A new approach for measuring subjectivity in Brazilian news

D. F. Lima and A. S. C. Melo and D. L. Carvalho and L. B. Marinho

Federal University of Campina Grande, Brazil
{diogoflorencio, allanmelo}@copin.ufcg.edu.br, daniella.carvalho@ccc.ufcg.edu.br,
lbmarinho@dsc.ufcg.edu.br

Abstract. With the advent of digital journalism, information democratization has become a reality since news articles are published as soon as the facts occur, and that they are accessible from any device connected to the internet. It is common sense the perception that some media outlets are more biased than others when it comes to the way of exposing the facts. However, automatic ways of measuring such biases is still an open research challenge. Under the assumption that journalistic texts must have an objective and impartial language, high levels of subjectivity in these texts may indicate bias. This paper proposes an initial analysis on the usage of subjectivity lexicons to characterize subjectivity in seven popular media outlets in Brazil. To better understand the obtained results, we carried out a correlation analysis between the levels of subjectivity, readability, and news popularity metrics. The adopted methods, along with the findings obtained from this research, may contribute to a better understanding of the linguistic characteristics of the news that readers consume daily in Brazil.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text analysis; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Bias, Machine Learning, Natural Language Processing, News, Subjectivity

1. INTRODUCTION

According to *Gallup and Knight Foundation*'s 2017 Survey on Trust, Media and Democracy, Americans believe that 62% of the news they consume is biased [1]. On the book *Bias: A CBS Insider Exposes How the Media Distort the News* [Goldberg 2001], Bernard Goldberg, who worked for about thirty years as a reporter for the American television network CBS, states that news coverage is often slanted to benefit parties, groups, people or ideas that are aligned with the journalists' ideology.

In general, journalistic news articles should make use of a formal, clear and objective language, crafted from the third person point of view, to establish an impersonal style. By contrast, biased news often adopt a more subjective language to influence or persuade readers through the use of emotional appeal and/or persuasion techniques [Wiebe et al. 2005; Mihalcea et al. 2007]. Under the premise that legibility and impartiality are characteristics of the journalistic genre, high levels of subjectivity combined with either low or high legibility in news may indicate some type of bias.

In previous work, subjectivity lexicons manually constructed by linguists [Amorim et al. 2018] were used to measure, based on *word embeddings*, subjectivity in news articles [Sales et al. 2019]. The present paper proposes the extension of those works by measuring the subjectivity of news published by seven popular media outlets in Brazil: *Carta Capital*, *El País*, *Estadão*, *Folha de São Paulo*, *O Antagonista*, *O Globo* and *Veja*. We correlate the levels of subjectivity with (i) the news size, (ii) readability metrics and (iii) the news popularity level. We perform a preliminary study typifying the

[1] https://www.knightfoundation.org/reports/perceived-accuracy-and-bias-in-the-news-media

subjectivity contained in the most popular Brazilian media outlets, exploring their impact through the readers' perspective, in terms of readability and popularity.

We defined the news size as the number of words in the article, and the popularity as the number of comments. For readability, the adopted metrics are the following: Coleman Liau Index [Coleman and Liau 1975], Smog Index [Nigam et al. 2016], Lix Index [Anderson 1983] and Dale Chall Score [Klare 1952]. Also, to guide this study, we defined the following research questions (RQ):

—**RQ1**: How do the levels of subjectivity between the media outlets compare to each other?
—**RQ2**: How strong is the correlation between subjectivity and readability?
—**RQ3**: How does the subjectivity correlates with the news size?
—**RQ4**: Is there a correlation between subjectivity and popularity?
—**RQ5**: Do the events that took place during 2018 (e.g. Brazilian general elections, FIFA World Cup) influence the subjectivity level of their corresponding news sections?

Note that from R1 to R4, we are interested in studying the subjectivity level in the news articles. With R5, on the other hand, we go deeper into exploring events that might have influenced the subjectivity level of those articles.

The adopted method, along with the research questions answers, shall contribute to a better understanding of the linguistic characteristics of the news that are daily consumed by Brazilian readers.

2. RELATED WORK

Media Bias is a broad concept that has been instantiated as distinct metrics throughout the years. For example, we found prior work approaching media bias as coverage bias [Soontjens et al. 2020], ideology and spin [Mullainathan and Shleifer 2002], gate-keeping [Al-Rawi 2019], and statement [Hamborg et al. 2019], among others. In this research, we instantiate media bias as subjectivity bias.

Subjectivity bias is also a widely studied concept in journalistic texts [Wilson et al. 2005; Chaturvedi et al. 2015; Yaqub et al. 2018]. Wilson et al. [2005], studied subjectivity in news, automatically identifying opinions, feelings and speculations contained in the text, using a *Naive Bayes* classifier. Chaturvedi et al. [2015], in turn, worked with subjectivity detection through the use of Convolutional neural networks based on *word embeddings*. We, in our turn, base our subjectivity computation on a set of 5 predefined lexicons, as described in section 4.

Frequently, the concept of subjectivity is associated with popularity. Bae and Lee [2012] analyzed the popularity of *tweets* through the bias contained in the text. In this case, the bias was defined in terms of the influence of positive and negative feelings. Flaounas et al. [2013], in turn, investigated the prediction of news popularity through the bias encompassed in the reported fact.

The following analysis of subjectivity is automatic and based on subjectivity lexicons. This methodology is employed in some related works [Amorim et al. 2018; Sales et al. 2019; Moraes et al. 2016; Jha et al. 2016; Lima et al. 2019]. Amorim et al. [2018], for example, utilizes the subjectivity calculation approach using lexicons to evaluate comments given by ENEM, a Brazilian high school leaving exam, essay reviewers. To assess subjectivity, the researcher uses lexicons of argumentation, sentiment, presupposition, modalization and valuation. Likewise, Sales et al. [Sales et al. 2019] based their research on the same lexicons to address subjectivity as a form of media bias when analyzing political news.

In a previous work [Lima et al. 2019], we performed a correlation analysis between the media outlets subjectivity levels and metrics such as readability and popularity to characterize subjectivity in five popular Brazilian media outlets. The presented research extends our previous work by (i) increasing the analysis coverage including more media outlets and news sections in the analysis, such as sports,

(ii) extending the correlation analysis for multiple news sections, and (iii) verifying whether important events (e.g., presidential elections, FIFA's Men World Cup) influenced the journalistic coverage.

3. DATABASE

We crawled news articles and comments, from January to December of 2018, from all sections of the following media outlets: *Carta Capital*[2], *El País*[3], *Estadão*[4], *Folha de São Paulo*[5], *O Antagonista*[6], *O Globo*[7] and *Veja*[8]. In 2018, sports and politics sections covered events such as the FIFA Men's World Cup[9] and the Brazilian general elections[10], notably interesting events to be analyzed due to a greater tendency for news regarding those topics to be presented in a more subjective way, according to the sports/political preferences of its editors. For example, the Brazilian press might tend to cover the FIFA World Cup under the Brazilian team's perspective, as a writer covering the elections is prone to do it under his/her political positioning.

Estadão, *Folha de São Paulo* and *O Globo* are *mainstream* media outlets consumed by readers with diverse sports/political preferences. *O Antagonista* is a declared right-wing media outlet[11] while *Carta Capital* openly supports the left-wing ideology[12]. Finally, *Veja* declares itself as opposed to the government regardless of its political positioning[13].

We used *Carta Capital*, *O Antagonista*, and *Veja* for validating our subjectivity analysis through a sanity check experiment. Due to their predefined ideological positioning, one might expect that some news articles present a higher subjectivity rates than media outlets that are supposed to be less biased.

The following information was collected for each news item: publication date, author, section, title, text, URL, and comments. Similarly, each comment contains its publication date, author, and text content. Table I presents the number of publications and comments collected for each media outlet.

	Carta Capital	El País	Estadão	Folha de SP	O Antagonista	O Globo	Veja
News	6,971	3,172	58,702	30,085	33,131	35,391	38,076
Comments	0	0	52,637	37,492	2,196,993	45,191	68,789

Table I. Number of publications and comments per media outlet.

Notice that, even though we only have crawled news from 2018, our database is composed of 205,528 news articles, which should be enough for granting the consistency of our results. Also, our database does not contain *Carta Capital's* and *El País'* comments as these media outlets did not publish, or did not make available, their comments sections at the time of our crawling. In order to create a shared database, the news from the different media outlets were grouped according to topics they address, such as politics and economics. Table II summarizes the distribution of some features related to morphological characteristics of our database.

[2]https://cartacapital.com.br
[3]https://brasil.elpais.com
[4]https://www.estadao.com.br
[5]https://www.folha.uol.com.br
[6]https://www.oantagonista.com
[7]https://oglobo.globo.com
[8]https://veja.abril.com.br
[9]`https://en.wikipedia.org/wiki/2018_FIFA_World_Cup`
[10]`https://en.wikipedia.org/wiki/2018_Brazilian_general_election`
[11]`https://pt.wikipedia.org/wiki/O_Antagonista`
[12]`https://pt.wikipedia.org/wiki/CartaCapital`
[13]`https://pt.wikipedia.org/wiki/Veja`

	mean	std	min	25%	50%	75%	max
Sentences	6.98	12.06	0	1	3	8	479
Unique Words	175	145.60	0	60	143	249	3242
Total words	319	334.82	0	78	229	454	14961

Table II. Morphological Characteristics of the News.

On Table II, we notice that, on average, a news article has around 7 sentences, 175 unique words and 319 total words. We can also observe a big standard deviation for all metrics (i.e., sentences, unique words and total words), which is justified due to the morphological variation associated to the journalistic coverage of each media outlet, as pictured by figure 6.

4. THEORETICAL FOUNDATION

This section introduces concepts that were adopted in our research and that are necessary to ease its understanding.

4.1 Subjectivity

Subjectivity can be defined as the interlocutor's ability to nominate himself as a subject [Benveniste 1971], introducing his own opinion to what is said, which may influence the readers' opinion and contribute to the construction of beliefs and values shared by people.

In previous work, Sales et al. [Sales et al. 2019] proposed the use of subjectivity lexicons to measure, using *word embeddings*, the subjectivity of journalistic texts. Those lexicons were built by [Amorim et al. 2018] through the manual analysis of expressions that often appear in texts where the interlocutor seems to express some subjectivity. Each lexicon encapsulates an aspect of subjectivity, more specifically these aspects are:

—**Argumentation (arg)** represents words and expressions that are related to an argumentative discourse, such as: "even" (*até*), "by the way" (aliás), "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.

—**Sentiment (sen)** gathers words and expressions that indicate the presence of the interlocutor's mood or feelings in the news. Words like "unfortunately" (*infelizmente*), "fortunately" (*felizmente*), and "preferably" (*preferencialmente*) are examples that belong to this set of aspects

—**Presupposition (pre)** contains expressions that suggest that the interlocutor accepts previous assumptions as true. Examples from this lexical set are: "nowadays" (*hoje em dia*), "to keep on" (*continuar a*), and factive verbs.

—**Modalization (mod)** This lexicon indicates that the writer exhibits a stance towards its own statement. Some examples of such markers are adverbs, auxiliary verbs, modality clauses, and some type of verbs, such as "undeniable" (*inegável*), "fundamental" (*fundamental*) and "fair" (*justo*).

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

Through those aspects the percentage of subjectivity present in the news articles was estimated.

4.2 Readability

For calculating readability, the following state-of-the-art metrics were considered: *Coleman Liau Index* [Coleman and Liau 1975], *Smog Index* [Nigam et al. 2016], *Lix Index* [Anderson 1983] and *Dale Chall Score* [Klare 1952]. Each metric is briefly explained below.

4.2.1 *Smog Index.* The *Smog Index* estimates the level of instruction needed to understand a text. It is based on the number of words with 3 or more syllables within a total of 30 sentences of the text, where 10 sentences are crafted from the beginning, 10 from the middle and 10 from the last part of the text. The metric is defined as:

$$SMOG = 1.0430 \times \sqrt{P \times \frac{30}{S}} \tag{1}$$

where P refers to the number of words with 3 or more syllables among the chosen sentences and S indicates the total amount of sentences in the text.

4.2.2 *Coleman Liau Index.* Similarly to *Smog*, *Coleman Liau Index* is a readability metric designed to assess the comprehensibility of a given text. *Coleman Liau Index* is based on the number of characters (letter or digit) and sentences employed for every 100 words utilized. Hence, the metric is determined as follows:

$$CLI = 0.0588 \times \bar{L} - 0.296 \times \bar{S} - 15.8 \tag{2}$$

where $\bar{L}$ represents the number of characters used and $\bar{S}$ the number of sentences formed for each 100 words.

4.2.3 *Lix Index. Lix Index* indicates the difficulty of reading a text based on the number of long words (the ones composed by more than 6 letters), the total of words and the number of sentences used in the text. A sentence is defined by the presence of a period, quotation mark or first letter capitalized. Its formula is presented below:

$$LIX = \frac{L}{B} + \frac{C.100}{L} \tag{3}$$

B determines the number of sentences in the text, C is the number of long words and L represents the total number of words in the text.

4.2.4 *Dale Chall Score.* The *Dale Chall Score*, in turn, evaluates the difficulty of reading a text based on the total of difficult words, the total of words and the total of sentences in the text. Difficult words are identified by their absence in a predefined list containing 3000 words that are considered easy for elementary school students to understand. Any word that is not on that list is considered difficult. In this work, we considered a list with words in Brazilian Portuguese. The metric is calculated as follows:

$$SCORE = 0.1579 \times (\frac{D}{L} \times 100) + 0.04696 \times (\frac{L}{S}) \tag{4}$$

D refers to the number of difficult words in the text, L is the total number of words and S indicates the number of sentences.

5. METHODOLOGY

A preparation was carried out on the raw text corpus in anticipation of the text mining assignments, where tasks such as the removal of numbers, whitespaces, punctuation marks and stopwords were performed.

After that, in order to better represent the semantic relations of words and expressions contained in the news, we trained a *word embedding* model based on the *word2vec skip-gram* [Mikolov et al. 2013] algorithm. The *window size* and *negative samples* are 5. As input, we used a corpus containing 205,528 news articles crawled from the selected media outlets (see I).

With the trained model, we measure the Word Mover's Distance (WMD) between each news article $n \in N$ and subjectivity lexicon $s \in S$, where $S = \{arg, pre, sen, val, mod\}$, as depicted in Figure 1.

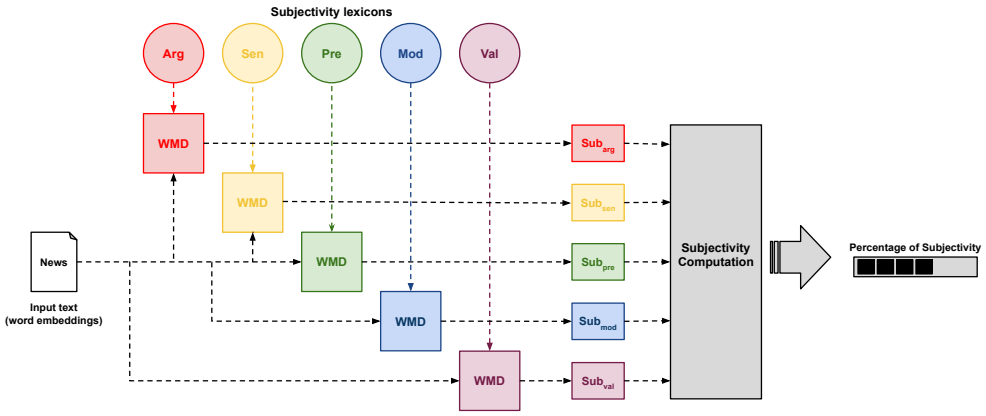

Fig. 1. Methodology representation. Given a news article and subjectivity lexicons, we 1) compute the Word Mover's Distance between its body text and each subjectivity lexicon, represented as sub_{lex} (e.g., sub_{arg}, sub_{sen}); and 2) use them to compute the percentage of subjectivity.

The WMD relies on the words of two predefined documents and their positions in the vector space, defined by a *word embedding* model, to calculate the lowest cost (euclidean distance) needed to overlap all the words in the first document to those in the second one. Since we measure the distance between a news article and a subjectivity lexicon, the resulting value might be considered as the amount of subjectivity contained in the news article. Thus, the amount of subjectivity of the type s in a news article n is given by $WMD(n, s)$.

More precisely, for each news article n, $||S||$ types of subjectivity are calculated. Therefore, the subjectivity sub_n of an article n can be represented by the $||S||$-dimensional subjectivity vector shown in Equation 5.

$$sub_n = \{WMD(n, s) | s \in S\} \tag{5}$$

We summarise the subjectivity vector into a single value by calculating the average of its values, as shown in Equation 6. Note that the subtraction in the Equation 6 is a mathematical tool to transform the distance value into a similarity value. Thereupon, the higher the $perc_n$ metric, the greater the subjectivity level in the news.

$$perc_n = 1 - \frac{\sum_{s \in S} WMD(n, s)}{||S||} \tag{6}$$

6. VALIDATION

In order to validate the lexicons usage in the given context, we check whether there is a significant difference of subjectivity between the informative and opinionated news articles. Naturally, the opinionated news articles are expected to present higher levels of subjectivity.

In addition to that, as a second step, we verify whether declared biased media outlets (e.g., O Antagonista) exhibited significantly higher subjectivity levels in their articles than the non-declared positioned media outlets (e.g., O Globo). We assume that, due to their declared stance, the declared-positioned media outlets are more prone to openly express their opinions on a reported topic than the non-declared positioned ones.

For each experiment, we compute the news articles subjectivity confidence intervals for the mean and discuss the results. The confidence intervals were computed with 5000 re-samples with random size, through *bootstraping* with replacement and 95% of confidence.

6.1 Opinionated vs. Informative

The first experiment is presents a comparison of the subjectivity levels in opinionated and informative news articles. Each news of our *dataset* was labeled as opinionated or informative, where opinionated news are considered to be those published in sections such as opinion blogs and columns of each media outlet. For that, we manually analyzed the sections of each media outlet and then defined URLs patterns that were used to classify the articles. The URLs patterns consist of keywords related to the opinion columns, such as the columnist name (e.g. Carlos Andreazza, Miriam Leitão, Merval Pereira) and names of opinion blogs (Chuteira FC, Futebol por Elas, Lance!), if one of those keywords is included in the URL, the article is considered as opinionated.

Opinionated news are expected to present greater subjectivity than informative news, given that this type of news aim to explicitly express the opinion and perspective of the interlocutor about the reported fact.

Figure 2 exhibits the confidence intervals of the average difference in subjectivity between opinionated and informative news. Confidence intervals containing the 0 value indicate that there is no significant difference between the subjectivity of opinionated and informative articles. In contrast, confidence intervals placed entirely over 0 indicate that there is a significant difference between the classes.

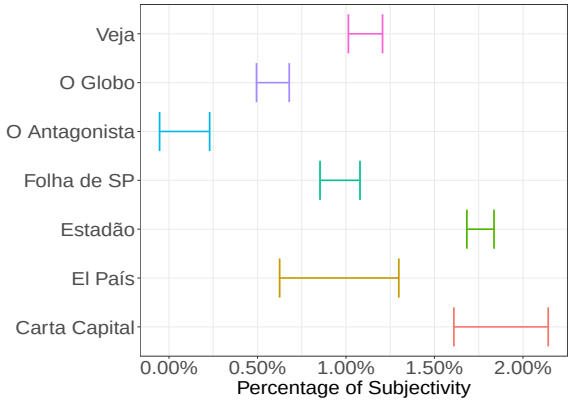

Fig. 2. Difference in subjectivity between opinionated and informative news.

Except for *O Antagonista*, the estimated range for the media outlets confidence intervals are positive and do not include 0. Thereby, one can conclude that there is, indeed, significant subjectivity difference between opinionated and informative news, and that our method is able to correctly model this difference.

Regarding *O Antagonista*, we suspect that its result does not show a significant difference due to the small size of its articles (Figure 6). In section 7.3, we further analyze the correlation between subjectivity and the news article's size.

6.2 Political Positioning

This experiment concerns the difference of subjectivity associated with the declared and non-declared positioned media outlets. In other words, we separate the declared positioned media outlets (i.e., VEJA, Carta Capital and O Antagonista) from the ones whose political position is undeclared. Intuitively, one can expect that media outlets with declared political positioning present higher subjectivity in their news, when compared to undeclared positioning ones, due to the idea of exposing their opinions on the reported facts.

Figure 3 shows the confidence intervals for the mean subjectivity (equation 6) contained in the news of the media outlets with declared and non-declared political positions. The results indicate that media outlets with declared political positioning exhibited higher values of subjectivity than those with non-declared positioning.

Fig. 3. Estimation of subjectivity by political position.

These results reinforce our hypothesis that positioned media outlets present higher subjectivity levels in their news than non-declared positioned ones and attest the efficiency of our method in capturing subjectivity in news articles.

7. RESULTS

In this section, we present the results for our previously introduced research questions. The following results are represented by confidence intervals with 95% of confidence computed using *bootstrapping* with 2000 re-samples and random size replacement.

7.1 Subjectivity by Media Outlet

The first experiment is related to the **RQ1**. Our goal is to estimate the subjectivity level for each media outlet and then compare them to each other in order to observe differences between declared

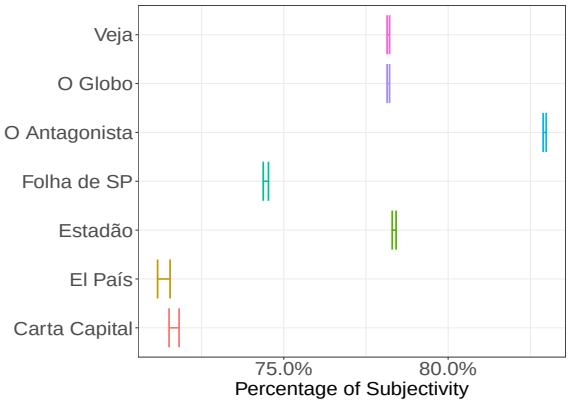

Fig. 4. Average subjectivity per media outlet.

and non-declared media outlets individually. With that goal, we estimate the confidence interval of the average percentage of subjectivity, for each media outlet, considering all news articles published in 2018.

Figure 4 presents the confidence interval of the average percentage of subjectivity for each media outlet. *O Antagonista* exhibits the highest subjectivity value followed by *VEJA*, *O Globo*, and *Estadão*.

Given their declared political positions, *O Antagonista*'s and *VEJA*'s results are somehow expected. On the other hand, *Carta Capital*'s low valued results, when compared to the other media outlets, are a surprise and require further research.

It is also interesting to note that although *O Globo* and *Estadão* are classified as non-declared media outlets, they do not present a significant difference in subjectivity when compared with *VEJA*. This result might suggest that *O Globo* and *Estadão* are not exempt from emitting their opinion, and perhaps influencing people, despite their non-declared positioning. For instance, considering the event of the Brazilian general elections, in theory, *VEJA* was expected to be significantly more subjective than *O Globo* or *Estadão* on its coverage, which was not observed in practice.

7.2 Readability vs. Subjectivity

The readability of a text refers to its typographic quality, reflecting its ease of reading. Regarding **RQ2**, in this experiment we estimate Spearman's correlation [Zar 1972] between the subjectivity and the readability metrics defined in section 4.2. The results are shown in Figure 5.

For any readability metric, the estimated correlation confidence intervals are presented inside the range -0.1 to -0.3. That is, we can conclude there is a weak correlation between subjectivity and readability. This shows that the proposed method captures well the subjectivity aspects present in the text, regardless of its reading complexity.

7.3 Subjectivity vs. Size

In this experiment, we answer **RQ3** by checking the Spearman correlation between the subjectivity of the news and its size (i.e., number of words). The main objective is to analyze whether the estimated subjectivity of a news article can be influenced by its size. With that goal, we first show the average news size for each media outlet and secondly, we present an individual analysis for each one of them by grouping into two categories: small and big news.

Figure 6 refers to the average size of the news article for each media outlet. *El País* publishes the

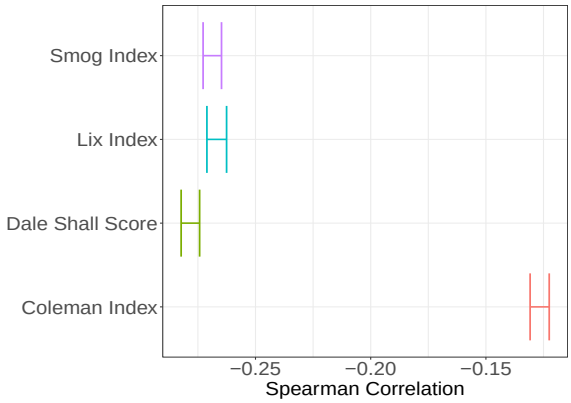

Fig. 5. Correlation between subjectivity and readability.

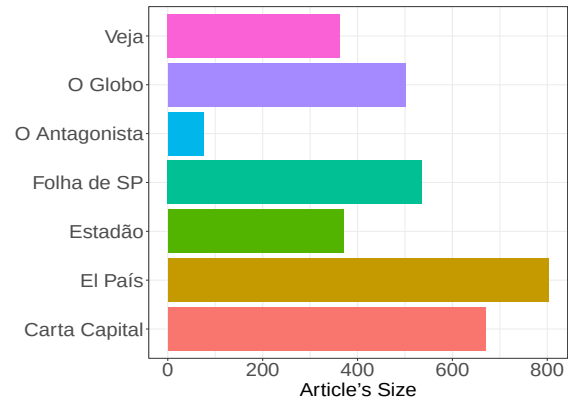

Fig. 6. Average Article's Size per media outlet.

longest articles – containing around 819 words in average –, followed by *Carta Capital* – with 756 words. *O Antagonista*, in contrast, publishes the shortest articles, with 75 words in average. These numbers might indicate how long those media outlets take to describe a fact or give their opinion about a topic.

We split the news considering the average size of each media outlet, resulting in two categories; one with the news that were below the average and one with the ones above. That was necessary because the average size of the news is quite different for each media outlet. As depicted in Figure 7, for most of the media outlets, regardless of the category, the news articles size is negatively correlated with the subjectivity; which is not the case for *O Antagonista* and *Carta Capital*.

Considering the small news category, only those two media outlets presented a positive correlation value between text size and subjectivity. Here, a positive correlation value implies that the higher the size of the news, the higher the subjectivity level.

Based on our results, we conclude that smaller news articles tend to be more subjective than long news articles. One possible reason for this is that when we use few words to describe an event while giving our opinion, it will be more difficult for a reader to distinguish between what is the fact and what is the opinion. Thus, the writer's opinion can easily be absorbed and spread by the reader.

A second possibility is related to the news article's motivation. News articles that are aiming to inform are concerned with providing more details about the reported fact, resulting in greater clarity

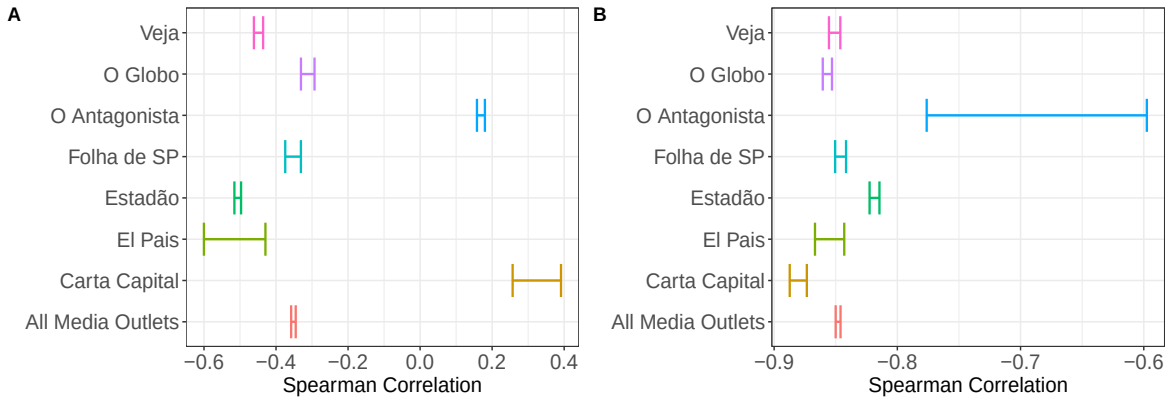

Fig. 7. Correlation between subjectivity and size.

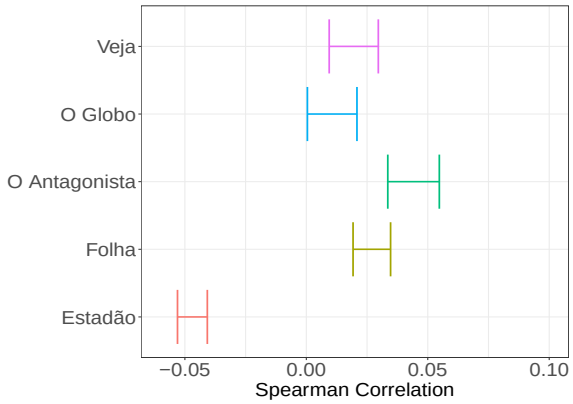

Fig. 8. Correlation between subjectivity and popularity.

to the reader and, probably, a diminished subjectivity level.

7.4 Popularity vs. Subjectivity

In order to answer **RQ4** we analyzed the existence of a correction between subjectivity and popularity. Here, we are interested in investigating whether high-subjective news articles tend to be more popular than low-subjective ones. To measure the popularity of an article, we consider the amount of comments that it has as a metric, therefore, only media outlets that provide comments on its publications were considered. The relation between subjectivity and popularity may be related to the adoption of emotional appeal [Wiebe et al. 2005; Mihalcea et al. 2007] aiming to persuade the readers to share its content and increasing its popularity. Figure 8 shows the estimated confidence interval of Spearman's correlation coefficient per media outlet.

Since no confidence interval includes the zero value, the results suggest the existence of a weak correlation between the news popularity and its subjectivity. *Estadão* is the only media outlet with negative results. We conclude, then, that subjectivity might influence the popularity of a news article even though weakly external factors are probably more decisive for determining its popularity.

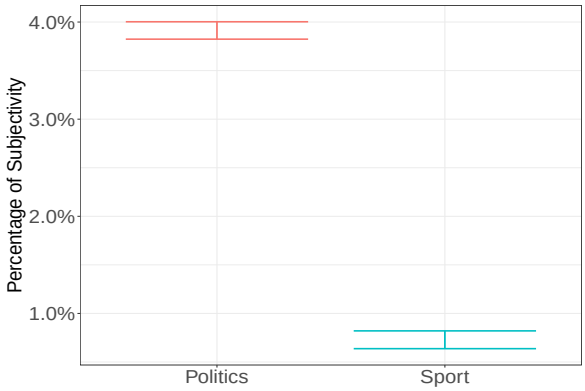

Fig. 9. Difference of the subjectivity in event coverage.

7.5 Subjectivity by Section

At last, we answer **RQ5** by analyzing the subjectivity distribution of some section throughout 2018. The objective here is to verify whether important events that happened in 2018 had influence over the journalistic coverage.

Based on the premise that events influence the news coverage, news addressing these events would be more subjective than the others. Thus, we consider events such as the Brazilian general elections and FIFA World Cup to estimate the news subjectivity from the political and sports sections of all media outlets.

In order to measure the impact of events on the news coverage, we split the news from each section into two groups: news addressing and news not addressing those events. For that, we have manually defined sets of keywords that characterize each of these events and checked whether the news article title contains one or more of these keywords. The keywords are distinct for each event:

—FIFA Men's World Cup: Seleção Brasileira (Brazilian National Team), Copa do Mundo (World Cup);
—Presidential elections: eleições presidenciais (presidential elections) and name of the candidates.

In the politics dataset, 64770 of the articles are related to the presidential elections, against 16881 articles that are not. Regarding the sports dataset, 21955 articles are related to the FIFA Men's World Cup, in contrast to 11286 articles that did not address this event. Figure 9 shows the confidence interval of the average difference between the news that are related to events and the unrelated ones, for each section.

Both confidence intervals are positive and do not contain 0, suggesting that such events have influenced and increased the subjectivity level by approximately 2%. This influence is probably related to the competitive aspect associated with the events and the high journalistic coverage portrayed by journalists who, occasionally, have opposite views about the events and might end up transmitting traces of their preferences when addressing them. However, further research is still necessary to evaluate how this subjectivity difference, of around 2%, is reflected in the news articles and how it impacts the readers.

Still, to elucidate how the difference in subjectivity occurs under the journalistic coverage perspective, we observe the following example, which consists of two news articles about the motivation of the murder of Marielle Franco, a Brazilian politician and activist who served as city councillor in the Municipal Chamber of Rio de Janeiro. The first article, published by *Estadão*, brings as headline

"Militia members killed Marielle because of lands, says general "[14] and presents around 77.4% of subjectivity, according to our metric. The second article, published by *O Globo*, presents the headline "Marielle Franco was killed by Escritório do Crime"[15] and presents 78.5% of subjectivity. *Escritório do Crime* is a Brazilian Militia organization specialized in contract killing. The difference in subjectivity between these two news items is 1.1%. However, one can feel a significant divergence of impact between these headlines.

8. CONCLUSION AND FUTURE WORK

It is common sense the perception that some media outlets are more biased than others in the way of exposing the facts. News with higher subjectivity can influence the opinion of readers and even in the construction of beliefs and values shared by public opinion.

In this paper, we conduct a subjectivity analysis on news articles from seven different Brazilian media outlets: *Estadão*, *Folha de São Paulo*, *El País*, *O Globo*, *Carta Capital*, *O Antagonista* and *Veja*. Their subjectivity levels were related to different aspects of the news (e.g., readability, popularity) through the application of state-of-the-art methods.

The presented analysis main findings are:

—*O Antagonista* and *Veja* were classified as the most subjective media outlets. Intuitively, this result was expected, since these are media outlets with a declared political alignment, which provides more evidence about the effectiveness of the method introduced;

—The subjectivity is weakly correlated with the readability of its text and very weakly correlated with popularity;

—The results show evidence that shorter news tend to be more subjective, which might be associated with the motivation of the news, as news that aim to inform are concerned with providing more details about the reported fact, resulting in greater clarity to the reader.

In future works, the proposed approach shall be extend and validated in other textual genres. In addition, its findings *features* will be included in a news recommendation system, seeking to reduce the subjectivity of the recommended news.

REFERENCES

Al-Rawi, A. Gatekeeping fake news discourses on mainstream media versus social media. *Social Science Computer Review* 37 (6): 687–704, 2019.

Amorim, E., Cançado, M., and Veloso, A. 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: Human Language Technologies, Volume 1 (Long Papers)*. pp. 229–237, 2018.

Anderson, J. Lix and rix: Variations on a little-known readability index. *Journal of Reading* 26 (6): 490–496, 1983.

Bae, Y. and Lee, H. Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology* 63 (12): 2521–2535, 2012.

Benveniste, E. Subjectivity in language. *Problems in general linguistics* vol. 1, pp. 223–230, 1971.

Chaturvedi, I., Cambria, E., Zhu, F., Qiu, L., and Ng, W. K. Multilingual subjectivity detection using deep multiple kernel learning. *Proceedings of Knowledge Discovery and Data Mining, Sydney*, 2015.

Coleman, M. and Liau, T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60 (2): 283, 1975.

[14] https://brasil.estadao.com.br/noticias/rio-de-janeiro,milicianos-mataram-marielle-por-causa-de-\
\terras-diz-general,70002645671

[15] https://oglobo.globo.com/rio/orlando-de-curicica-diz-que-marielle-franco-foi-morta-pelo-escritorio\
\-do-crime-23092732

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., and Cristianini, N. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital journalism* 1 (1): 102–116, 2013.

Goldberg, B. *Bias: A CBS Insider Exposes How the Media Distort the News*. Regnery Publishing, 2001.

Hamborg, F., Donnay, K., and Gipp, B. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20 (4): 391–415, 2019.

Jha, V., Shreedevi, G., Shenoy, P. D., and Venugopal, K. Generating multilingual subjectivity resources using english language. *Int. J. Comput. Appl* 152 (9): 41–47, 2016.

Klare, G. R. A table for rapid determination of dale-chall readability scores. *Educational Research Bulletin*, 1952.

Lima, D. F., Sales, A., and Balby, L. An analysis of subjectivity in brazilian news. KDMILe, Fortaleza, Ceara, Brazil, pp. 81–88, 2019.

Mihalcea, R., Banea, C., and Wiebe, J. Learning Multilingual Subjective Language via Cross-Lingual Projections. *Proceedings of ACL* 1 (1): 14–21, 2007.

Mikolov, T., Le, Q. V., and Sutskever, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Moraes, S. M., Santos, A. L., Redecker, M., Machado, R. M., and Meneguzzi, F. R. Comparing approaches to subjectivity classification: A study on portuguese tweets, 2016.

Mullainathan, S. and Shleifer, A. Media bias. Tech. rep., National Bureau of Economic Research, 2002.

Nigam, S., Kumar, N., Mandal, N., Padma, B., and Rao, S. Real time ambient air quality status during diwali festival in central, india. *Journal of Geoscience and Environment Protection* vol. 4, pp. 162–172, 2016.

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

Soontjens, K., Van Remoortere, A., and Walgrave, S. The hostile media: politicians' perceptions of coverage bias. *West European Politics*, 2020.

Wiebe, J., Wilson, T., and Cardie, C. Annotating Expressions of Opinions and Emotions in Language. *Empirical Methods in Natural Language Processing* 1 (1): 164–210, 2005.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. pp. 34–35, 2005.

Yaqub, U., Sharma, N., Pabreja, R., Chun, S., Atluri, V., and Vaidya, J. Analysis and visualization of subjectivity and polarity of twitter location data. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, pp. 67, 2018.

Zar, J. H. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association* 67 (339): 578–580, 1972.