Journal of Universal Computer Science, vol. 26, no. 2 (2020), 173-199
submitted: 9/1/19, accepted: 24/2/20, appeared: 28/2/20

Analysing Bias in Political News

Gabriel Domingos de Arruda
(EACH/USP, São Paulo - SP, Brazil
gda.gabriel@gmail.com)

Norton Trevisan Roman
(EACH/USP, São Paulo - SP, Brazil
norton@usp.br)

Ana Maria Monteiro
(UNIFACCAMP, Campo Limpo Paulista - SP, Brazil
anammont@cc.faccamp.br)

Abstract: Although of paramount importance to all societies, the fact that media can be biased is a troubling thought to many people. The problem, however, is by no means easy to solve, given its high subjectivity, thereby leading to a number of different approaches by researchers. In this work, we addressed media bias according to a tripartite model whereby news can suffer from a combination of selective coverage of issues (Selection Bias), disproportionate attention given to specific subjects (Coverage Bias), and the favouring of one side in a dispute (Statement Bias). To do so, we approached the problem within an outlier detection framework, defining bias as a noticeable deviation from some mainstream behaviour. Results show that, in following this methodology, one can not only identify bias in specific outlets, but also determine how that bias comes about, how strong it is, and the way it interacts with other dimensions, thereby rendering a more complete picture of the phenomenon under inspection.
Key Words: Bias detection, Bias in news, NLP applications
Category: J.4, H.4.m

1 Introduction

Media plays a fundamental role in all societies, providing information necessary for the individual decision making process and for the support of (or opposition to) government decisions. However, the very same information that helps people in this process could be manipulated so as to drive the thoughts of entire populations, insomuch that media has sometimes been referred to as the "fourth state" (*e.g.* [Dallmann et al., 2015]). As a result, it is widely believed that newswire media is ideologically slanted [Budak et al., 2016], as revealed in a study according to which 78% of the public in the United States believed the US media to be biased [Urban, 1999].

Bias, however, is not a particularity of the medium only, but instead, it could be something that also lies in the eye of the beholder. In this sense, it has already been noticed that people usually find it difficult to objectively identify bias, as shown by [Schmitt et al., 2004], who found that neutral and partisan people classify news differently, to the extent that an article that is considered neutral by some person is taken to be positive (or, conversely, negative) by someone else, depending on this person's previous alignment to the reported matter.

On the other hand, it might just be the case that bias comes up as a problem which is inherent to the production of news [Park et al., 2009], to the extent that it would not be possible for some newswire producer (and consumer) to be completely neutral. Yet, given media's importance, it becomes crucial to be able to identify and recognise biased reports of facts, if we are to live in an open society [Dallmann et al., 2015]. Quantifying bias, however, can be methodologically challenging [Budak et al., 2016]. Still, even though one cannot really quantify it within a single source, it is possible to gather evidence for it, by referring to different sources of information [Park et al., 2009], so as to try to cover different aspects of the same issue, and thereby increase the odds of having a more complete description of the facts.

In following this idea, much of the extant research approaches bias within a tripartite model (*e.g.* [Saez-Trumper et al., 2013,Dallmann et al., 2015]), according to which bias could be introduced in news in three different ways: through slanted selection, coverage and statement. Selection Bias (sometimes also called "gatekeeping" [D'Alessio and Allen, 2000] or "filtering" [Budak et al., 2016]) refers to the selective coverage of the issues to be presented to the public [Budak et al., 2016], that is which stories are reported and which are not to be brought to the public's attention [D'Alessio and Allen, 2000]. Even though some degree of selectivity is always expected, since one cannot simply report everything [Saez-Trumper et al., 2013], this dimension refers to an unbalanced way of doing so.

Coverage Bias, in turn, measures the disproportionate amount of attention given to some subject in comparison to others, thereby accounting for the relative amount of coverage (time or space) each person or story receives in the reporting of some issue [D'Alessio and Allen, 2000, Saez-Trumper et al., 2013, Dallmann et al., 2015]. Finally, Statement Bias (sometimes also known as "framing" [Budak et al., 2016]) deals with the way facts are reported [Saez-Trumper et al., 2013] and, more specifically, how the news producer's opinions are woven into the text [D'Alessio and Allen, 2000, Dallmann et al., 2015], by reporting more favourable (or unfavourable) news about some party, for example, being usually measured through the analysis of the sentiments associated with the report.

In this article, we present the results of applying this tripartite model to a corpus of political news in Brazilian Portuguese (*cf.* [de Arruda et al., 2015]). Collected from five big newswire outlets, the corpus was annotated by four volunteers, who had to identify, for each paragraph in the news, the person who is the main subject of that paragraph, along with the paragraph's polarity towards that person. Since news were collected during the 2014 presidential and state elections, the corpus focus on the main candidates running for governor of the state of São Paulo and president of Brazil. As such, annotators had to choose among these candidates for the main subject of each paragraph, ruling out any other person outside this scope. The choice for this kind of data followed a similar reasoning as that of [D'Alessio and Allen, 2000], who pointed out that such campaigns can be both small and sufficiently scrutinised to build a fertile ground for the analysis of bias and, more specifically, Selection Bias.

Selection Bias, however, is not the only type of bias that our model was able to verify through this corpus. In analysing the polarity related to each candidate, we could also identify Statement Bias, by comparing how different sources portray the same candidates. Through the computation of the amount of news related to each candidate, it was also possible to account for Coverage Bias, should some outlet astray from the mainstream behaviour, as demonstrated by a comparison amongst all analysed outlets. Finally, regarding Selection Bias, this same approach was used to capture lacks of coverage too, that is when some outlet decides to omit news about some party. In this case, bias would become evident from the identification of which candidates are reported and which are excluded from news, when comparing each outlet against its counterparts.

To accomplish our goal we relied on a couple of outlier detection techniques, in order to determine, for each dimension, whether some outlet might be considered biased (*i.e.* an outlier) in that dimension. To the best of our knowledge, we are the first to apply such techniques to this tripartite model and to make this analysis based on entire paragraphs, as opposed to statements (*e.g.* [Saez-Trumper et al., 2013]) or the polarity of surrounding words (*e.g.* [Dallmann et al., 2015]). By working at the paragraph level, we hoped that, given the broader context, annotators would have made more trustworthy decisions, thereby building a more precise corpus and so leading to a more accurate assessment of bias. Also, the fact that our model can be fully automated, as it will be shown in the forthcoming sessions, makes it an interesting choice for the analysis of bias in large amounts of data.

This model could then be used not only to identify bias in specific outlets, but also to determine how that bias comes about, by pointing out which dimensions have these outlets as outliers. Moreover, since our outlier detection procedure tries to determine how far some data point lies from the median value of its counterparts (see Section 2 for details), our method provides both the identification of a biased outlet and a measure of this bias' amplitude, along with its direction, that is through the overstatement or understatement of some fact, or through its exaggerated or minimised coverage, for example. Finally, even though the model was applied to a corpus of news in Brazilian Portuguese, we see no reason why it could not be used with any other language, provided that a corpus similar to that of [de Arruda et al., 2015] can be found in the target language.

2 Materials and Methods

Within the context of this research, which relies on a corpus of news gathered through Twitter messages, in which tweets are collected and their links to the original news followed[1], the above mentioned dimensions can be redefined as follows:

- *Selection Bias*: the preference for choosing facts related to some specific politician, measured by the amount of references each producer makes to each candidate at the news texts, and so assessing how often that politician was referenced by the analysed media.
- *Coverage Bias*: the preference for giving more attention to some specific politician, assessed through the amount of distinction given to that politician by the news outlet. Usual measures for this kind of bias are the size of columns, pictures and headlines [D'Alessio and Allen, 2000]. Even though, throughout the internet, time and space constraints are not much of an issue, potentially rendering such measures ineffective, Twitter squish them to the limit, by demanding messages to be no longer than 140 characters[2]. Hence, a possible measure for Selection Bias is to determine whether some candidate mentioned in the news text was also mentioned in the tweet leading to that text. This, in turn, would raise this candidate's prominence, according to the inverted pyramid principle, whereby information and facts in news should be organised in a decreasing order of importance [Park et al., 2012].
- *Statement Bias*: the preference for expressing more favourable (or, conversely, unfavourable) opinions towards some specific politician, measured by the proportion of positive, negative and neutral paragraphs associated to each candidate in news texts by each producer.

As it turns out, it is not possible within our model to determine whether some isolated outlet can be considered biased. This conclusion could come only through a comparison amongst sets of outlets. Since the corpus of political news at hand comprises texts from the same time period, collected through the same news distribution tool (*i.e.* Twitter), it would not be so strange to expect little to no difference, amongst the analysed outlets, regarding who is reported and how this person is reported. As such, in order to identify possible differences between outlets, we relied on the identification of outliers, defined as an observation presenting a clear deviation from the remaining observations in the sample where it occurs [Grubbs, 1969]. Thus, an outlet will be considered biased at some

[1] See [de Arruda et al., 2015] for details on this procedure. Tools used to calculate statistics can be found at `https://github.com/gdarruda/metricas-vies`

[2] Even though in 2017 it was announced that this limit would raise to 280 characters, by the time of this research it was still 140.

dimension if, for that dimension, it strongly deviates from its counterparts (*i.e.* if it is an outlier).

To do so, we assumed data to be normally distributed around the "true value" of each dimension, and adopted the commonly used z-score metric [Cousineau and Chartier, 2010] to identify outliers, whereby one excludes data points lying beyond a threshold corresponding to a number of standard deviations from the mean. Using the mean as a central tendency indicator, however, may not be the best choice, due to the fact that outliers are already included in the assumed distribution, with great influence to its mean and standard deviation [Leys et al., 2013]. This, in turn, reduces the odds of identifying them in small data samples. As an alternative, one can replace the mean with the median, which is less sensitive to outliers, and use the Median Deviation [Hampel, 1974], that is the median of the absolute deviations from the median (*i.e.* $median(|x_i - median(X)|)$, where $X = \{x_i, 1 \leq i \leq N\}$ is the dataset at hand), to estimate data deviation.

To use Median Deviation as a consistent estimator for data deviation, however, one has to multiply it by a constant scale factor b which, for the Normal distribution, has the value of $b \approx 1.4826$ [Leys et al., 2013]. This estimator, called Median Absolute Deviation, and which can be mathematically stated as $MAD = b \times median(|x_i - median(X)|)$, can then be used as our measure of statistical data dispersion. As a didactic example, consider the dataset $D = \{2, 5, 4, 1, 8, 8, 7, 1000\}$,[3] which clearly presents an outlier in its last element. To calculate its MAD, one builds the new set $D' = \{4, 1, 2, 5, 2, 2, 1, 994\}$, by subtracting each point in D from its median value 6 and taking the absolute value of the result. Next, one takes the median of this new set – 2 – and multiply it by b, which leads to $MAD = 2.97$. Using 6 as a central tendency indicator and 2.97 as a deviation value seems much more appropriate than using the 129.38 mean and 351.80 standard deviation counterparts of the original dataset.

Once calculated the dataset's MAD, it is necessary to define a threshold beyond which a data point will be considered an outlier. To do so, we followed [Miller, 1991], who suggests adopting 2.5 or 2.0 standard deviations as a cut-off value, instead of the commonly adopted 3.0 for Normal distributions. In our case, however, we took 2.0 MADs around the median, instead of Standard Deviations around the mean, as our threshold.

3 Results

In following our procedure for Selection Bias, we calculated the proportion of paragraphs in which each candidate was pointed out as the paragraph's target entity by the majority of annotators, related to the overall number of paragraphs

[3] Adapted from [Leys et al., 2013].

in news texts by the same outlet. Table 1 summarises these results. In this table, G_1 to G_3 refer to the three main candidates running for governor of the state of São Paulo, whereas P_1 to P_3 refer to those running for president of Brazil and N_1 to N_5 refer to the five analysed Twitter profiles[4].

Table 1: Proportion of references to each candidate in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	14.63%	2.44%	0.00%	14.63%	12.20%	19.51%
N_2	1.69%	1.98%	4.52%	2.82%	23.45%	12.99%
N_3	4.03%	3.76%	3.23%	8.33%	20.56%	20.56%
N_4	5.60%	0.00%	0.00%	0.00%	23.20%	16.00%
N_5	0.00%	0.00%	0.00%	3.28%	16.94%	24.59%
Mean	5.19%	1.64%	1.55%	5.81%	19.27%	18.73%
Median	4.03%	1.98%	0.00%	3.28%	20.56%	19.51%

As it would be expected, candidates for governor received less attention than their presidential counterparts, even though G_1 was referenced, in average, almost as much as P_1, who was running for president. Also, magazines (*i.e.* N_4 and N_5) seem to let these candidates aside, focusing in those running for president instead, something that was not observed amongst newspapers (*i.e.* N_3 and N_1) and the online news portal N_2. That could be related to the fact that, even though all these news producers are actually situated in the state of São Paulo, magazines are supposed to address a national readership, thereby moving away from the local political scenario. Still, remarkably one of these magazines – N_4 – was second only to N_1 in referencing G_1.

From the computation of MAD values for Table 1, we can determine each news producer's deviation from the median for each candidate. These values, shown in Table 2, were computed as follows. Let us assume we are currently analysing P_3, for example. From Table 1, we have the relative amount of reference this candidate had across all producers, which leads to the dataset $X = \{0.1951, 0.1299, 0.2056, 0.1600, 0.2459\}$[5], with $M = 0.1951$ as its median value. Subtracting M out of this sequence, and taking the absolute value of the results, we come to $X' = \{0.0652, 0.0351, 0.0000, 0.0105, 0.0508\}$, with the new median $M' = 0.0351$ which, when multiplied by 1.4826 gives us $MAD = 0.0520$. The number of deviations can then be calculated from $(x_i - M)/MAD$, $\forall x_i \in X$,

[4] See [de Arruda et al., 2015] for more details about candidates and outlets.
[5] Here these values are shown as proportions, instead of percentages.

leading us to $\{0.00, -1.25, 0.20, -0.67, 0.98\}$, as shown in Table 2.

Table 2: Deviations from the median in news texts.

	Target Entity					
	Governor			**President**		
Twitter Profile	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$
N_1	3.06	0.17	∞	2.34	-1.96	0.00
N_2	-0.67	0.00	∞	-0.09	0.67	-1.25
N_3	0.00	0.67	∞	1.04	0.00	0.20
N_4	0.45	-0.75	∞	-0.67	0.62	-0.67
N_5	-1.16	-0.75	∞	0.00	-0.85	0.98

In this table, we see that only N_1 exceeded the two deviations from the median threshold, with G_1 and P_1. In both cases, the deviation was positive, indicating that, during the analysed period, these candidates were mentioned more often in this newspaper than in the remaining news producers. One interesting point about this table is that it shows a clear drawback of this model, which is the possibility of one having infinity values for the deviation, as occurred with G_3. In this case, these values were due to a zero MAD, which rendered any other value an outlier, since the ratio between deviations from the median and MAD produces infinity values. In this research, we have ignored such cases (we will come back to this topic in Section 4).

Moving on to the analysis of Coverage Bias, we calculated, for each news producer, the amount of paragraphs in news texts related to each specific candidate (*i.e.* paragraphs having that candidate as their target), whose source tweet (*i.e.* the tweet providing the link to the news) also mentioned that same candidate. This value was then divided by the total number of paragraphs in news texts by that outlet. This measure indicates then the overall proportion of paragraphs in each outlet whose target entities were also mentioned in the paragraph's source tweet, thereby increasing their prominence in the news. Table 3 shows the results. As somewhat expected, here too candidates running for president were more often cited than those running for governor. One interesting result, however, comes from the comparison between Selection (Table 1) and Coverage (Table 3), which shows a remarkable similarity amongst candidates for governor. This high similarity, however, disappears when we move to the presidential side of these tables. The reasons for this are still to be determined, and we will discuss this further in Section 4.

The amount of deviations from the median for Coverage Bias, in terms of MAD, can be seen in Table 4. Once more we have infinity values for one candi-

Table 3: Proportion of direct references to candidates in tweets by each producer.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	14.63%	2.44%	0.00%	29.27%	12.20%	29.27%
N_2	1.69%	1.98%	4.52%	5.65%	42.09%	23.73%
N_3	4.30%	3.76%	3.23%	14.52%	32.93%	40.32%
N_4	5.60%	0.00%	0.00%	0.00%	33.60%	27.20%
N_5	0.00%	0.00%	0.00%	6.56%	26.23%	53.01%
Mean	5.24%	1.64%	1.55%	11.20%	29.41%	34.71%
Median	4.30%	1.98%	0.00%	6.56%	32.93%	29.27%

date, who had no direct reference in tweets by three of the five outlets, thereby rendering its MAD zero. As with the results for Selection, once again N_1 exceeded the two deviation threshold, positively for G_1 and P_1, meaning these were mentioned more often in this outlet than in its counterparts, and negatively for P_2, indicating that, when compared to the remaining news producers, this candidate was undermentioned in tweets by this outlet. Another producer crossing the two-deviation line was N_5, which mentioned P_3 more often than its counterparts.

Table 4: Deviations from the median in the tweets.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	2.67	0.17	∞	2.34	-2.09	0.00
N_2	-0.67	0.00	∞	-0.09	0.92	-0.67
N_3	0.00	0.67	∞	0.82	0.00	1.35
N_4	0.34	-0.75	∞	-0.67	0.07	-0.25
N_5	-1.11	-0.75	∞	0.00	-0.67	2.89

Finally, regarding Statement Bias, results for the proportion to the overall number of paragraphs in news texts, by each news producer, related to each candidate (*i.e.* that had that candidate as their target entity), of those classified as positive, neutral and negative news for that candidate by the majority of annotators, are shown in Tables 5, 6 and 7, respectively. In these tables, a value of 62.50% in Table 5 for P_3 in N_1, for example, with 12.50% in Table 6, and

25.00% in Table 7, means that, from all paragraphs in the news texts by N_1, whose target was P_3, 62.50% were classified as positive, 12.50% as neutral, and 25.00% as negative news for that candidate.

Table 5: Proportion of positive references to candidates in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$
N_1	50.00%	0.00%	0.00%	50.00%	20.00%	62.50%
N_2	0.00%	57.14%	56.25%	40.00%	25.30%	50.00%
N_3	26.67%	28.57%	4.17%	35.48%	35.95%	40.52%
N_4	14.29%	0.00%	0.00%	0.00%	31.03%	35.00%
N_5	0.00%	0.00%	0.00%	50.00%	3.23%	33.33%
Mean	18.19%	17.14%	12.08%	35.10%	23.10%	44.27%
Median	14.29%	0.00%	0.00%	40.00%	25.30%	40.52%

Table 6: Proportion of neutral references to candidates in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$
N_1	0.00%	0.00%	0.00%	50.00%	80.00%	12.50%
N_2	100.00%	42.86%	43.75%	30.00%	45.78%	30.43%
N_3	26.67%	35.71%	16.67%	38.71%	29.41%	24.18%
N_4	0.00%	0.00%	0.00%	0.00%	44.83%	35.00%
N_5	0.00%	0.00%	0.00%	50.00%	16.13%	17.78%
Mean	25.33%	15.71%	12.08%	33.74%	43.23%	23.98%
Median	0.00%	0.00%	0.00%	38.71%	44.83%	24.18%

Corresponding deviations from the median can, in turn, be found in Tables 8, 9 and 10, respectively. As it turns out, in Table 8 the deviation threshold was crossed, at the negative side, for P_2 by N_5, meaning this candidate was subject to fewer positive news by this producer than by other outlets. Even though the threshold was also crossed by N_4 with P_1, we do not consider this as an indication of bias, since this value comes up as a result of the fact that N_4 made no reference to this candidate whatsoever, as can be observed by summing up

Table 7: Proportion of negative references to candidates in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$
N_1	50.00%	100.00%	0.00%	0.00%	0.00%	25.00%
N_2	0.00%	0.00%	0.00%	30.00%	28.92%	19.57%
N_3	46.67%	35.71%	79.17%	25.81%	34.64%	35.29%
N_4	85.71%	0.00%	0.00%	0.00%	24.14%	30.00%
N_5	0.00%	0.00%	0.00%	0.00%	80.65%	48.89%
Mean	36.48%	27.14%	15.83%	11.16%	33.67%	31.75%
Median	46.67%	0.00%	0.00%	0.00%	28.92%	30.00%

the results for this candidate and magazine across Tables 5 to 7. It makes then little sense to speak of Statement Bias when no statement was made in first place.

At the positive side, P_3 received more positive reports from N_1 than from other outlets. As for neutral references (Table 9), the only outlet to cross the two deviation threshold, at the negative side, was N_4 with P_1, but once again we cannot take this result as an indication of bias, for it was caused by an absence of references to this candidate in this magazine. Finally, negative reports (Table 10) were those with the highest deviations at both negative and positive sides. Negative references to P_2 were seen less often in news by N_1, whereas considerably more often in N_5. N_5 was also off the threshold in such references for P_3. Again, zero medians had some candidates receive infinity values in these tables.

Table 8: Deviations from the median of positive references to candidates.

	Target Entity					
	Governor			President		
Twitter Profile	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$
N_1	1.69	∞	∞	0.67	-0.62	2.06
N_2	-0.67	∞	∞	0.00	0.00	0.89
N_3	0.58	∞	∞	-0.3	1.25	0.00
N_4	0.00	∞	∞	-2.70	0.67	-0.52
N_5	-0.67	∞	∞	0.67	-2.6	-0.67

Table 9: Deviations from the median of neutral references to candidates.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	∞	∞	∞	0.67	1.54	-1.23
N_2	∞	∞	∞	-0.52	0.04	0.66
N_3	∞	∞	∞	0.00	-0.67	0.00
N_4	∞	∞	∞	-2.31	0.00	1.14
N_5	∞	∞	∞	0.67	-1.26	-0.67

Table 10: Deviations from the median of negative references to candidates.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	0.06	∞	∞	∞	-3.41	-0.64
N_2	-0.81	∞	∞	∞	0.00	-1.33
N_3	0.00	∞	∞	∞	0.67	0.67
N_4	0.67	∞	∞	∞	-0.56	0.00
N_5	-0.81	∞	∞	∞	6.09	2.41

3.1 Infinity Values

As shown in the analysis above, even though MAD and median may be better tools for detecting outliers, specially given their relative stability in the presence of such deviations, they come with the undesirable side effect of producing many infinity values, which rules part of the dataset out of the analysis as a whole, thereby reducing the potential extension of conclusions. As an alternative, we considered rolling back to mean and standard deviation, and calculated z-scores for the same datasets. Results for Selection Bias are shown in Table 11 (the z-score equivalent of Table 2), whereas results for Coverage are shown in Table 12 (the z-score equivalent of Table 4), and for Statement in Tables 13 to 15 (the z-score equivalents of Tables 8 to 10).

As it turns out, the only outlets to reach the two z-score threshold in these tables were N_5 for P_3, regarding Coverage Bias (a difference already pointed out in our analysis through MAD), and N_3 for G_3, regarding Statement Bias and, more specifically, the number of negative references to this candidate. It is noticeable, however, that this last value was reported as infinity when using MAD and median. Still, even though MAD may produce infinity values, going back to z-scores does not help much, for mean and standard deviation move towards

Table 11: Candidate citation z-scores in news texts by each producer.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	1.85	0.55	-0.80	1.71	-1.67	-1.28
N_2	-0.69	0.23	1.53	-0.58	0.98	-0.78
N_3	-0.23	1.46	0.86	0.49	0.31	0.64
N_4	0.08	-1.12	-0.80	-1.13	0.93	-0.10
N_5	-1.02	-1.12	-0.80	-0.49	-0.55	1.53

Table 12: Candidate citation z-scores in tweets by each producer.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	1.85	0.55	-0.80	1.78	-1.73	-0.52
N_2	-0.70	0.23	1.53	-0.55	1.27	-0.42
N_3	-0.19	1.46	0.86	0.33	0.35	-0.52
N_4	0.07	-1.12	-0.8	-1.1	0.42	-0.52
N_5	-1.03	-1.12	-0.80	-0.46	-0.32	2.00

the outliers' direction, reducing the odds of detecting them (as happened in our dataset), unless outliers are weight-balanced along the positive and negative sides of the scale, which is not to be expected as an usual feature.

Nevertheless, and despite the reduction in the amount of deviating outlets that can be detected, the fact that we have access to real values (instead of infinity scores) can show us some interesting patterns, such as the change in the

Table 13: Z-scores for positive references to candidates in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	1.70	-0.75	-0.55	0.81	-0.27	-1.43
N_2	-0.97	1.75	1.99	0.27	0.19	1.54
N_3	0.45	0.50	-0.36	0.02	1.14	0.58
N_4	-0.21	-0.75	-0.55	-1.90	0.70	-0.24
N_5	-0.97	-0.75	-0.55	0.81	-1.76	-0.44

Table 14: Z-scores for neutral references to candidates in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	-0.65	-0.81	-0.71	0.88	1.72	-0.32
N_2	1.93	1.40	1.85	-0.20	0.12	0.82
N_3	0.03	1.03	0.27	0.27	-0.65	-0.32
N_4	-0.65	-0.81	-0.71	-1.83	0.07	1.35
N_5	-0.65	-0.81	-0.71	0.88	-1.27	-1.52

Table 15: Z-scores for negative references to candidates in news texts.

	Target Entity					
	Governor			President		
Twitter Profile	**G_1**	**G_2**	**G_3**	**P_1**	**P_2**	**P_3**
N_1	0.41	1.87	-0.50	-0.81	-1.28	1.16
N_2	-1.11	-0.70	-0.50	1.37	-0.18	-1.48
N_3	0.31	0.22	2.00	1.07	0.04	-0.24
N_4	1.5	-0.70	-0.50	-0.81	-0.36	-0.50
N_5	-1.11	-0.70	-0.50	-0.81	1.79	1.07

z-score sign for a single entity only, across all news by some outlet[6], for example. This phenomenon can be observed in Table 11, when dealing with Selection Bias, for N_3 and N_5. In this case, N_3 lies above the mean in references to all candidates, when compared to the remaining outlets, whereas falling below the mean only for G_1, which means this outlet made an above the mean amount of references to all candidates but this one, for whom it stayed under the mean. Conversely, N_5 lies below the mean for all candidates, except for P_3, meaning this candidate was mentioned more often in news texts by this outlet.

Throughout the tables, we see this also happening with Coverage (Table 12), for N_5 and P_3, and Statement (Tables 13 to 15). Regarding this last kind of bias, we see in Table 13 that this is a widespread phenomenon for positive references to candidates, to the extent that it was not observed in only one outlet – N_1. For the remaining outlets, N_2 was below average only for G_1 (meaning it had fewer positive news about this candidate than the mean amount of positive news by all outlets), N_3 was below average only for G_3, N_4 was above average only for P_2, and N_5 was above average only for P_1.

As for neutral references to candidates (Table 14), N_2 was below average

[6] Something that could not be properly observed with MAD, given the existence of these infinity values.

only for P_1, who also happened to be the only candidate for whom N_5 was above average. The remaining outlets did not present this behaviour for neutral comments. Moving on to negative references (Table 15), this phenomenon can be seen in N_2, which was above average only for P_1 (meaning they had more negative reports about this candidate than the average amongst all outlets), N_3, which was below average for P_3, and N_4, which was above average for G_1. Table 16 summarises these results, pointing out, for each analysed dimension and outlet, the only candidate to present a z-score with opposite sign to that of his/her counterparts in the same outlet.

Table 16: Candidates presenting a z-score sign different from all others.

Outlet	**Selection**	**Coverage**	**Statement**		
			Positive	**Neutral**	**Negative**
N_1	–	–	–	–	–
N_2	–	–	G_1 (-0.97)	P_1 (-0.20)	P_1 (1.37)
N_3	G_1 (-0.23)	–	G_3 (-0.36)	–	P_3 (-0.24)
N_4	–	–	P_2 (0.70)	–	G_1 (1.5)
N_5	P_3 (1.53)	P_3 (2.00)	P_1 (0.81)	P_1 (0.88)	–

However interesting, these results raise the question as to what extent they are relevant, that is to what extent one can actually expect such behaviour by chance (or, similarly, what are the odds of having n out of m candidates to present positive z-scores in the same outlet, with all others lying at the negative side). In an attempt to shed some light into this question, let us assume there to be a probability p of some outlet making an above the mean amount of references to a specific candidate. Conversely, $(1 - p)$ would be the probability of making a below the mean amount of references (for the sake of simplicity, an exact mean amount is taken to be on the positive side of the scale). Under these circumstances, the probability of having exactly n out of m candidates lie above the mean, assuming independence amongst them, is given by [Mitchell, 1997]

$$P(above = n, below = (m - n)) = \frac{m!}{n!(m-n)!} p^n (1-p)^{m-n}$$

Furthermore, let us assume that news are uniformly distributed amongst candidates, and that outlets are unbiased (*i.e.* they randomly choose news from this distribution). Within this set-up, one can expect news to be balanced around the mean for each candidate, that is one can expect $p = 0.5$, which leads to a $P(above = 1, below = 5) \approx 9.4\%$ chance of having a single candidate present a positive z-score while the remaining five are at the negative side (the same applies to a negative z-score analysis). Although an approximately 20% chance[7] may not be high enough to rule out the possibility of such a result being but random fluctuation, one must remember that we have come to this number through some very strong (and unconfirmed) assumptions about data distribution and independence. Also, it might just be the case that this phenomenon comes from the natural shifting in the mean by outliers. Still, we think this is something that deserves future investigation.

4 Discussion

Even though we approached bias within a tripartite model, according to which one should look at the problem under three different viewpoints (*i.e.* the slanted selection, coverage and statement of news), it becomes paramount to understand that, however different, these dimensions are by no means to be taken in isolation. Let us take, for example, the issue raised in Section 3 regarding Statement Bias. In this case, our bias threshold was crossed by N_4 for P_1. However, in analysing the data in Tables 1 and 3, one sees no reference to this candidate by this outlet. But then how can someone be reported in a slanted way when that someone was not reported at all? It turns out that the outlier limit was crossed not because there were slanted mentions about the candidate, but only because the absence of any mention pushed the candidate beyond this line, and this should be considered before deeming an outlet biased.

Now, let us consider Coverage Bias. In analysing only the data in Table 4, one might come to the conclusion that P_1 and P_3 were treated similarly by N_1 and N_5, since both had similar deviations from the median regarding their Coverage in tweets (2.34 and 2.89, respectively). However, when one turns to Tables 8 and 9, and look at the way these candidates were mentioned and, more precisely, the amount of positive and neutral news about them, one sees that they lie in opposite sides of the median, although not so far as to have them taken as outliers in this dimension too. In this case, even though bias was detected in one dimension, one should be very cautious so as not to overstate this conclusion, by making unfunded claims. These examples, in turn, indicate that dimensions are actually complementary to each other, meaning that any assessment of bias must come through a joint analysis of them.

In fact, an analysis of the pairwise (linear) correlation between dimensions has shown there to be a strong correlation between Selection and Coverage only (Tables 1 and 3, respectively), indicating that candidates highlighted in tweets are usually those mentioned more often in their corresponding news texts. Weak correlations could also be observed between Statement (Tables 5, 6 and 7) and

[7] The chance of a single candidate lie above **or** below the mean while all others lie on the opposite side.

both Selection and Coverage, but only regarding positive news[8], indicating that the more often a candidate is mentioned, the more positive news we find about him/her. Table 17 illustrates these results. In this table, we present the values for the Pearson correlation coefficient, along with its associated p-value, for all dimension pairs. In this research, we followed [Hinkle et al., 2003], and took values between $0.3 < r \leq 0.5$ to represent a low (positive) correlation, between $0.5 < r \leq 0.7$ to be a moderate correlation, and values between $0.7 < r \leq 1.0$ to indicate a high correlation. All values in the table are reported at the 95% confidence level.

Table 17: Pearson correlation coefficient (r,p-value) between dimensions.

	Selection	Coverage	Statement		
			Positive	Neutral	Negative
Selection	(1, 0)	(0.97, $\ll$ 0.01)	(0.49, < 0.01)	(0.23, 0.22)	(0.27, 0.14)
Coverage	–	(1, 0)	(0.46, 0.01)	(0.20, 0.29)	(0.22, 0.23)
Positive	–	–	(1, 0)	(0.34, 0.06)	(-0.08, 0.69)
Neutral	–	–	–	(1, 0)	(-0.27, 0.15)
Negative	–	–	–	–	(1, 0)

Another point raised in Section 3 was the difference between news related to candidates running for governor and those running for president, whereby one sees a greater amount of attention being paid to the presidential run than to its state government counterpart. Table 18 illustrates these results for Selection and Coverage – the dimensions responsible for capturing the amount of attention given (and references made) to each candidate. As it turns out, this difference was found to be of statistical significance for both Selection and Coverage (Mann-Whitney $W = 0.00, p < 0.01$, at the 95% confidence level, for both dimensions). The reasons for this difference, however, remain uncertain, even though it might just be the case, as already pointed out, that the presidential run was preferred over state government because of its importance to a broader readership, whereas the state campaign would be restricted to the state of São Paulo only.

In this same table, another interesting result comes from the comparison between the figures for the state government run in both Selection and Coverage, which are almost identical, to the extent that there is a virtually perfect correlation between both sets (Pearson's $r \approx 1, p << 1$, at the 95% confidence level). On the other hand, even though at the presidential side we can still observe a smaller however high correlation, this correlation is not of statistical significance

[8] The weak correlation observed between Neutral and Positive statements was not of statistical significance ($p = 0.06$).

Table 18: Distribution of references to candidates for governor and president.

	Selection		Coverage	
Twitter Profile	**Governor**	**President**	**Governor**	**President**
N_1	17.07%	46.34%	17.07%	70.74%
N_2	8.19%	39.26%	8.19%	71.47%
N_3	11.02%	49.45%	11.29%	87.77%
N_4	5.60%	39.20%	5.60%	60.80%
N_5	0.00%	44.81%	0.00%	85.80%

(Pearson's $r = 0.75, p = 0.14$, at the 95% confidence level). This, in turn, could be an indication that the overall correlation observed between Statement and Selection might in fact have been mostly due to the state government side of the data, there being more variation in its presidential counterpart. Once again, the reasons for this phenomenon are still to be determined.

These results, along with the analyses they allow us to do, illustrate one of the greatest advantages of this methodology: the fact that one can not only define bias in a more objective way, that is through an analysis of outliers, whereby one can determine whether some news outlet departs from its counterparts along the analysed dimensions, but also that, in comparing the figures for each outlet and candidate across dimensions, some political and social phenomena can be unveiled, which could be of interest to a broader readership other than people interested in bias only. As a result, the gains obtained through this methodology as a whole outvalue the sum of its parts, to the extent that conclusions may arise through the joint analysis of all dimensions that could not be reached should these dimensions be taken in isolation.

Nevertheless, there are some drawbacks to the methodology too. As noticed in Section 3, the use of median and MAD led to many infinity values when calculating the number of deviations from the median for each outlet. We understand that these values should not be regarded as an indication of bias, since any value would produce the same output, given that the problem lies in the (zero) MAD part of the equation. In fact, should someone render infinity deviations as an indication of the existence of outliers, there would be situations, such as that of Table 9, in which all candidates are outliers, which is nonsense. To get matters worse, rolling back to mean and standard deviation did not help much, given their high sensitivity to outliers. However, some more extreme deviations could still be observed with this methodology, meaning that a joint analysis using both approaches might be more appropriate to the overall assessment of bias.

As also noted in [Dallmann et al., 2015], another important limitation of this method lies in the fact that it heavily depends on how data was sampled. Besides questions related to sample size, which can limit the strength of any conclusion, there is a potential threat to data validity regarding the method followed for choosing news producers to integrate the dataset. Naturally, the methodology used for sampling outlets depends on the overall goal of the study, that is whether someone is interested in analysing bias across all existing outlets, or just within some subset of these (perhaps with some specific feature). Whatever the procedure, one must bear in mind that if the sample is biased, then bias may not be detected. In other words, if one chooses only outlets known to favour one candidate over another, then that will become the estimated population distribution of news about these candidates, and no outlier regarding this pattern will be detected.

In this work, we tried to reduce the impact of this issue by attempting to balance outlets out according to their editorial view, that is by choosing, amongst those with a large readership, outlets widely believed to have different political alignments. That, however, is a highly subjective assessment made by two of the researchers, and a more objective methodology must be tested. If it turns out that such a methodology is in fact impractical, given the natural subjectivity of the task, at least it should be tried some methodology that relies on higher amounts of people judging outlets, so as to reduce any individual bias or misunderstanding. We leave this as a future work direction. Still, and even though our choices may have weakened our conclusions regarding bias itself, they do not rule the methodology out as a whole. On this regard, we firmly believe this to be a course to be pursued in the analysis of media news.

5 Comparison to the Related Literature

Although the detection and measurement of media bias has been an issue studied for over a decade, its precise definition and, consequently, the methodology followed to detect it are still by no means standard. Current approaches vary from the analysis of the terms authors use to compose their sentences and texts (*e.g.* [Fortuna et al., 2009,Recasens et al., 2013]), to more elaborated frameworks (*e.g.* [Mehler et al., 2006,Saez-Trumper et al., 2013,Iyyer et al., 2014,Morstatter et al., 2018]), whereby one first define bias in terms of a set of dimensions (usually, from one to three) and tries to identify this bias in news texts according to these dimensions. To get matters more complicated, even when authors agree on the number of dimensions to be analysed, they may not agree on what these dimensions are (*e.g.* [Budak et al., 2016,Morstatter et al., 2018]), thereby making it harder to compare results by different approaches.

Perhaps the most straightforward approach, defining bias in terms of the wording of sentences and texts has the advantage of posing very few requirements regarding annotated corpora. Examples under this approach vary from the identification of differences in the lexicon used in multiple reports of the same event, along with a measure of topic intersection [Fortuna et al., 2009], to the analysis of human edits aimed at removing bias from Wikipedia[9] articles [Recasens et al., 2013], in order to gather linguistic cues that might lead to the detection of words that would induce bias in texts. Although interesting, these methods rely on the assumption that texts can be naturally biased, and that this bias would become evident from the words chosen to compose them.

That, however, seems to be too strong an assumption, since differences in lexical choices can derive from writing style, for example. Even though using human judgments to detect "loaded" words may seem more appropriate an approach, that still relies on the assumption that such words exist, and that bias can be taken as a common sense notion, making these judgments reliable. Once again, this does not seem to be the case, given the low accuracy scores (from around 34 to 59%) obtained by humans and the tested computational model, as reported in [Recasens et al., 2013]. In this work, we understand bias to be something that cannot be analysed on the basis of raw differences between texts, but rather something that must be defined beforehand. That means before one can compare different sources, one must first determine what to compare in them.

As a matter of fact, this seems to be the approach adopted by much of the extant research. Differences arise, however, when it comes to defining what to look for in the available sources before comparing them. More commonly, researchers define bias along a single dimension, thereby focusing on a single aspect of the issue, and try to determine differences between sources under this perspective. This is the case, for example, with the work by [Iyyer et al., 2014], who try to identify ideological differences between texts. Relying on annotated corpora, along with lists of words associated to different political alignments, they built a classifier to tell liberal from conservative texts. Although the applied technique is expected to use all available evidence to determine features that render some text biased, it still depends on annotations that, ultimately, are based on the assumption of the existence of some common-sense notion of bias.

Somewhat departing from this need for a common definition, but still focusing on a single dimension of the issue, the work by [Mehler et al., 2006] analyses references to named entities across different geographical locations, developing a variance-based model to estimate the frequency of references to these entities in different cities. A similar approach is also applied by [Ward et al., 2009], who also try to determine differences in the frequency of named entities in news, but this time using a set of classifiers and a juxtaposition score for co-reference association, and focusing on aspects related to ethnicity and culture.

Adding one dimension to the analysis, [Morstatter et al., 2018] argue for the existence of two types of bias: agenda setting, which refers to the systematic selection of stories (very close to Selection Bias), and which could be assessed,

[9] `https://www.wikipedia.org/`

amongst other ways, by counting the sources referenced in news, and framing (or second-order agenda setting), which is defined as the reinforcement of specific aspects of a story. Framing seems to comprise both Coverage and Statement, being sometimes close to Selection too, since it "can take many forms, from emphasis of information to the selective presentation of information within a text" [Morstatter et al., 2018]. The problem is approached by first automatically detecting frames in news (from a set of 10 previously identified frame types), and then determining their polarity. Hence, even though taking bias as a two-dimensional problem, the focus still lies in only one dimension: framing.

In order to detect frames, the authors compare several different classifiers, along with ensembles involving some of them. Polarity is detected in a similar way, by splitting each frame in two different sub-frames, with a positive and a negative representative for each of them. Although heavily relying on a corpus annotated with frame types, the authors report a high inter-annotator agreement. Still, accuracy results were rather low, with a top value around 0.43 for the model with polarity. However interesting, this approach also takes an absolute viewpoint on bias, thereby rendering something as inherently biased or not. We, on the other hand, understand bias as something that must be defined relatively to other sources as deviations from some mainstream behaviour.

Approaching bias as a two dimensional problem, [Budak et al., 2016] deal with what they call issue filtering, that is the selective coverage of issues (which seems to involve both Coverage and Selection), and issue framing, which determines how issues are presented (akin to Statement Bias). To do so, the authors build some classifiers, trained over a corpus of human annotated news, to first tell political news from others. Next, they train different classifiers to identify each article' topic, along with its polarity towards the Democratic or Republican Party. Articles were then assigned a partisanship score reflecting their left or right leaning, and outlets were assigned the average of their articles' scores, weighted by each article's popularity. This assignment, in turn, allowed the authors to rank and compare outlets, identifying deviations from the mainstream.

Instead of applying some outlier detection technique to these scores, however, to determine the strength of the observed differences, the authors opt for a more qualitative comparison of the overall bias in outlets. A more detailed comparison can be found when they analyse issue framing and issue filtering separately, where a regression model was fit to this last dimension. Through this model, it was possible to identify deviations in coverage for Republican and Democrat scandals. Nevertheless, the authors render these differences to be non representative of issue filtering as a whole, also concluding that news outlets are surprisingly similar, something they attribute to the analysed period not coinciding with any election term. Still, through a visual inspection of the presented data, one notices that the authors' focus on the absolute, instead of relative, magnitude of the observed differences might have actually concealed outliers.

Moving on to tripartite approaches to bias, we find that authors usually rely on the same set of dimensions, to wit, Selection (or Gatekeeping), Coverage and Statement. Being the first ones to define bias within this tripartite model, [D'Alessio and Allen, 2000] present us with a meta-analysis covering 59 quantitative studies. Focusing on US presidential elections since 1948, the authors apply a d' test to identify differences in the political alignment of outlets, along with a χ^2 test of homogeneity to determine the strength of deviations. Even though claiming that only a small overall bias could be found in television networks (with newspapers and magazines showing virtually no bias at all), the authors also found studies indicating substantial ideological bias by some newspapers, reporters and editors. These, however, balance each other out, so the overall bias in the industry remains neutral.

Although being able to spot some outliers, the authors' focus on the overall alignment of outlets led them to use methods less adapted to outlier detection (such as the χ^2 test, for example). That, however, was not a problem for them, since they were interested in determining whether the media as a whole would present some overall pro-Republican or pro-Democratic bias. Still, our work can be seen, under a methodological viewpoint, as an extension to theirs, even though we focus on online news, as opposed to more traditional media, and account for candidates running for president and stated governor, as opposed to parties in a presidential election, in an attempt not to determine any overall bias in the communication industry, but instead, to highlight deviant behaviours by outlets towards specific candidates.

Another work to approach bias as a three dimensional problem can be found in [Saez-Trumper et al., 2013], who follow the definitions by [D'Alessio and Allen, 2000] to analyse a set of 80 international online news sources, along with social media communities around them, looking for the way a group of 10 heads of state was mentioned in news. In this work, Selection was identified by determining the amount of overlap amongst stories posted by these producers, as measured by the Jaccard coefficient. The similarity matrix was then projected in two dimensions and differences between countries were identified. Coverage, in turn, was measured by the amount of words in articles covering the same story[10]. Along with word distributions, the authors also calculate distributions for mentions to people, comparing them and analysing correlations between different geographical regions and political leaning.

Finally, Statement was measured, with the aid of a dictionary containing the valence of each word, by the amount of positive and negative expressions mentioning people. Within each news source, each person from the heads of

[10] Within social media, this bias was measured by the number of tweets with links to these articles.

state list was assigned a valence corresponding to the average of the sentiment in all statements on all articles mentioning that person in that source. Valence distributions were then used to compare social communities and news sources. Despite the fact that we measure Statement in a similar way, instead of assuming words to be naturally valenced (through a specialized dictionary), we rely on human assessments of the valence of paragraphs containing the target person, that is we try to determine the valence of the context in which that person was cited. Even though this may turn out as a more subjective approach, we find it more appropriate than assigning sentiment to words without context.

Our measures of Selection and Coverage, in turn, differ from those of [Saez-Trumper et al., 2013] mainly because we focus on finding deviant reports related to candidates in online news, not accounting for the social communities that arise around them. That led us to focus on direct references to these candidates, as opposed to making a broader analysis of stories. But even when dealing with a similar target, such as mentions to people, the main difference between our work and theirs lies in that, instead of dealing with a heterogeneous set of news producers, from at least eight different geographical locations worldwide, and clustering them so as to determine whether any cross-country patterns arise, we opted for finding outliers from a presumably homogeneous set. This, however, only reflects the way different goals shape the decisions we make towards them.

One last work to deal with the model presented in [D'Alessio and Allen, 2000] is that of [Dallmann et al., 2015], who studied political and economical news from four German online newspapers, in an attempt to determine the existence of bias towards political parties. Even though agreeing on this tripartite model, the authors focus in two of the dimensions only: Coverage and Statement. Coverage was approached by determining how often political parties and their members appear in news headlines and texts. These measures, however, do not integrate to build some overall indicator of Coverage, being analysed separately instead. Statement, in turn, was measured through sentiment analysis, whereby a four-word window was set around mentions to parties, with a sentiment score being assigned to this set. The overall sentiment towards some party was then calculated as the sum of all individual sentiments in the article.

Along with sentiment analysis, the authors also determined a list of keywords related to political orientation, counting the number of their occurrences in party manifestos and news texts. A cosine similarity between news and manifestos was then calculated. Once again, these metrics were analysed independently. Although displaying Coverage in terms of deviations from the mean, and applying a t-test to each party and newspaper to determine if deviations were significant, the authors apparently refrain from following the same procedure for Statement too. Also, and even though the t-test can give us an idea of the relevance of the observed differences, the fact that it was applied to deviations from the mean makes this procedure less suitable for outlier detection, given the sensitivity of the mean to them.

Overall, our procedure differs from the related work in that we set out to detect deviant behaviour by individual outlets regarding specific politicians, in a presidential and state government campaign. That goal, in turn, led us to use techniques more specific to outlier detection, and to take a broader view of the problem. As such, instead of focusing on a single or a couple of dimensions, as much of the extant research does, we approached bias within the tripartite model introduced by [D'Alessio and Allen, 2000]. But even amongst those who adopt this broader viewpoint, we could find no research approaching bias as an outlier detection problem. That, however, is not to be taken as a negative feature of theirs, or even a positive feature of ours, but instead as a consequence of different research objectives. Hence, we present our work not as a counterpart to the related research, but as a complement to it, so as to give others a wider set of options when making their own choices.

6 Conclusion

What is bias? This is a very important and yet tricky question, since one has no clear answer to it. Even though we could put it simply as "offering a partial perspective on facts" [Saez-Trumper et al., 2013], truth is that one has no way of telling that for sure, given the lack of references complying with some acceptable standard of fairness against which to compare news content [Shoemaker and Reese, 1996]. In this work, we understood bias as something that cannot be taken as a boolean feature of texts, whereby these are inherently biased or not, but instead as a noticeable deviation from some mainstream behaviour. "Deviant" labels, however, are not to be assigned to texts, but to their producers, and this assignment, in turn, must not be taken as a moral assessment of that producer, but only as a measure of how that producer fits amongst its counterparts regarding the reported issue.

This understanding of bias led us to define it according to the tripartite model introduced in [D'Alessio and Allen, 2000], and which resulted from observations on the way media bias their news, within an outlier detection framework, using the median as a central tendency indicator and MAD as a measure of deviation. To the best of our knowledge, we are the first to approach bias this way and, more specifically, to apply outlier detection techniques to this tripartite model. Also, we seem to be the first ones to base our analysis on paragraphs, instead of statements, sentences, words, or even whole texts. These were actually straightforward decisions to us, given our understanding of bias as deviant (*i.e.* outlier) behaviour and the focus on mentions to candidates in political news, which might benefit from a broader context as that provided by paragraphs.

As a result, some very interesting phenomena could be observed, beyond the plain identification of which news producers figured as outliers. Amongst these, perhaps the most important lesson is that bias is not to be analysed from a single dimension only. Instead, it should be determined from the joint analysis of all dimensions. This becomes even more important if we recall that dimensions were not found to be highly correlated, with the exception of Selection and Coverage, but mainly for the State government part of the data. This, in turn, is an indication that, by restricting the analysis to a single dimension, one would be bound to miss important information from the remaining ones, which could lead to the overstatement of results or even to wrong overall conclusions.

Another interesting result, and which is somewhat inline with those of [Mehler et al., 2006], was that nationwide elections are given more attention than local ones, which are better covered by newspapers, whereas magazines tend to focus on a broader readership. Even though the targeted readership might explain this behaviour, the real reasons for it remain to be addressed in future research. Finally, when using mean and standard deviation, instead of median and MAD, we could observe yet another phenomenon, which is the alignment in z-score valences for all but a single producer, thereby rendering this producer an outlier in relation to this pattern. Even though we could not see this as something highly improbable, mainly given the small number of outlets analysed, it still remains as a puzzling result that deserves some future attention.

These results, in turn, illustrate one of the main advantages of our approach, which is the fact that we were able not only to define bias in a more objective way (*i.e.* as deviant behaviour), but also to possibly detect different bias strategies, such as omission, for example. The method then captures both slanted presentation of facts and their omission, which would still render the omitting producer an outlier. Furthermore, through this method other phenomena could be observed which might be of interest to other knowledge areas, such as psychology or social and political sciences, for example. Still, some drawbacks exist, mainly related to the use of MAD which, despite being resilient to the presence of outliers, produces many infinity values. Even though rolling back to median and standard deviation did solve this problem, it came at the price of not detecting as many outliers as it should, given the sensitivity of the mean to them.

One possibility to overcome this problem would be to carry out the analysis using both median and mean, as we did in our work. Another possibility would be to slightly shift zero MADs by adding some small factor to them, akin to Laplace Smoothing for example. That, however, would have to rely on a complete statistical analysis of the meaning of this new data set, so as to determine the appropriateness of the idea, which we leave for future work. Another drawback to this work is its high dependency on the way the dataset was sampled, to the extent that, should this dataset be biased towards the same direction, then no bias would be detected at all. This is something inherent to any statistical analysis, and should be accounted for by all who decide to experiment using this methodology.

This feature, however, becomes an issue only if sampled outlets do not match the underlying assumptions about media fairness. On this regard, even though in this work we assumed that there should be an equity in the treatment of the same candidate by different outlets, that is reports about some candidate should be similar, in what refers to the analysed dimensions, across producers, our method does not depend on this assumption. As mentioned in Section 4, the adopted definition of bias should be reflected in the way outlets are sampled. If sampled outlets are expected to be naturally slanted towards some direction, then any deviation from this behaviour could be deemed as biased. The assumption, then, is that this mainstream behaviour reflects an accepted notion of fairness, whatever that might be. This, in turn, frees our method from cultural assumptions, making it suitable to different viewpoints.

As for future work directions, besides those already mentioned, we believe it would be interesting to fully automate this process. Even though most of this work was carried out automatically (namely, news collection and relative frequency calculations from Twitter and newswire websites), data for sentiment analysis was manually annotated, mainly to avoid errors introduced by automatic techniques and so help validate our outlier-detection approach to the problem. Nevertheless, we understand that automating this last part is a necessary feature to increase the method's usefulness. To this end, sentiment classification and entity resolution techniques could come out very handy. Regarding the model's generalization, even though our analysis relies on tweets and their associated texts as its primary source of information, it can be readily adapted to other sources, such as newswire texts for example. In this case, the only dimension to be adapted is Coverage, in which more traditional measures could be used.

Finally, the model could be applied not only to the identification of bias, but also to determine how this bias takes place (by analysing the dimensions in which the outlet figures as an outlier, along with its side around the median), how strong it is (by determining how far this outlier lies from the median), and the way it interacts with other dimensions, thereby rendering a more complete picture of the phenomenon under inspection. Moreover, despite the fact it was primarily applied to a corpus in Brazilian Portuguese, the method could be adapted to virtually any language, provided an equivalent corpus exists.

References

[Budak et al., 2016] Budak, C., Goel, S., and Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80:250–271. Special Issue.

[Cousineau and Chartier, 2010] Cousineau, D. and Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, 3(1):59–68.

[D'Alessio and Allen, 2000] D'Alessio, D. and Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of Communication*, 50(4):133–156.

[Dallmann et al., 2015] Dallmann, A., Lemmerich, F., Zoller, D., and Hotho, A. (2015). Media bias in german online newspapers. In *26th ACM Conference on Hypertext & Social Media (HT'15)*, pages 133–137, Guzelyurt, Northern Cyprus.

[de Arruda et al., 2015] de Arruda, G. D., Roman, N. T., and Monteiro, A. M. (2015). An annotated corpus for sentiment analysis in political news. In *10th Brazilian Symposium in Information and Human Language Technology (STIL 2015)*, pages 101–110, Natal, RN – Brazil.

[Fortuna et al., 2009] Fortuna, B., Galleguillos, C., and Cristianini, N. (2009). *Text Mining: Classification, Clustering, and Applications*, chapter Detection of Bias in Media Outlets with Statistical Learning Methods. Chapman and Hall/CRC, New York.

[Grubbs, 1969] Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

[Hampel, 1974] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

[Hinkle et al., 2003] Hinkle, D. E., Wiersma, W., and Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences*. Houghton Mifflin, 5 edition.

[Iyyer et al., 2014] Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 1113–1122, Baltimore, Maryland, USA.

[Leys et al., 2013] Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

[Mehler et al., 2006] Mehler, A., Bao, Y., Li, X., Wang, Y., and Skiena, S. (2006). Spatial analysis of news sources. *IEEE transactions on visualization and computer graphics*, 12(5):765–772.

[Miller, 1991] Miller, J. (1991). Reaction time analysis with outlier exclusion: bias varies with sample size. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 43A(4):907–912.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

[Morstatter et al., 2018] Morstatter, F., Wu, L., Yavanoglu, U., Corman, S. R., and Liu, H. (2018). Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2).

[Park et al., 2009] Park, S., Kang, S., Chung, S., and Song, J. (2009). Newscube: delivering multiple aspects of news to mitigate media bias. In *SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, pages 443–452, Boston, MA, USA.

[Park et al., 2012] Park, S., Kang, S., Chung, S., and Song, J. (2012). A computational framework for media bias mitigation. *ACM Transactions on Interactive Intelligent Systems*, 2(2).

[Recasens et al., 2013] Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1650–165, Sofia, Bulgaria.

[Saez-Trumper et al., 2013] Saez-Trumper, D., Castillo, C., and Lalmas, M. (2013). Social media news communities: Gatekeeping, coverage, and statement bias. In *22nd ACM international conference on Information & Knowledge Management (CIKM '13)*, pages 1679–1684, San Francisco, CA, USA.

[Schmitt et al., 2004] Schmitt, K. M., Gunther, A. C., and Liebhart, J. L. (2004). Why partisans see mass media as biased. *Communication Research*, 31(6):623–641.

[Shoemaker and Reese, 1996] Shoemaker, P. J. and Reese, S. D. (1996). *Mediating the Message: Theories of Influences on Mass Media Content.* Longman, USA, 2 edition.

[Urban, 1999] Urban, C. D. (1999). Examining our credibility : perspectives of the public and the press. Technical report, ASNE Foundation. Report for the ASNE Journalism Credibility Project.

[Ward et al., 2009] Ward, C. B., Bautin, M., and Skiena, S. (2009). Identifying differences in news coverage between cultural/ethnic groups. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Milan, Italy.