padding with the average of the distances. Thus, a maximum limit of 100 sentences per document is established for all experiments. In this way, a document containing ten sentences will be represented by a vector containing the ten semantic distances and the remaining dimensions of the vector are filled with the average of these ten distances.

4. Dataset Characterization

In this section, we describe the datasets we use in the experiments. The datasets are composed by fake and legitimate news from both Portuguese and English languages.

4.1. *Data collection*

Table 1. Number of trustful news per news domain.

Domain	**Estadão**	**Folha de S.P.**	**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

Our dataset of legitimate news in Portuguese is composed by a total of 207,914 articles collected from two of the major mainstream news sites in Brazil: *Estadão* and *Folha de São Paulo*. The news are dated from 2014 to 2017. We have built a Web crawler to collect the news automatically from these news sites. The crawler was developed with the ability to identify the news in four major domains: Politics, Sports, Economy and Culture. Table 1 shows the distribution of topics present in legitimate news.

Regarding fake news, the dataset comprises fake news that were disseminated in Brazil from 2010 to 2017. We have collected these news from two fact-checking services, namely: *e-Farsas* and *Boatos*. These services keep track of some of the most shared fake news articles that circulate in the Web, providing verifiable evidence of falsity. An example is the fake news tracked by e-Farsas saying that the hitherto presidential candidate Bolsonaro (current president of Brazil) requested a rifle to retaliate outlaws in a slum area.[m] A single post in Facebook, sharing this news, had around 27,000 shares and more than 20,000 comments. We collected a total of 121 fact-checked fake news from more than 40 different news sources. Although small, this data set has two interesting properties: (i) it contains highly shared fake news, which means that they potentially deceived lots of people; and (ii) the fake news come from highly diverse sources with possibly different characteristics. These two properties make the problem even more challenging, i.e., the fake news were believed to be true by a lot of people and just knowing the characteristics of a few fake news dissemination sources should not be of much help. It is also important to remark that in the real world the number of existing fact-checked fake news, in comparison to the number of legitimate ones, is intrinsically orders of magnitude smaller, so we are trying to reproduce the real world with all its intrinsic challenges.

The legitimate news in English were collected from the Kaggle dataset entitled "All the News"[n]between the years of 2016 and 2017. We collected news from *The Guardian* (1798 articles), from *New York Times* (1598 articles) and 2598 articles from *CNN*. The English fake news is composed by political fake articles from *Snopes*[o](103 articles from fake and mostly-fake categories), political fake news from [5] (75 articles), the top fake news collected by Buzzfeed[p] (41 articles) and the dataset extracted from the BSDetector[q](1425 articles).

4.2. *Subjectivity Distribution*

[m]http://www.e-farsas.com/bolsonaro-pediu-um-fuzil-para-revidar-contra-os-bandidos-ao-ser-recebido-tiros-em-cidade-de-deus.html
[n]https://www.kaggle.com/snapcrack/all-the-news/version/4
[o]https://github.com/sfu-discourse-lab/Misinformation_detection
[p]https://github.com/BuzzFeedNews/2017-12-fake-news-top-50
[q]https://www.kaggle.com/mrisdal/fake-news

Table 2 and Table 3 show the descriptive statistics of the semantic distances reported by WMD for the five lexicons in Portuguese, considering the fake and legitimate news, respectively. In each column, the statistics of the semantic distances between the sentences of the documents and each of the five lexicons used are presented.

Table 2. Descriptive statistics of semantic distances for fake news in Portuguese

	Arg	**Pre**	**Sen**	**Val**	**Mod**
n sentences	1,262	1,262	1,262	1,262	1,262
mdia	0.872157	0.865573	0.866132	0.869331	0.874076
std	0.014562	0.011595	0.009140	0.011721	0.012182
min	0.811058	0.796779	0.816884	0.822704	0.808073
25%	0.863432	0.860071	0.860033	0.862856	0.867496
50%	0.873016	0.866392	0.865291	0.869716	0.875169
75%	0.881896	0.872832	0.872063	0.877078	0.881241
max	0.907376	0.894934	0.895981	0.900131	0.902114

Table 3. Descriptive statistics of semantic distances for legitimate news in Portuguese

	Arg	**Pre**	**Sen**	**Val**	**Mod**
n sentences	1,048,576	1,048,576	1,048,576	1,048,576	1,048,576
mdia	0.873015	0.865627	0.86633	0.869616	0.874199
std	0.0133645	0.0108445	0.00889849	0.0109894	0.0112677
min	0.775375	0.764377	0.809926	0.779943	0.776655
25%	0.864464	0.860394	0.860672	0.863438	0.867933
50%	0.87357	0.866169	0.86578	0.870071	0.874846
75%	0.882205	0.872328	0.871784	0.876617	0.881418
max	0.927491	0.916601	0.910878	0.919109	0.927155

From the Table 2 and Table 3, it is possible to observe that, for the case of legitimate and fake news in Portuguese, the values for the statistics presented for the two datasets are quite similar, with little difference for the two sets. This similarity can be seen in Figures 1 and Figure 2, which present boxplots for both datasets.

Table 4 presents a summary of descriptive statistics related to the semantic distances for the fake news dataset in English, totaling 161,400 sentences. Similarly, Table 5 presents the summary for legitimate news in English, totaling 577,700 sentences from legitimate news. In these tables, when comparing the averages of each of the lexicons, we can observe that the fake news sentences tend to have a lower distance values when compared to the legitimate news sentences. This demonstrates a tendency of a higher level of subjectivity related to the fake news, when compared to the legitimate news. The only exception is the lexicon "hedges", in which the fake news presented higher values (less subjectivity) than the legitimate news. Figures 3 and Figure 4 show the distribution of data through boxplots. When analyzing the medians defined in the boxplots, it is possible to observe the subtle differences in each lexicon, for both datasets, where it is possible to notice that the medians of the fake news tend to be smaller than the medians of the legitimate news, demonstrating that the fake news sentences are semantically more similar to the subjectivity lexicons.

5. Results and Discussions

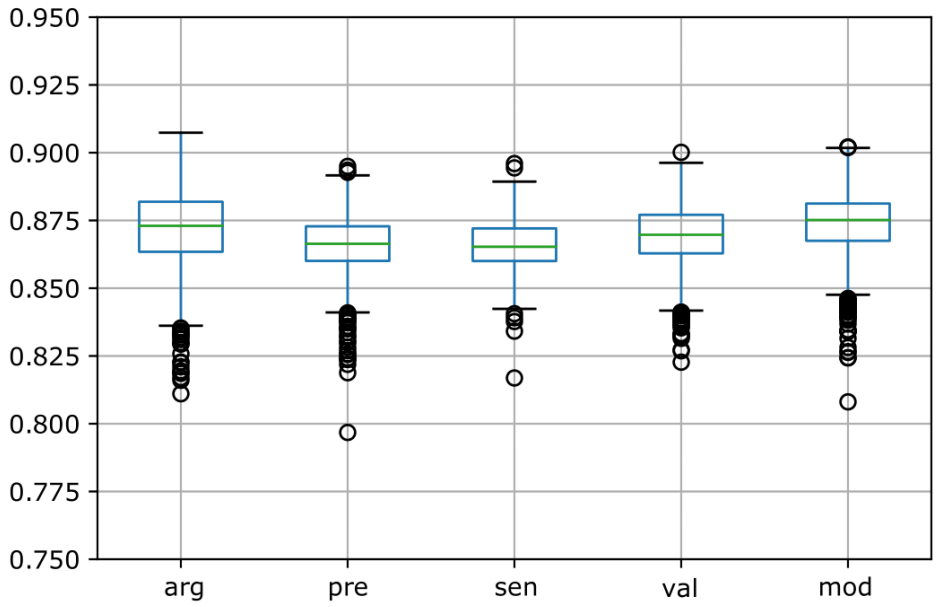

Fig. 1. Boxplot for the semantic distances of fake news in Portuguese.

Table 4. Descriptive statistics of semantic distances for fake news in English.

	assertives	factives	hedges	implicatives	negative	positive	report	bias	positive_gold	negative_gold
n sentenas	161,400	161,400	161,400	161,400	161,400	161,400	161,400	161,400	161,400	161,400
mdia	0.848723	0.856772	0.840404	0.852624	0.795093	0.811316	0.814766	0.800111	0.801980	0.782325
std	0.017431	0.018439	0.017629	0.017137	0.021775	0.015803	0.013720	0.018127	0.043943	0.046436
min	0.743138	0.714145	0.730930	0.739973	0.704506	0.729462	0.732502	0.690459	0.630992	0.581031
25%	0.839179	0.847457	0.830128	0.843336	0.780995	0.802083	0.806709	0.789052	0.780129	0.761946
50%	0.848435	0.856724	0.840212	0.852669	0.796023	0.811782	0.813971	0.800781	0.808245	0.789917
75%	0.858291	0.867225	0.850557	0.862148	0.809183	0.820992	0.822361	0.811519	0.832919	0.815393
max	0.958710	0.966396	0.941242	0.950850	0.876647	0.879597	0.909514	0.887666	0.912833	0.881969

In this section, we will describe the experiments performed in order to answer the four research questions previously described.

5.1. *Statistical Analysis*

Table 6 shows the statistical analysis comparing the values of the semantic distances between the sentences of the fake and legitimate news. In the analysis, the Mann-Whitney hypothesis test is performed, which consists of a non-parametric hypothesis test involving two independent samples, being robust in scenarios for samples with different sizes [29]. Even so, due to the large difference in size between the fake (1,262 sentences) and legitimate (1,048,576 sentences) news datasets, for this analysis, a stratified approach was also conducted, where 50 repeated tests were performed, where in each repetition, a sample of 1,262 distances of fake and legitimate news sentences were randomly selected to perform the hypothesis test. The value of 1,262 was established because it is the number of sentences in the fake news dataset,

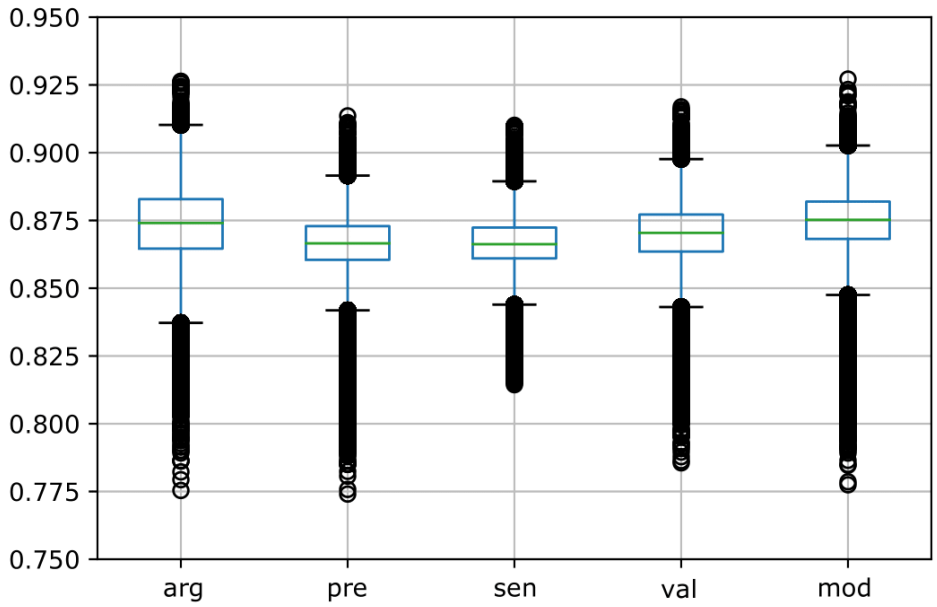

Fig. 2. Boxplot for the semantic distances of legitimate news in Portuguese..

Table 5. Descriptive statistics of semantic distances for legitimate news in English.

	assertives	factives	hedges	implicatives	negative	positive	report	bias	positive_gold	negative_gold
n sentenas	577,700	577,700	577,700	577,700	577,700	577,700	577,700	577,700	577,700	577,700
mdia	0.848964	0.857577	0.838923	0.853414	0.800162	0.813679	0.815159	0.80309	0.810614	0.792114
std	0.0164179	0.0175222	0.0170778	0.0160289	0.0209407	0.0160071	0.0134946	0.0164412	0.0389415	0.0402623
min	0.725402	0.697192	0.715651	0.726747	0.711466	0.731334	0.731048	0.709798	0.62922	0.578198
25%	0.840313	0.849224	0.830369	0.845632	0.788118	0.805208	0.807188	0.794083	0.791524	0.773882
50%	0.848829	0.857694	0.838786	0.853366	0.800783	0.813881	0.814977	0.802924	0.815124	0.796762
75%	0.857466	0.866482	0.847478	0.861681	0.813147	0.822689	0.822715	0.812097	0.836303	0.819765
max	0.958709	0.966396	0.941241	0.959407	0.89213	0.900637	0.909513	0.894564	0.920157	0.895628

thus allowing tests with the same number of samples for both datasets. Thus, the first column of Table 6 shows the number of repetitions, for each lexicon, where the null hypothesis of the test is rejected (p-value <0.05). The null hypothesis assumes the same distribution for the distances of fake and legitimate news. In this first column, it is possible to visualize that the lexicon of sentiments presented the highest amount of rejections for H0, obtaining 38 rejections among the 50 repetitions performed. In the second column, a standard execution is presented, without stratification, where the distances of the sentences from the fake and legitimate news are conventionally tested, considering that the alternative hypothesis represents that the two samples come from different distributions (*two-sided*). In this column, we show the values where the p-value is less than 0.05. In this case, it can be seen that the result shows that there are significant differences between the sentences of fake and legitimate news for the lexicons of Argumentation (p-value = 0.009) and Sentiment (p-value = 0.001). Now, knowing that there are significant differences for two lexicons, it is imperative to know, in a

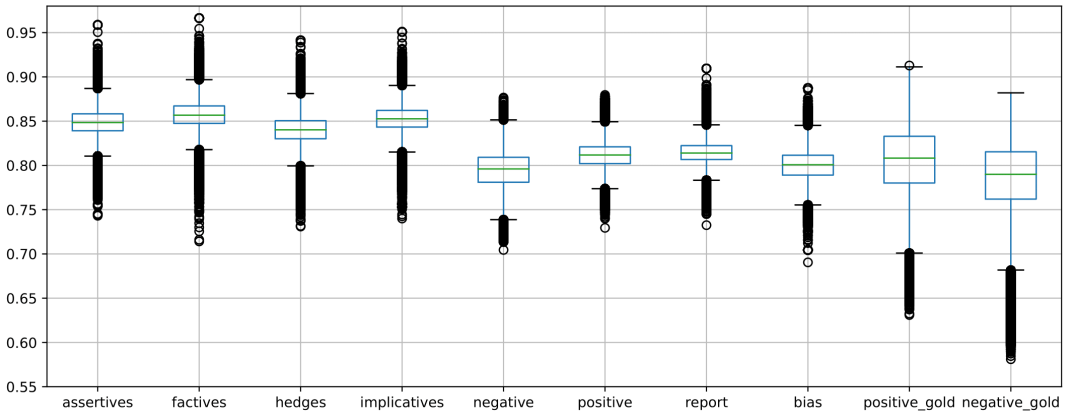

Fig. 3. Boxplot for the semantic distances of fake news in English.

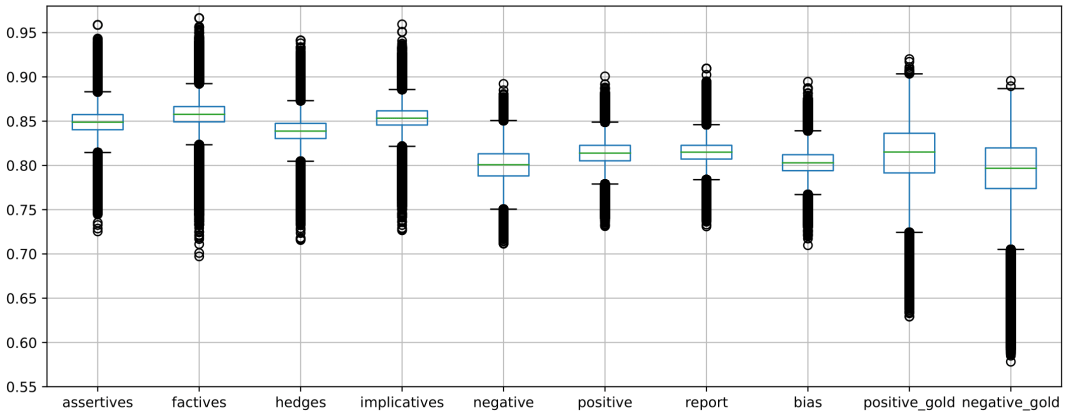

Fig. 4. Boxplot for the semantic distances of legitimate news in English.

significant way, the type of these differences. For example, it is important to know whether the fake news sentences are more subjective (smaller distances to the subjectivity lexicons) than the distances from the legitimate news sentences.

For such an analysis, the third column of Table 6 presents the same test presented in the second column, however, the alternative hypothesis is modified to consider that the distances of the fake news sentences are greater (less subjective) than the distances of the legitimate news. In this case, we can see that this alternative hypothesis could not be verified for any of the five lexicons. The fourth column, on the other hand, presents the opposite, and the alternative hypothesis is accepted when the distances of the fake news are shorter (more subjective) than the legitimate news. For this scenario, corroborating with the evaluation presented in the previous columns of the table, it can be seen that for the lexicons of argumentation and sentiment, the null hypothesis could be rejected, meaning that the fake news sentences are statistically more subjective than the legitimate ones. In practice, these results can indicate

Table 6. Hypothesis tests comparing the semantic distances of the sentences for each lexicon, considering the fake and legitimate news in Portuguese. The results show the number of tests in which a significant difference (p-value <0.05) were reported between the semantic distances of fake and legitimate news (column # H0 rejected). The second column (H1 two-sided) shows the results (significant p-values) of a traditional test (not stratified), comparing the distances of fake and legitimate sentences, where the alternative hypothesis is to consider that both samples belong to different distributions. In the third column (H1 (Fake >Legitimate)) the same evaluation is performed, but now the alternative hypothesis (H1) consists of stating that the values of the semantic distances of the fake news sentences are greater (less subjective) than the legitimate news distances. The fourth column (H1 (Fake <Legitimate)) performs the same test, however, considering that the alternative hypothesis considers that the semantic distances of the fake news sentences are smaller (more subjective) than the distances of the legitimate news.

	# H0 rejected	H1 two-sided	H1 (Fake >Legitimate)	H1 (Fake <Legitimate)
Argumentation	22	0.009	-	0.004
Presupposition	1	-	-	-
Sentiment	38	0.001	-	0.0005
Valuation	11	-	-	-
Modalization	3	-	-	-

that the fake news, in semantic terms, seem to be both more argumentative and emotional, when compared to the legitimate news. This results allow us to answer the questions **Q1** and **Q2** positively for only to of the five lexicons used in Portuguese.

Now, considering the same evaluation for English news, the Table 7 shows the hypothesis tests for the English dataset, following the same methodology as the hypothesis tests performed for Portuguese evaluations. In this table, it is possible to verify that, for all 50 randomized tests performed, it was possible to verify significant differences for all lexicons used. It is also possible to notice when analyzing the fourth column, that for all the subjectivity lexicons (with the exception of the subjectivity lexicon "hedges"), the semantic distances of the fake sentences were statistically smaller than those of the legitimate news. This contributes to the hypothesis that fake news seems to be more subjective than legitimate news, even for the English scenario. In this table, too small p-values were rounded, showing only 0.00. These subtle but yet significant differences can be seen in the presented boxplots, where the medians of fake news (Figure 3) are lower than the medians of legitimate news (Figure 4) , demonstrating that fake news seems to be more subjective than legitimate news. This result also allows us to answer questions **Q1** and **Q2** arguing that it is possible to find significant differences between fake and legitimate news in terms of subjectivity. One hypothesis that we envision for the "Hedges" lexicon case is that, possibly, writers and journalists from major media outlets, when trying to maintain a certain level of impartiality in the documents, make use of these terms in order to reduce the level of commitment to facts not yet fully clarified. In opposite, the authors of fake news would tend to reproduce untrue facts as being completely true, avoiding the use of this linguistic resource.

5.2. *Classification Results*

In this section, we describe the experimental setup of the fake news classification experiments, the main results and discussions.

5.2.1. *Experimental Setup*

Table 7. Hypothesis tests comparing the semantic distances of the sentences for each lexicon, considering the fake and legitimate news in English. In the results, it is possible to observe that for all the lexicons used, there were significant differences (p-value <0.05). For these cases, most tests showed that fake news had shorter semantic distances than legitimate news, which demonstrates a greater semantic similarity between fake news and subjectivity lexicons, which denotes a greater subjectivity of fake news.

	# H0 rejected	H1 two-sided	H1 (Fake >Legitimate)	H1 (Fake <Legitimate)
assertives	50	0.00	-	0.00
factives	50	0.00	-	0.00
hedges	50	0.00	0.00	-
implicatives	50	0.00	-	0.00
negative	50	0.00	-	0.00
positive	50	0.00	-	0.00
report	50	0.00	-	0.00
bias	50	0.00	-	0.00
positive_gold	50	0.00	-	0.00
negative_gold	50	0.00	-	0.00

As classification models, we have used XGBoost and Random Forests, which are well known for their strong predictive power and for providing state-of-the-art performance in a wide range of complex domains. We have used these classifiers in two settings: (i) with our proposed subjectivity vectors as input features, and (ii) with classic BoW/TF-IDF representations. The latter resembles many related works that rely only on content representation of words for fake news detection.

We used the scikit-learn Machine Learning library (v0.19)[r]for training these classifiers. We used the default settings of the algorithms since our main goal is to compare the aforementioned classification settings (i) and (ii) on an equal footing. We also exclude sentences with less than 3 words and documents with more than 100 words. For the models based on TFIDF, we also exclude terms that appears in less than 1% of documents. We also replicate a scenario described in [30] where the authors performed a study on fake news during the 2016 US presidential election reporting a proportion of one fake news for each four legitimate news spread by some Facebook pages. We followed this same proportion in our experiments generating several samples, through bootstraping, where the class distributions followed this proportion. For the classifications, we perform an evaluation using a random sampling approach, reporting an average result based on 100 repetitions. In each repetition, we consider splits of 70%-30% for training and testing based on the number of fake news in Portuguese, since it is the shortest dataset. Hence, in each train/test evaluation, since we have 121 fake news in Portuguese, we use 85 (fake news) and 340 (legitimate news) for training and 36 (fake news) and 144 (legitimate news) for testing. This setup respects the 4:1 distribution between legitimate and fake news, and it is also applied to the English experiments.

As evaluation metrics, we use the Precision, Recall, F-measure and PR-AUC. The PR-AUC metric represents the area under the curve (AUC) considering Precision and Recall. This metric has the advantage of considering the performance of the models when varying their classification thresholds, giving a more comprehensive view of the models. The classic metrics Precision, Recall and F-measure evaluate the models in a more restricted way, using their default classification threshold. In this research, the four evaluation metrics will be

[r]http://scikit-learn.org/stable/

reported, however, the PR-AUC metric will be used as the main evaluation metric.

To calculate the semantic distances using WMD, a model was generated implementing word embeddings from a Wikipedia dump for Portuguese. For the experiments in English, we use the pre-trained model called Google News embeddings [s].

5.2.2. *Results for Portuguese News*

Table 8 shows the average results in terms of PR-AUC, Precision, Recall and F1-Score for the classification of fake news using the Average Subjectivity per Document (ASD), considering 100 repetitions for each model. In this scenario, each document is represented as a vector of five dimensions, where each dimension represents the average distance of the document's sentences in relation to each of the five subjectivity lexicons used in Portuguese. These results are worse than those originally presented in [11]. This difference is due to improvements in the sentence tokenization process implemented in this work, allowing a more precise sentence tokenization. The improvements in this pre-processing step revealed an important concern regarding the usage of the WMD, which is the correlation between the size of the text and the reported distances. Basically, the WMD tends to report short distances for longer text snippets. So, minor issues in sentence tokenization could inject some noise in the evaluations.

Table 8. Average results of the fake and legitimate news classification using the ASD features for the Portuguese language.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.27(±0.04)	0.13(±0.07)	0.30(±0.15)	0.08(±0.05)
RF	0.26(±0.04)	0.15(±0.06)	0.32(±0.13)	0.10(±0.05)

Table 9. Average results of the fake and legitimate news classification using the VSS features for the Portuguese language.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.30(±0.04)	0.13(±0.06)	0.36(±0.17)	0.08(±0.04)
RF	0.26(±0.04)	0.14(±0.07)	0.32(±0.13)	0.09(±0.05)

The Table 9, shows the results for the same classification, however, using the Vector Subjectivity per Sentence (VSS). Comparing both results, we can note that the models using VSS tend to show a better classification results. This is probably due to the highest complexity of such features, that keep the entire information about the sentence's subjectivity, instead of lose it by the usage of the simple averages. This result can indicate that improvements in the feature engineering process can bring improvements in the classification.

Considering the models based on BoW/TFIDF, the Table 10, shows the results for such models. In the results, we can note a higher performances compared to models based on subjectivity features. Due to the ability to represent specific terms in the documents, these models are able to generate a representation completely based on the occurrence of words in the documents, which in certain cases can be useful, as in problems of information retrieval. However, when used in textual classifications, these models can generate biased representations, as they are based on the occurrence of specific terms in data. To demonstrate this

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

hypothesis, experiments were conducted in order to take advantage of the variability in the legitimate news in Portuguese, which are subdivided into four different subjects (i.e. Sport, Politics, Economy and Culture). For this scenario, the experiments were performed in a design called "cross-domain", where the models are trained using legitimate news on a given topic, for example Culture. For the test set, the legitimate news are used from different topics, such as, for example, Politics. In this way, it is possible to evaluate the classification of legitimate and fake news considering variations in the legitimate news domain.

Table 10. Average results of the fake and legitimate news classification using the Bow/TFIDF features for the Portuguese language.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.68(±0.07)	0.47(±0.09)	0.81(±0.09)	0.33(±0.08)
RF	0.51(±0.06)	0.27(±0.09)	0.76(±0.14)	0.17(±0.07)

In the first experiment using the "cross-domain" approach, the Table 11 presents the average results of the evaluations using the VSS features. It is possible to observe in the results that there are no significant variations in terms of PR-AUC. In fact, it was also possible so see some improvements in terms of F1-Score and Recall, when compared to the results without the cross-domain evaluation (Table 9).

Table 11. Average results of the fake and legitimate news classification using the VSS features for the Portuguese language using the cross-domain approach.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.30(±0.05)	0.18(±0.08)	0.32(±0.12)	0.14(±0.07)
RF	0.25(±0.03)	0.17(±0.07)	0.29(±0.11)	0.13(±0.06)

Now, considering the models based on BoW/TFIDF, the Table 12 shows the same cross-domain experiment, but considering models based on TFIDF features. In these results, although they are still superior to the models based on subjectivity, it can be noted a significant decrease in performance, when compared to the same models without the cross-domain evaluation (Table 10). In terms of PR-AUC, for the XGBoost model, there was a reduction of approximately 32% in performance using cross-domain. For Random Forest, the reduction in PR-AUC was 25% in the average result. These results suggest that the classical text classification models based on Bow/TFIDF may suffer from a strong bias present in the dataset to which they are trained. A similar reduction in performance could not be seen in the models based on subjectivity, denoting that this models could generalize better.

Table 12. Average results of the fake and legitimate news classification using the BoW/TFIDF features for the Portuguese language using the cross-domain approach.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.46(±0.15)	0.42(±0.10)	0.40(±0.17)	0.57(±0.19)
RF	0.38(±0.13)	0.33(±0.12)	0.47(±0.22)	0.31(±0.13)

With the presented results, since the performance of the classical models based on BoW/TFIDF were significantly better than the models based on subjectivity, we answer the **Q3** in the direction that the models based on the proposed features could not achieve better results than the BoW/TFIDF models for the Portuguese language. However, in respect to the **Q4**, we

show that the models based on subjectivity seems to achieve a better generalization, since the models based on VSS features using XGBoost did not show a significant reduction in performance using the cross-domain evaluation. In contrast, the models based on BoW/TFIDF showed reduction in performance of more than 30%.

In fact, we can demonstrate the limitations of the models based on lexical features such as BoW/TFIDF by trying to find possible explanations regarding the classification models. To do this, we use the SHAP library [31]. SHAP is based on the *Shapley values*, which are values that represent the importance of each feature for the correct prediction of the model. For example, features that have a significant impact the classification of a model, are considered relevant features. This analysis allows to obtain a more objective understanding of the classification decisions of the implemented models, generating insights about the classification problem.

For this evaluation, in order to compare the explanations provided by the models using subjectivity features and those using BoW/TFIDF features, we use the XGBoost model, since it was the one that showed the best overall results. For the analysis using the proposed subjectivity features, we choose the model using the ASD features. This choice was made in order to improve the visualizations provided by the SHAP library, since the model using ASD have only five features to analyze. For both models using ASD and BoW/TFIDF, we selected the best models found in the random sampling executions.

The Figure 5 shows the summary plot for the model using the ASD features. In the image, the y-axis represents the five subjectivity features used in order of relevance for the classification. On the x-axis, we have the range of values that represent the SHAP values, which express the weight that each feature exerts to determine the classification of a sample. The positive values represent a greater weight for the classification of the target class (class 1), which in this case, consists of fake news classification. The negative values represents a greater tendency for the classification of a legitimate news (class 0). Therefore, points displaced more to the right mean a higher chance for the model to classify the sample as a fake news, considering a particular feature. Each point on the graph represent a sample, in this case, a news. The color of the point represents the numerical value of a given feature.

It can be seen in Figure 5, that the lexicon of sentiment (sen_avg) was the one that presented the greatest relevance for the classification choices of this particular model, given that this feature is at the top of the y axis. Additionally, we can see that, even in this feature, there is a prevalence of blue dots positively correlated with values greater than zero in the x axis. This means that shorter semantic distances relative to the lexicon of sentiment are correlated with fake news predictions. Similar behavior can also be observed for the argumentation lexicon (arg_avg), showing a trend that fake news seems to adopt a language that is both more emotional and argumentative when compared to legitimate news.

The Figure 6, on the other hand, shows the same analysis, but now using the model based on BoW/TFIDF features. For this model, the features are stemmed words present in the vocabulary of the news used in the training process. Despite that this model shows a better classification result, we argue that the most relevant features do not properly address the problematic of fake news, but they just reveals aspects of the subject of the documents. This hypothesis could explain the severe performance reduction in the cross-domain evaluation. For example, the top relevant feature of this model using BoW/TFIDF features is the word "ano" (year). The occurrence of this term in documents is represented by the red dots, since

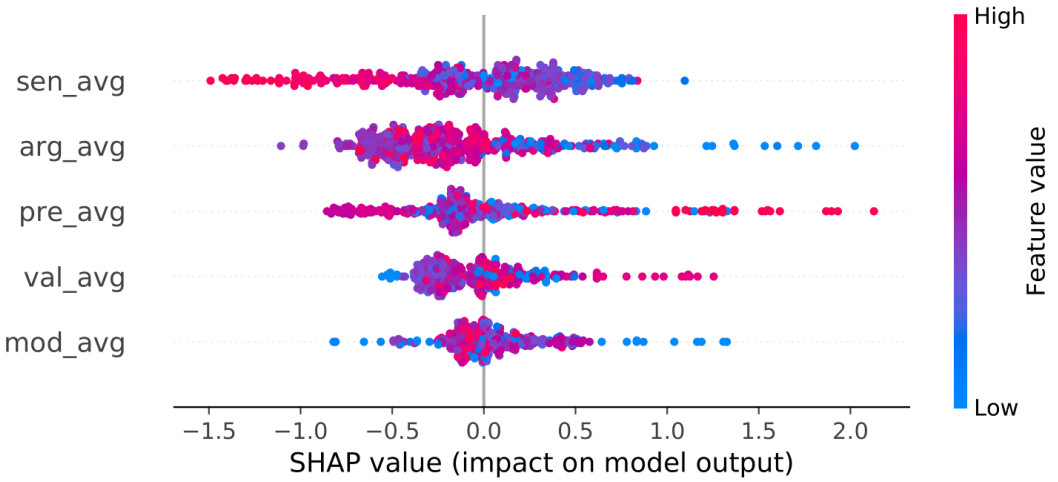

Fig. 5. Summary plot generated through SHAP, showing the weight that the features exert on the model classification decision. On the y-axis, the five subjectivity features that form the vector representation of a document are listed, in order of importance. On the x-axis we see the shap values, where values greater than zero represent a greater chance for the classification of the target class (class 1), which in this case is fake news. Negative values (less than zero) represent a greater chance for the classification of legitimate news (class 0).

the TFIDF weight when a term is present in document is always grater than zero, resulting in red dots in the Figure. We can see that this term is totally related to legitimate news (shap values less than zero in x-axis). Other examples are the terms "lul" (Lula) and "bolsonar" (Bolsonaro), both related to political personalities in Brazil. According to the SHAP values, both terms are related fake news classifications.

5.2.3. *Results for English News*

Table 13 presents the average results of the classification of fake news for the English dataset, using the ASD features. At first, it is possible to notice that the results presented are systematically better when compared to the results reported for the news in Portuguese, presented in the Table 8. This result indicates that, possibly, both the lexicons used for the news in English and the embeddings used seem to be better suited for the representation of news documents. It is important to note that the lexicons adopted in these experiments have twice dimensions of subjectivity in relation to the lexicons in Portuguese, which have only five dimensions. Also, the embeddings used for the experiments in English are made from news documents, while the embeddings used in Portuguese come from a Wikipedia dump. These differences can improve the news representation in the English language, thus improving the results.

Table 14 shows the results of the classifications using the VSS features. As observed for the results in Portuguese, the use of VSS features also seems to improve the classification results, especially for the XGBoost model. This demonstrates that a more complex feature engineering process can bring good results for this type of features.

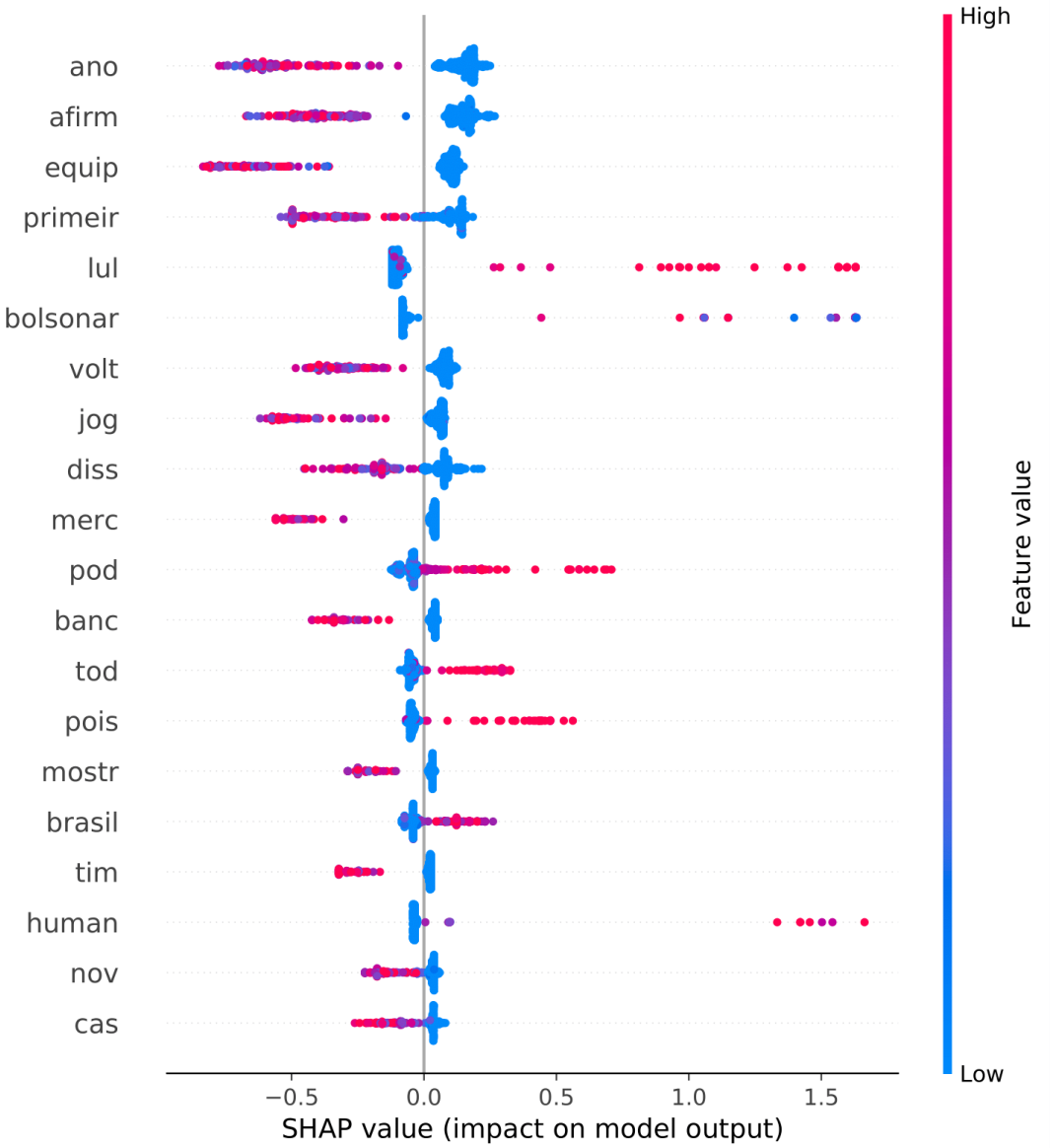

Fig. 6. Summary plot generated through SHAP showing the most relevant features for the classification of the model based on BoW and TFIDF in the Portuguese language. Since the features are TFIDF weights, the red dots means that the term is occurring in a document (TFIDF grater than zero).

Considering the models using the BoW/TFIDF features in English news, the Table 15 shows the results for the classification models using such features. Again, this scenario is the one with the best results. However, it is important to highlight that these models tend to suffer with the same issue observed in the previous analysis using SHAP for the Portuguese language. We can observe this in the Figure 7, where the top two features for the classification

Table 13. Average results of the fake and legitimate news classification using the ASD features for the English language.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.43(±0.06)	0.33(±0.08)	0.58(±0.11)	0.23(±0.07)
RF	0.40(±0.06)	0.32(±0.08)	0.57(±0.12)	0.23(±0.07)

Table 14. Average results of the fake and legitimate news classification using the VSS features for the English language.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.50(±0.07)	0.36(±0.10)	0.63(±0.12)	0.25(±0.09)
RF	0.38(±0.06)	0.32(±0.09)	0.55(±0.12)	0.23(±0.08)

of the best model based on BoW/TFIDF are "donald" and "cnn". The former term is related to fake news classifications, and the last one is totally related to legitimate news classifications.

Based on the classification experiments, we answer the **Q3** in the same way we did in the Portuguese experiments, reporting that the models based on the proposed subjectivity features are not yet ready to outperform the BoW/TFIDF models. Since the legitimate news from the English dataset used do not have distinctions regarding the news subject, we cannot perform the cross-domain in the same design used with the Portuguese news, meaning that we cannot answer the **Q4** using the English dataset at this moment. This must be addressed in future works.

6. Conclusions

This work has the main objective of investigating how textual subjectivity based on lexicons can help in the understanding and also in the identification of fake news. Initially, this research proposes the application of subjectivity lexicons and semantic distances as a way to extract subjectivity from the news. The semantic distances proved to be efficient for the extraction of subjectivity in news documents, considering two different languages, that are the Brazilian Portuguese and English. We also propose two different ways of using the subjectivity features, and we demonstrate that the VSS features achieved better results.

Another important finding is that the classical models based on BoW/TFIDF seem to be biased in the context of the documents that they are trained, reporting high classification results, but not learning nuances that truly distinguish fake and legitimate news. In such direction, the usage of models based on subjectivity can be a reliable alternative.

Future work include the improvement of the proposed models, by applying different transformations on features, and also using Deep Learning models. We also plan to explore different approaches in subjectivity extraction.

Table 15. Average results of the fake and legitimate news classification using the BoW/TFIDF features for the English language.

	PR-AUC	**F1**	**Precision**	**Recall**
Xgboost	0.81(±0.04)	0.70(±0.05)	0.82(±0.06)	0.61(±0.07)
RF	0.60(±0.07)	0.46(±0.10)	0.84(±0.10)	0.32(±0.09)

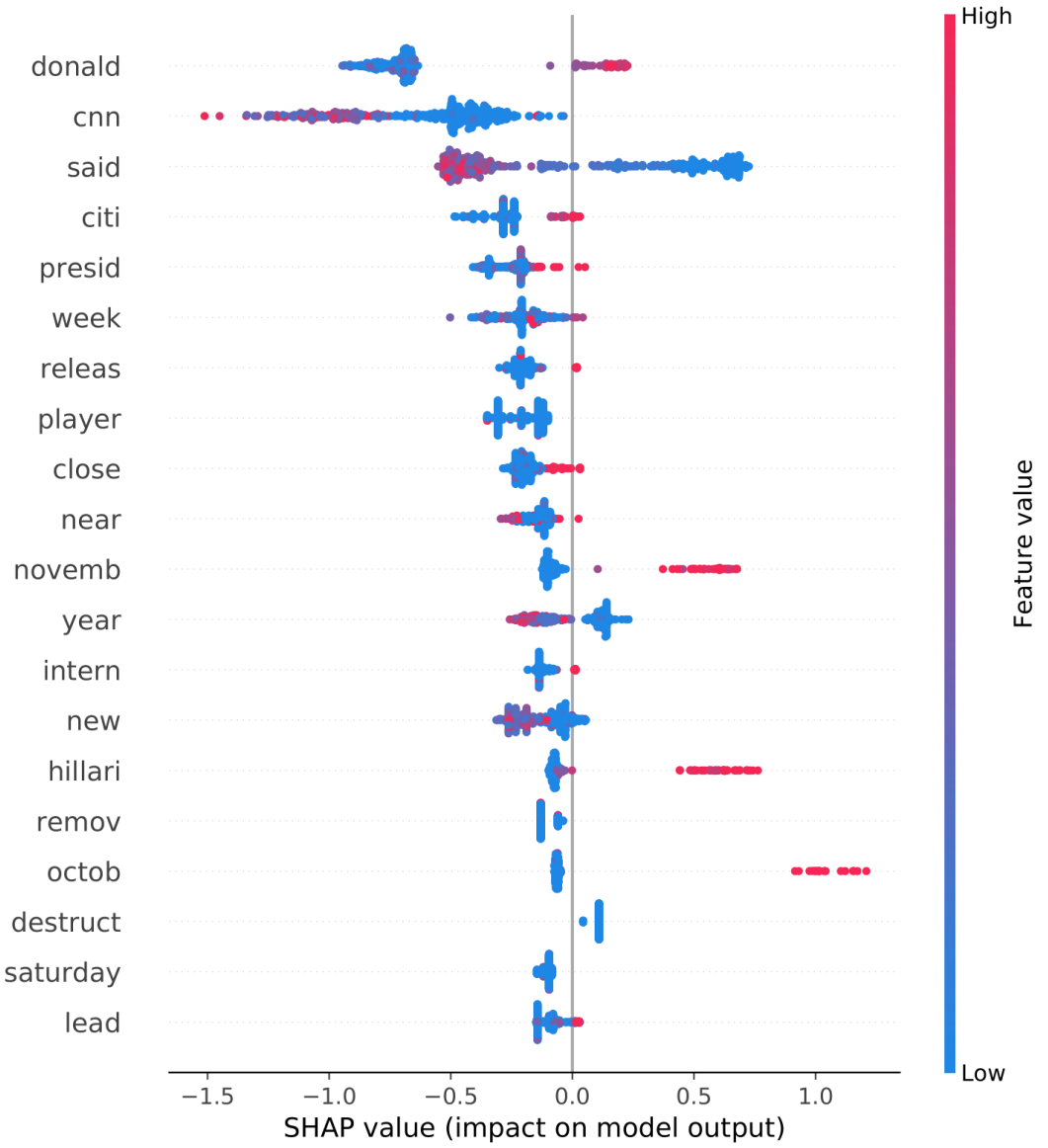

Fig. 7. Summary plot generated through SHAP showing the most relevant features for the classification of the model based on BoW and TFIDF in the English language.

References

1. Changjun Lee and Jieun Shin and Ahreum Hong (2018), *Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea*, Telematics and Informatics, Vol. 35, pp. 245 - 254.
2. Hunt Allcott and Matthew Gentzkow (2017), *Social Media and Fake News in the 2016 Election*, Journal of Economic Perspectives, vol 31(2), pages 211-236.
3. Regina Marchi (2012), *With Facebook, blogs, and fake news, teens reject journalistic "objectivity"*,

Journal of Communication Inquiry, vol 36, pages 246-262.
4. Ferrara, Emilio and Varol, Onur and Davis, Clayton and Menczer, Filippo and Flammini, Alessandro (2016), *The Rise of Social Bots*, Commun. ACM, vol 59, pages 96-104.
5. Benjamin Horne and Sibel Adali (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*, International AAAI Conference on Web and Social Media.
6. Peter Bourgonje, Julian Moreno Schneider, Georg Rehm (2017), *From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles*, Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism.
7. Hadeer Ahmed, Issa Traore, Sherif Saad (2017), *Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques*, Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pages 127-138.
8. Wiebe, Janyce and Wilson, Theresa and Bruce, Rebecca and Bell, Matthew and Martin, Melanie (2004), *Learning Subjective Language*, Computational Linguistics, Vol. 30, pages 277-308.
9. Mihalcea, Rada and Banea, Carmen and Wiebe, Janyce (2007), *Learning Multilingual Subjective Language via Cross-Lingual Projections*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 976–983.
10. Evelin Amorim, Marcia Cancado, Adriano Veloso (2018), *Automated Essay Scoring in the Presence of Biased Ratings*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 229–237.
11. Caio Jeronimo and Leandro Marinho and Claudio Campelo and Adriano Veloso and Allan Sales (2019), *Fake News Classification Based on Subjective Language*, Proceedings of the 21st International Conference on Information Integration and Web-based Applications and Services - iiWas.
12. Shu, Kai and Sliva, Amy and Wang, Suhang and Tang, Jiliang and Liu, Huan (2017), *Fake News Detection on Social Media: A Data Mining Perspective*, SIGKDD Explor. Newsl., Vol 19, pages 22–36.
13. Wang, William Yang (2017), *Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection*, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 422–426.
14. Potthast, Martin and Kiesel, Johannes and Reinartz, Kevin and Bevendorff, Janek and Stein, Benno (2018), *A Stylometric Inquiry into Hyperpartisan and Fake News*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 231–240.
15. Tanushree Mitra and Eric Gilbert (2015), *CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations*, International AAAI Conference on Web and Social Media.
16. Janze, Christian and Risius, Marten (2017), *Automatic Detection of Fake News on Social Media Platforms*, Pacific Asia Conference on Information Systems PACIS.
17. Tacchini, Eugenio and Ballarin, Gabriele and Della Vedova, Marco L and Moret, Stefano and de Alfaro, Luca (2017), *Some like it hoax: Automated fake news detection in social networks*, arXiv preprint arXiv:1704.07506.
18. Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea (2018), *Automatic Detection of Fake News*, Proceedings of the 27th International Conference on Computational Linguistics.
19. Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, Oto A. Vale (2018), *Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results*, Computational Processing of the Portuguese Language, pages 324–334.
20. Rashkin, Hannah and Choi, Eunsol and Jang, Jin Yea and Volkova, Svitlana and Choi, Yejin (2017), *Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking*, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.
21. S. B. Parikh and P. K. Atrey (2018), *Media-Rich Fake News Detection: A Survey*, IEEE Conference

on Multimedia Information Processing and Retrieval (MIPR), pages 436-441.
22. Recasens, Marta and Danescu-Niculescu-Mizil, Cristian and Jurafsky, Dan (2013), *Linguistic Models for Analyzing and Detecting Biased Language*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1650–1659.
23. Wilson, Theresa and Wiebe, Janyce and Hoffmann, Paul (2005), *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 347354.
24. Choi, Yoonjung and Deng, Lingjia and Wiebe, Janyce (2014), *Lexical Acquisition for Opinion Inference: A Sense-Level Lexicon of Benefactive and Malefactive Events*, Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 107–112.
25. Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff (2013), *Distributed Representations of Words and Phrases and their Compositionality*, Advances in Neural Information Processing Systems 26, pages 3111–3119.
26. Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey (2013), *Efficient Estimation of Word Representations in Vector Space*, CoRR.
27. Kusner, Matt J. and Sun, Yu and Kolkin, Nicholas I. and Weinberger, Kilian Q. (2015), *From Word Embeddings to Document Distances*, Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, pages 957–966.
28. Huang, Gao and Quo, Chuan and Kusner, Matt J. and Sun, Yu and Weinberger, Kilian Q. and Sha, Fei (2016), *Supervised Word Mover's Distance*, Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 4869–4877.
29. Mann, H. B. and Whitney, D. R. (1947), *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other*, Annals of Mathematical Statistics, Vol. 18, pages 50–60.
30. Silverman, Craig (2016), *This analysis shows how viral fake election news stories outperformed real news on Facebook*, BuzzFeed news, Vol. 16.
31. Lundberg, Scott M and Lee, Su-In (2017), *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems 30, pages 4765–4774.