An Information Theory Approach to Detect Media Bias in News Websites

Victoria Patricia Aires
New York University
Federal University of Amazonas
victoria.aires@icomp.ufam.edu.br

Juliana Freire
New York University
juliana.freire@nyu.edu

Fabiola G. Nakamura
Federal University of Amazonas
fabiola@icomp.ufam.edu.br

Altigran Soares da Silva
Federal University of Amazonas
alti@icomp.ufam.edu.br

Eduardo F. Nakamura
Federal University of Amazonas
nakamura@icomp.ufam.edu.br

ABSTRACT

News websites and portals are, together with social media, major sources of information nowadays. However, such types of media may be biased regarding, especially, political and ideological leaning/orientation. Hence, the awareness of such bias, leaning, or orientation is a key factor for the readers (content consumers) to decide how much content/opinion they accept or reject from a given source. Over the years, especially nowadays, biased information has been used as a tool to control and manipulate public opinion, ultimately leading to the proliferation of fake news. Consequently, it is important to develop methods to automatically identify and inform the reader about the eventual political and ideological bias of the sources. The majority of current research focuses on polarity detection or a bi-class problem, such as left vs. right-wing leaning or Democratic vs. Republican. In addition, most of them are based on a large number of features (lexical or bag-of-words), resulting in computationally intensive methods. In this work, we introduce Poll (POLitical Leaning detector), a strategy based on Information Theory concepts to detect media bias in news websites/portals considering bi-class and multi-class problems. Our strategy reduces the feature space to as little as the number of classes being considered, significantly reducing the overall computational cost. Compared to a representative baseline, our strategy yields a macro accuracy of up to 76% for a four-class problem compared to 22% for the baseline under the same conditions. For some classes, we could reach an F1 of 0.80 against 0.28 from the baseline.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Information systems** → *Content analysis and feature selection.*

KEYWORDS

media bias detection, news analysis, classification, online news

WISDOM '20, August 24th, 2020, San Diego, CA

ACM Reference Format:
Victoria Patricia Aires, Juliana Freire, Fabiola G. Nakamura, Altigran Soares da Silva, and Eduardo F. Nakamura. 2020. An Information Theory Approach to Detect Media Bias in News Websites. In *WISDOM '20: Workshop on Issues of Sentiment Discovery and Opinion Mining, August 24th, 2020, San Diego, CA.* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

With the popularization of the web and social media, news websites and portals have become major sources of information for the population. For the sake of simplification we will use the terms news websites and news portals indistinctly hereafter, given this simplification does not impact on the paper contributions. Compared to traditional vehicles such as newspapers, television, and radio, news portals are fast, have a worldwide reach, and allow interactions among the readers [1, 6, 11]. However, as with traditional vehicles, news from these portals may be biased [12, 19]. This is a fundamental problem that is currently of major importance to our society, partially because the readers tend to spend little time on reading and even less (or none) assessing the quality of the source [4, 5].

A starting point to mitigate the effect of bias in news dissemination is to identify the ideology of the content provider. A proper ideological identification allows the reader to reason about it and decide whether or not that ideology introduces bias (information that is incomplete, incorrect, misleading, or fake).

This first step is often approached as a bi-class problem that consists in classifying an information source as left/right (ideology), or as Democratic/Republican (partisan) [20]. However, in practice it might be useful to consider a multi-class problem [10], such as a four-class scenario: left (extreme or moderate) and right (extreme or moderate). The particularities of the moderate classes (vocabulary, style, citation patterns) make the boundaries of moderate left and moderate right a little fuzzy. These characteristics make the multi-class problem more challenging.

In this work, we propose a novel method that analyzes the content of articles to determine the political orientation/leaning and intensity of ideology of news portals (up to four classes). The proposed method uses basic Information Theory concepts to identify key content (Shannon Entropy [22]) and quantifying the differences (Jensen-Shannon divergence [9, 15]) between a target portal and news portals of known orientation to guide the detection of the political orientation of the target portal. Compared to more

Table 1: Summary of related works.

Work	Strategy	Classes	Performance	Target
Efron [7]	Hyperlink co-citations	Liberal/Conservative	77.50% (accuracy)	web documents
Krestel et al. [13]	TF-IDF vectors and Cosine similarity	Left/Right German political parties	not reported	German news outlets
Rao and Spasojevic [20]	Word Embeddings and LSTM	Democratic/Republican	87.57% (accuracy)	tweets
Elejalde et al. [8]	Rank difference	Liberal/Conservative	not reported	Chilean news outlets
Ribeiro et al. [21]	Audience demographics from social media	Liberal/Moderate/ Conservative	not reported	news sources
Gordon et al. [10]	Word Embeddings	Democratic/Republican	not reported	tweets
Baly et al. [3]	A varied set of features including lexical features	Left/Center/Right	41.74% (accuracy)	news sources

traditional solutions, our strategy represents a significant reduction in the (dimension) feature space used to characterize the news website, which aims at reducing the computational cost while keeping/improving the detection efficacy. In some cases, we were able to reach an accuracy score of 86%, by using as few as four features, against 43% from the baseline, which uses 282 features, and an F1 of 0.80, against 0.28 from the same baseline.

The key contributions of this paper are threefold: (i) we propose a classification strategy using Information Theory concepts such as Shannon entropy to compute more reliable features to classify media bias in news websites; (ii) we quantify the performance of the method computing different features and dissimilarity measures in two distinct datasets, under different classification tasks; and (iii) we show that our approach is robust and outperforms a more traditional baseline, accurately classifying media bias for binary scenarios and, more importantly, a multi-class scenario.

The rest of the paper is organized as follows: Section 2 summarizes the related work; Section 3 describes the steps that compose our method; Section 4 includes details about our experiments and results; and Section 5 presents our conclusions and future work.

2 RELATED WORK

Ideological bias has been detected by using different strategies, mostly based on the analysis of text content. Krestel et al. [13] used political text samples, such as speeches and statements, webpages of political parties and articles from news websites to characterize how the discourse of German news sources are similar to those of German parties.

Efron [7] discovers the political orientation of a web document by using co-citation data within a probabilistic model. Efron's model assesses the probability of co-citation between a set of reference documents, whose political orientation is well-known, and a target document, whose political orientation must be discovered. The decision is based on the premise that documents with stronger co-citation are more likely to be politically aligned.

Elejalde et al. [8] use tweets to automatically compute the political and socioeconomic orientation of Chilean media news portals, mapping the opinions expressed in tweets in a political survey to obtain the ideological bias of the portals. Ribeiro et al. [21] relied on ads to infer the political bias of news sources on social media such as Facebook and Twitter. The authors show that the ideological orientation (liberal or conservative) of a news source is related to the political preference of the audience.

Baly et al. [3] developed a method to predict factuality and bias of news media. They experimented with a varied set of features including lexical attributes to model headline and content of news articles, and information extracted from Twitter and Wikipedia. They showed that their approach is better suited to factuality (low, mixed and high) than to media bias, where they got a accuracy score of 41.74%. In this specific case, they performed two tasks: 3-way (left, center and right) and 7-way (extreme left, left, left center, center, right center, right and extreme right).

Directly related to political orientation, we have the political disaffection defined by Monti et al. [18] as "the lack of confidence in the political process, politicians, and democratic institutions, but with no questioning of the political regime." The authors show that the amount of tweets of disaffection along time is a strong indicator of political inefficacy. The detection of political disaffection, in this case, can be augmented with a bias detector to, among other things, understand the disaffection directed to a specific political ideology.

Word embeddings have also been used to detect political bias in tweets [10, 20]. In particular, Rao and Spasojevic [20] could define if a tweet leans towards the Democratic or Republican party with an accuracy as high as 87.57%. Gordon et al. [10] do not assess the performance of the classifier but use the word embedding to find that because of Trump's tweets, the Republican candidates category reaches a bias score of 0.97 (an indicator of the bias intensity with maximum value of 1.00).

The related work is summarized in Table 1. Most researches offer a case study or a very specific characterization, analyzing only a limited set of news sources. Methods with a more general approach like those of Efron [7] and Baly et al. [3] also have limitations. In the first case, the method does not perform well when classifying web pages with only a few hyperlinks. In the latter, the set of features is very large, 282 in total. This can have implications for processing time and explainability. Also, the majority of works focus on a bi-class problem, classifying only left and right-wing leanings.

Thus, the major difference of our work is that we focus on using Information Theory as a dimension reduction strategy to detect the political leaning of a news website (not social media) regarding the intensity (extreme vs. moderate) and ideology (left vs. right). As a result, we can have up to four classes: left, left center, right center, and right, which makes the task harder compared to bi-class

approaches. Noteworthy, when dealing with a three-class problem (harder than bi-class, but simpler than four-class problems) Baly et al. [3], reported an accuracy of 42% for their method.

3 POLITICAL LEANING DETECTOR USING A INFORMATION THEORY APPROACH

In this section, we describe the POLitical Leaning Detector (Poll), a novel method we propose to detect media bias in news websites by using Information Theory concepts, more specifically, Shannon entropy and statistical divergence. Our method is similar to TF-IDF, but the key difference is that we have a strategy to select the most useful terms to characterize the speech of each bias class, using entropy to quantify the importance of terms. Figure 1 summarizes the steps that compose our approach which are discussed in the next subsections.

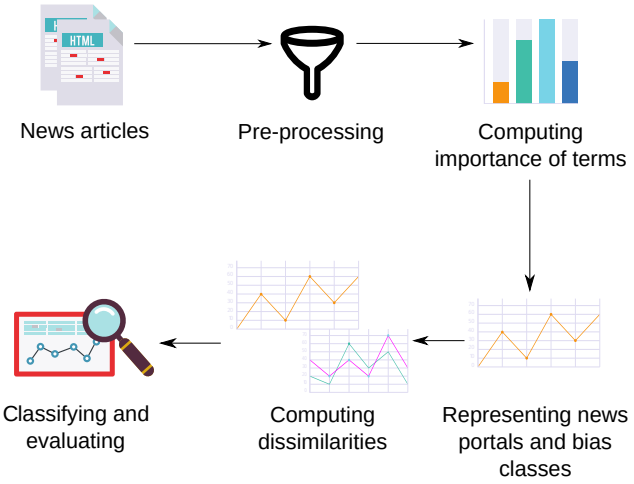


Figure 1: Overview of Poll (POLitical Leaning Detector), a information theory-based method to detect media bias in news websites.

3.1 News Articles and Preparing the Data

As a starting point for our method, we need a collection of news articles belonging to websites of known leaning/bias. We pre-process the data by transforming the text of the articles (from both title and content) to lower case, remove numbers, special characters and punctuation, and words that are not in English. This step is necessary because we identified some noise, like words similar to the name of functions in programming languages and HTML tags. Since we will quantify the importance of the terms in the next step, we do not remove stop words. If they are irrelevant to the context, the method will filter them.

3.2 Computing the Importance of Terms

To calculate the importance of terms in the vocabulary, we computed Shannon entropy [22], a quantifier from Information Theory that measures the amount of information carried by a variable (or randomness, from a statistical perspective). Given a probability mass function (pmf) $p = (p_1, p_2, \ldots, p_n)$ over a sample space of size n, i.e. $\left(\sum_{i=1}^{n} p_i\right) = 1$, the Shannon entropy is given by [22]

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1}$$

We use the Shannon entropy to quantify how useful a term is to distinguish two or more classes of bias by running through the following steps:

(1) Compute the frequency of all the terms in our reference corpus and discard the low-frequency ones (less than 10 in our datasets), as those terms might be *noisy terms*, which yields our vocabulary V.
(2) Given a problem of N classes[1], for each term $t \in V$ compute $p^{(t)} = (p_1^{(t)}, p_2^{(t)}, \ldots, p_N^{(t)})$, which is the pmf of the term t over the sample space of our bias classes.
(3) For each term $t \in V$, compute $H\left(p^{(t)}\right)$, which represents the importance of the term for distinguishing among the bias classes.
(4) Select a subset $V_R \subseteq V$, called vocabulary of reference, of the m most relevant terms, based on the $H\left(p^{(t)}\right)$ values computed in the previous step. The naive strategy is to keep the m terms of lowest entropy. This is the strategy we adopt in this paper, and the value of m is further specified in Section 4.

To understand how we use entropy in this work, let us check the two extreme cases. A term that evenly occurs across all the classes of our problem will have the maximum entropy $\log N$, and that term will be useless to distinguish among those classes (random occurrence). On the other extreme, a term that occurs only in a single class will have entropy zero (minimum), and that term will correctly identify the target class (assuming our sample corpus perfectly describes the reality).

Figure 2 shows an example comparing the entropy of the terms *trump* and *soros* in one dataset. In this example, we can see that *trump* is more evenly cited than *soros*, which is mostly concentrated within the class right. Thus, the term *soros* should be better than *trump* to distinguish among those four classes.

3.3 Representing a News Portal

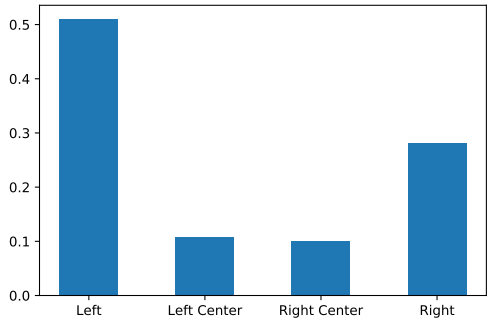
For every news portal/website w, we obtained a collection of articles/pages. In this work, we represent w by a pmf, in which each term $t \in V_R$ is mapped onto a bin of the pmf, representing the expected probability (normalized frequency) of t in an *average* article published in w.

Let us say that $|V_R| = n$. Given a website w with a corpus D of documents/articles collected from w, for every $t \in V_R$ and $d \in D$,

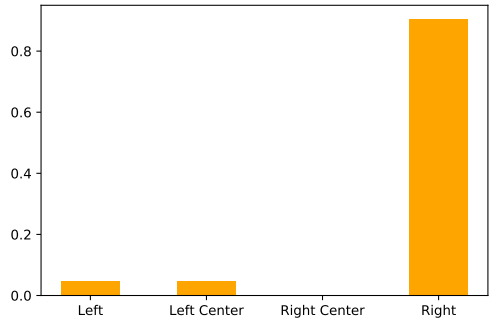
$$F_t^{(w)} = \sum_{d \in D, t \in d} f_{t,d}, \tag{2}$$

in which $f_{t,d}$ is the raw count (frequency) of t in d. Then, portal w is represented by the pmf $p^{(w)} = \left(p_1^{(w)}, p_2^{(w)}, \ldots, p_n^{(w)}\right)$, in which

[1]For instance, $N = 4$ for the problem of classifying a news portal as having a left, left center, right center, or right political orientation/bias.



(a) Citation of the term *trump*, referring to Donald Trump. The entropy score of this term was 1.69.



(b) Citation of the term *soros*, referring to George Soros. The entropy score of this term was 0.55.

Figure 2: Example of citation of two terms by sources belonging to each four bias classes in dataset News-July. The values are normalized.

$$p_t^{(w)} = \frac{F_t^{(w)}}{\sum\limits_{t' \in V_R} F_{t'}^{(w)}} \tag{3}$$

and, consequently, $\sum\limits_{t \in V_R} p_t^{(w)} = 1$.

3.4 Representing a Bias/Leaning Class

Now let us consider we have the bias/leaning classes represented by $B = \{b_1, b_2, \ldots, b_N\}$ in which every $b \in B$ is a class (e.g. left, left center, right center, and right, so that $N = 4$). We represent every class $b \in B$ by a pmf that is computed analogously to the pmf for each portal, but instead of using the documents of a target portal, we consider the documents for a target class.

Let us say that $|V_R| = n$. Given a class b with a corpus D_b of documents/articles collected from news portals of class b, for every $t \in V_R$ and $d \in D_b$,

$$F_t^{(b)} = \sum_{d \in D_b, t \in d} f_{t,d}, \tag{4}$$

in which $f_{t,d}$ is the raw count (frequency) of t in d. Then, class b is represented by the pmf $p^{(b)} = \left(p_1^{(b)}, p_2^{(b)}, \ldots, p_n^{(b)}\right)$, in which

$$p_t^{(b)} = \frac{F_t^{(b)}}{\sum\limits_{t' \in V_R} F_{t'}^{(b)}} \tag{5}$$

and, consequently, $\sum\limits_{t \in V_R} p_t^{(b)} = 1$.

Noteworthy, the classes represented by eq. (4) and (5) include only the documents of reference. The target documents represented by eq. (2) and (3) are not included in the computation of the $p_t^{(b)}$ pmfs that represent the bias classes.

3.5 Computing Dissimilarities Between News Portals and Bias Classes

After obtaining the pmfs for each news portal, we calculate a dissimilarity matrix that will model how different every news portal is with respect to every bias class.

For every news portal k and bias class b, we compute $D(k\|b)$, in which $D(\cdot\|\cdot)$ is a divergence, i.e., given a space of probability distributions S, with common support, $D(\cdot\|\cdot) : S \times S \to \mathbf{R}$ is a function such that

- $D(p\|q) \geq 0$, for all $p, q \in S$ and
- $D(p\|q) = 0$ if, and only if, $p = q$.

The objective of $D(\cdot\|\cdot)$ is to account the difference between two pmfs (shapewise). In this work, we consider three important divergences: Cosine distance, Jaccard distance, and Jensen-Shannon divergence.

The cosine distance is commonly used for Information Retrieval problems [2]. It is equivalent to the Pearson correlation, being proportional to the angle between two points in a vector space (sample space, in our case). The cosine distance between p and q is given by

$$\cos(p, q) = 1 - \frac{\sum\limits_{i=1}^{n} p_i q_i}{\sqrt{\sum\limits_{i=1}^{n} p_i^2} \sqrt{\sum\limits_{i=1}^{n} q_i^2}}. \tag{6}$$

The Jaccard distance [14, 16] between p and q is widely used in Biology domains and also in Computer Science to measure differences between vectors in $\mathbf{R}^n$ spaces, and it is given by

$$\text{jac}(p, q) = 1 - \frac{\sum\limits_{i=1}^{n} \min(p_i, q_i)}{\sum\limits_{i=1}^{n} \max(p_i, q_i)}. \tag{7}$$

The Jensen-Shannon divergence [15] between p and q relates to the concept of entropy and can be defined as

$$\text{jsd}(p,q) = \frac{H(p) + H(q)}{2} - H\left(\frac{p+q}{2}\right) \tag{8}$$

in which $H(\cdot)$ is the Shannon Entropy as defined in eq. (1). In general, the Jensen-Shannon divergence is a strong measure to account the difference between pmfs. In a simplistic way, the Jensen-Shannon divergence accounts the amount of bits that differs the pmfs being compared [15] and it is closely related to the concept of mutual information [9].

3.6 Classifying and Evaluating

After obtaining a dissimilarity matrix that accounts for the differences between the speech of each news portal and each class of bias, we feed a classifier with this matrix as features. This classifier will use these dissimilarity scores to distinguish the classes among each other.

4 EXPERIMENTS & RESULTS

In this section, we describe the results obtained when applying the method to two different datasets and detail the experimental setup used to evaluate the performance of the classifier.

4.1 Datasets

As discussed in Section 3, we need two types of data: news articles belonging to news websites with previously known political bias; and the assigned orientation (leaning or bias) of these sources. We used the labels obtained from Media Bias Fact Check (MBFC) [17], a fact-checking website that classifies news websites regarding ideological bias and credibility of factual reporting. Their methodology, although subjective, is based on a quantified system. They define five labels of political orientation/bias: left, left center, center, right center, and right. Figure 3 shows an example of a website labeled by MBFC. In this work, we will not consider websites from the center class, since we want to focus on a polarized field of discourse, specifically, left (extreme and moderate) and right-wing (extreme and moderate).

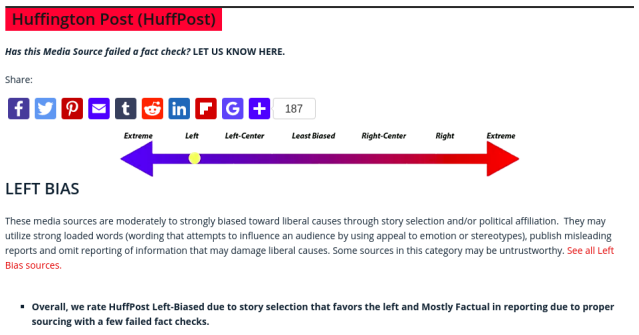


Figure 3: An example of a website labeled by Media Bias Fact Check [17].

Because most of the available news datasets include numerous articles from a few websites, we considered that they would not be beneficial to our application. Our aim is to classify political bias of news portals, so we maximized the number of websites in our tests. Thus, we decided to build datasets more appropriated for this task. Using the websites labeled by MBFC as seeds in a crawling process, we created two different datasets, **News-July** and **News-February**, collecting news articles from these seeds in different time spans. It is important to highlight that we did not restricted the crawl to a specific topic, i.e., we crawled news articles about arbitrary subjects. After crawling these two datasets, we sampled the articles to balance the number of websites belonging to each bias class and the number of articles of each source. We list the details of each dataset in Table 2.

Table 2: Datasets built by crawling seeds from Media Bias Fact Check [17].

Dataset	Time window	Number of websites	Articles per website	Total of articles
News-July	June 17-19, 2019	248	20	4960
News-February	February 14-15, 2020	576	20	11520

4.2 Experimental Setup

Once the data was gathered, we defined an experimental setup to tune and evaluate the performance of our method. There are some aspects to consider, such as model of the vocabulary, the dissimilarity metric, and the features. Our choices are discussed in the next paragraphs.

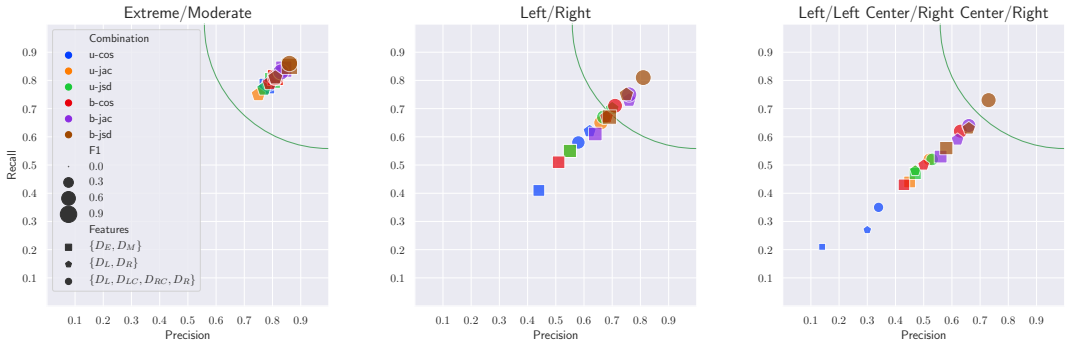
Modeling the vocabulary. Among the ways of modeling the terms in the text, we decided to test unigrams and bigrams. The idea is to analyze if a better representation of the context has a positive impact on the performance.

Selecting terms. Like explained in Section 3, we consider only the m terms of lowest entropy. We empirically determined $m = 10,000$ as the best value for our scenarios, based on the number of terms in each vocabulary for both News-July and News-February. Other datasets and domains might have a different value.

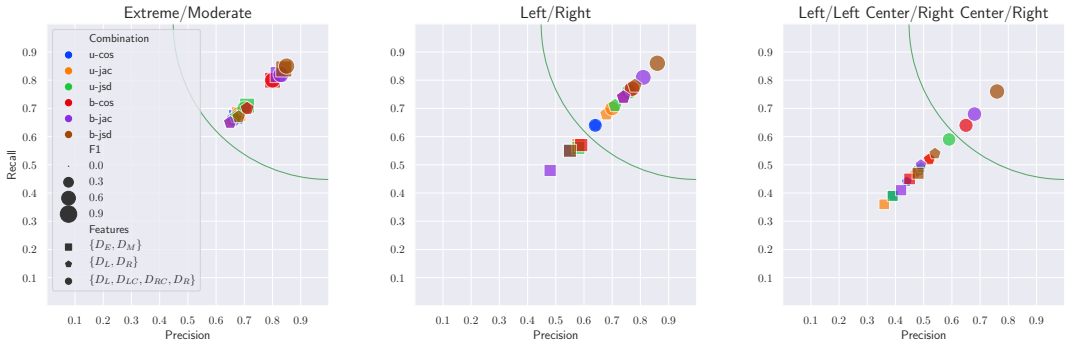
Dissimilarity measure. Given two pmfs in a sample space of size n (number of terms in our vocabulary of reference V_R), $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$, we assess the performance of three dissimilarity measures as our divergence $D(p\|q)$: the cosine distance, Jaccard distance, and Jensen-Shannon divergence, presented in Section 3.5.

Features. After computing the frequency histograms and having the dissimilarity measures, we define which sets of classes will compose the dissimilarity matrix. We chose the following alternatives to perform our tests:

- Extreme/Moderate (D_E, D_M): in this case, for every target document we compute two features: (1) D_E, the divergence of the document's pmf to the extreme (left and right) class' pmf; and (2) D_M, the divergence of the document's pmf to the moderate (left and right) class' pmf.
- Left/Right (D_L, D_R): in this case, for every target document we compute two features: (1) D_L, the divergence of the document's pmf to the left (extreme and moderate) class' pmf;



(a) Dataset News-July.



(b) Dataset News-February.

Figure 4: Performance obtained by Poll using different strategies in terms of precision, recall and F1.

and (2) D_R, the divergence of the document's pmf to the right (extreme and moderate) class' pmf.

- Left/Left Center/Right Center/Right (D_L, D_{LC}, D_{RC}, D_R): fin this case, for every target document we compute four features: (1) D_L, the divergence of the document's pmf to the extreme left class' pmf; (2) D_{LC}, the divergence of the document's pmf to the moderate left class' pmf; (3) D_{RC}, the divergence of the document's pmf to the moderate right class' pmf; and (4) D_R, the divergence of the document's pmf to the extreme right class' pmf.

Tasks. We are interested in comparing our method, Poll, with three different classification tasks:

- Extreme/Moderate.
- Left/Right.
- Left/Left center/Right center/Right.

With these tasks, we can evaluate if the discourse of extreme sources is more similar than the ones of more moderate sources. We can also evaluate if using the corresponding set of features leads to better results when performing each task.

Classifier and evaluation. For the classifier, we chose the Support Vector Machine (SVM) model with RBF kernel, $C = 1.0$ and remaining parameters set to the default of the scikit-learn library[2].

We conducted the experiments using leave-one-out cross validation (LOOCV) and computed the metrics: precision, recall, F1, and accuracy.

Last, we compared Poll with the method proposed by Baly et al. [3] described in Section 2 as a baseline. This method represent news articles by calculating a set of 141 features like POS tags, sentiment scores, bias, subjectivity, and morality. They compute these features for both title and body, which leads to a set of 282 features in total, that are given as input to a supervised method, specifically, a SVM classifier. To perform the experiments, we implemented the method using the code shared by the authors[3] to model and classify our two datasets. The performance was evaluated applying the same setup described to evaluate Poll, i.e., leave-one-out cross validation (LOOCV) and using precision, recall, F1 and accuracy as performance measures.

We selected this baseline because, like our approach, they focus on automatic detecting media bias of news websites; they also do not restrict articles to a single subject and period of time, and they use several features that are more common in text classification, more specifically, in fake news detection. In addition, this baseline was designed to work with multi-class problems as well. Thus, we can determine if our method can perform better than a more traditional method by applying a smaller set of features independent

[2]https://scikit-learn.org/stable/

[3]https://github.com/ramybaly/News-Media-Reliability/

Table 3: Classification results for two datasets when performing three classification tasks. Best results for each task are bold.

(a) Performance for Extreme/Moderate task.

Method	Dataset	Performance						
		Precision		Recall		F1		Accuracy
		Extreme	Moderate	Extreme	Moderate	Extreme	Moderate	
Baly et al. [3]	News-July	0.42	0.44	0.36	0.49	0.39	0.46	43%
	News-February	0.43	0.43	0.44	0.42	0.43	0.42	43%
Poll	News-July	**0.87**	**0.84**	**0.84**	**0.88**	**0.86**	**0.86**	**86%**
	News-February	**0.86**	**0.83**	**0.83**	**0.87**	**0.84**	**0.85**	**85%**

(b) Performance for Left/Right task.

Method	Dataset	Performance						
		Precision		Recall		F1		Accuracy
		Left	Right	Left	Right	Left	Right	
Baly et al. [3]	News-July	0.50	0.50	0.48	0.53	0.49	0.52	50%
	News-February	0.42	0.44	0.37	0.50	0.39	0.47	43%
Poll	News-July	**0.84**	**0.78**	**0.76**	**0.85**	**0.80**	**0.82**	**81%**
	News-February	**0.86**	**0.87**	**0.87**	**0.86**	**0.86**	**0.86**	**86%**

(c) Performance for Left/Left Center/Right Center/Right task.

Method	Dataset	Performance												
		Precision				Recall				F1				Accuracy
		Left	Left Center	Right Center	Right	Left	Left Center	Right Center	Right	Left	Left Center	Right Center	Right	
Baly et al. [3]	News-July	0.29	0.26	0.34	0.25	0.26	0.31	0.34	0.23	0.27	0.28	0.34	0.24	28%
	News-February	0.26	0.22	0.19	0.22	0.30	0.22	0.19	0.19	0.28	0.22	0.19	0.21	22%
Poll	News-July	**0.77**	**0.75**	**0.63**	**0.79**	**0.79**	**0.66**	**0.73**	**0.74**	**0.78**	**0.70**	**0.68**	**0.77**	**73%**
	News-February	**0.82**	**0.73**	**0.69**	**0.80**	**0.78**	**0.73**	**0.77**	**0.75**	**0.80**	**0.73**	**0.73**	**0.77**	**76%**

of context. Noteworthy, although the method proposed by Baly et al. [3] incorporates other sources of information and different tasks, the authors also allow to execute the method using only features extracted from the articles (title and content), which is exactly our context. Also, it provides the possibility to choose other bias classes. So, we can use these settings to compare the baseline directly to our proposed method, and verify how it compares to a traditional method while using information obtained from news articles only, i.e., without relying on external sources.

4.3 Experiment 1: Variations of Poll

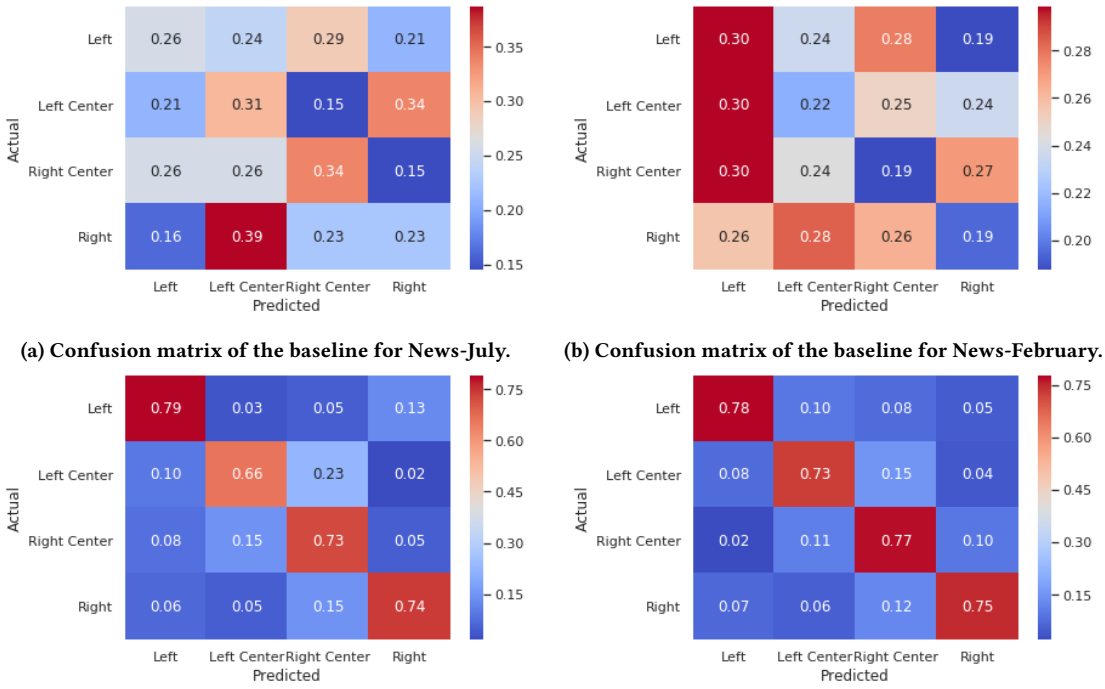
In the first experiment, we compare the results obtained by Poll when using different vocabulary models, dissimilarity measures, and sets of features for each classification task and dataset. Figure 4 illustrates the performances in terms of precision, recall, and F1. In this figure, each method refers to combining unigrams (u) or bigrams (b) with a divergence: cosine distance (cos), Jaccard distance (jac) and Jensen-Shannon divergence (jsd).

Comparing the performances obtained by unigrams and bigrams, we see that bigrams outperformed unigrams. This result indicates that a better representation of the context leads to more representative probability mass functions and a better characterization of the discourses of the bias classes.

In terms of the experimented divergences, the results highlight some key differences in performance. In general, the combinations that used cosine distance got the lower balances between precision, recall and F1. Computing Jaccard distance and Jensen-Shannon divergence led to similar results, but Jensen-Shannon performed better in all cases. These results make sense, since we are working with probability mass functions: Jensen-Shannon divergence is more sensitive to differentiate between these distributions, thus being more suitable than more common metrics.

In terms of which task was easier or more challenging, we see that classifying extreme/moderate sources was easier, followed by left/right and left/left center/right center/right as the most challenging. This is illustrated in each plot, where the green semi-circle highlights that classifying extreme/moderate resulted in a better classification, leading to similar values of precision, recall, and F1. In comparison, the performances when classifying left/right and four-classes were more nuanced. This confirms the intuition that a multi-class problem is more challenging than the binary cases, and that the speeches of extreme sources are more similar between them than to those of moderate sources.

Comparing the three sets of features, the results show that using dissimilarities based in four classes was the best strategy in all tasks for both datasets. Besides that, there seems to be a correlation



(a) Confusion matrix of the baseline for News-July.

(b) Confusion matrix of the baseline for News-February.

(c) Confusion matrix of Poll for News-July.

(d) Confusion matrix of Poll for News-February.

Figure 5: Confusion matrices obtained by each method when performing the multi-class problem. Values are normalized by rows.

between the task and the corresponding set of features. When performing the extreme/moderate task, using extreme/moderate divergences led to a good balance of precision, recall and F1. The same occurred when performing the left/right task using left/right features. In the third class, using the four features set was the best strategy, especially in the News-February dataset.

On a side note, Figure 4 also shows that Poll, regardless the term size and the divergence being used is well balanced regarding precision and recall, because the points are close to a implicit 45° line between the precision and recall axes.

From this first experiment, we conclude that: (i) using bigrams as the vocabulary model leads to better results; (ii) the best dissimilarity measure for this context was Jensen-Shannon divergence; (iii) there seems to be a relation between the task and the set of features, but the best choice for all tasks was to compute left/left center/right center/right dissimilarities and (iv) the best combination for Poll is: bigrams as terms, Jensen-Shannon divergence as the divergence, and D_L, D_{LC}, D_{RC}, D_R as features. For now on, when we refer to Poll, we are referring to this specific combination. This result is used to compare Poll to the baseline in the next experiment.

4.4 Experiment 2: Poll vs. the Baseline

After the first experiment, we verified how Poll, combined with bigrams, Jensen-Shannon divergence, and four-class features, performs against a more traditional baseline. Our tests included the three classification tasks. Table 3 summarizes the results.

In the first task (Table 3a), where we classify extreme and moderate sources, the results show that the baseline has low precision, recall, F1, and accuracy. This means that the method is not able to distinguish between the classes, especially in the case of dataset News-July, where it had more confusion between extreme and moderate sources. Poll, in contrast, achieved high scores and balanced results in precision, recall, and 0.86 of F1 for both classes, indicating that our features were able to discriminate well between classes. In absolute terms, our method performed almost twice as better than the baseline for all performance metrics, achieving a macro accuracy of 86% versus 43% obtained by the baseline in the same situation.

Similarly, in the second task (Table 3b), in which we classify left and right sources, the results show that the baseline was more successful classifying the right class than the left class. But even so, precision, recall, F1, and accuracy were low (close to 0.50). Again, Poll performed almost twice as better than the baseline, with balanced results for both classes (F1 equal to 0.86) and a macro accuracy score of 86% against 43% achieved by the baseline in the same case.

The four-class problem (Table 3c), classifying all four classes of bias, is where our method really stands out. Figure 5 shows the confusion matrices obtained by each method for each dataset. The baseline had trouble to distinguish between the four bias classes, performing poorly. Poll, on the other hand, was able to distinguish between the four bias classes. Our method performed better when classifying both extreme classes (left and right), with F1 equal to 0.80 and 0.77 for these classes in the best case. But even to the

moderate classes, that tend to be more similar between them, we got good results, with F1 of 0.73 for both left center and right center classes in the same case. Compared to the baseline, Poll performed almost 3.5 times better, achieving a macro accuracy score of 76% versus 22% in the same dataset.

These results are related to the strategy each method applies to represent news websites and bias classes. The baseline uses 282 textual features that are probably very similar for all four bias classes. So, they are not very useful to characterize the discourse of each ideological bias/orientation. Our strategy, on the other hand, reduces the number of features by focusing on capturing particularities of the discourses of each bias class. The results show that this strategy leads to more representative features, allowing a classifier to accurately distinguish the four bias classes.

So, with the second experiment, we conclude that: (i) Poll was successful in accurately classifying two binary problems and a multi-class problem; (ii) Poll outperformed the baseline with balanced scores of precision, recall, and F1, reaching accuracy scores above 73% (multi-class) and 81% (binary), using only four features, against 282 from the baseline whose accuracy was as low as 22%.

5 CONCLUSIONS & FUTURE WORK

In this paper, we presented Poll (POLitical Leaning Detector), a new approach to detect media bias in news websites. Our approach applies concepts from Information Theory to quantify the importance of terms in news articles and better characterize the speech of websites with a particular ideological leaning.

To evaluate the effectiveness of the method, we performed experiments to classify two datasets composed by news of different periods and discussing several topics, without restriction to a single subject. Thus, we showed that our approach accurately classifies the bias of news websites in three different situations: separating more extreme and more moderate sources; left and right sources; and a more detailed classification, separating four classes of bias (left, left center, right center, and right sources). We observed that our method outperformed a more traditional approach that uses 282 textual features like sentiment scores and POS tags, achieving accuracy scores 2 to 3 times higher than the baseline. Furthermore, our approach obtains these results using a set of only 2–4 features. This result shows that our proposed method effectively captures the particularities of the discourse used by websites of each political bias/orientation.

As future work, we plan to explore other strategies to select terms. Instead of using a fixed number of terms, we can investigate more sophisticated possibilities and analyze the impact of these choices on the final classification. Also, we intend to check how the method performs when classifying other collections of news, like past news and more recent news, and also news about specific topics.

REFERENCES

[1] Victória Patrícia Aires, Fabíola G. Nakamura, and Eduardo Freire Nakamura. 2019. A Link-based Approach to Detect Media Bias in News Websites. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 742–745. https://doi.org/10.1145/3308560.3316460

[2] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[3] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium.

[4] Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics* 76 (Jan. 2015), 87–100. https://doi.org/10.1016/j.pragma.2014.11.010

[5] Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tilburg University, The Netherlands, 79–88. https://doi.org/10.18653/v1/W18-6509

[6] Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media bias in german online newspapers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 133–137.

[7] Miles Efron. 2004. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. ACM, 390–398.

[8] Erick Elejalde, Leo Ferres, and Eelco Herder. 2017. The nature of real and perceived bias in chilean media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 95–104.

[9] D. M. Endres and J. E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory* 49, 7 (July 2003), 1858–1860. https://doi.org/10.1109/TIT.2003.813506

[10] Joshua Gordon, Marzieh Babaeianjelodar, and Jeanna Matthews. 2020. Studying Political Bias via Word Embeddings. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 760–764. https://doi.org/10.1145/3366424.3383560

[11] Ruth A Harper. 2010. The Social Media Revolution: Exploring the Impact on Journalism and News Media Organizations. 2, 03 (2010). http://www.inquiriesjournal.com/a?id=202

[12] Markus Knoche, Radomir Popović, Florian Lemmerich, and Markus Strohmaier. 2019. Identifying Biases in Politically Biased Wikis through Word Embeddings. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (Hof, Germany) *(HT '19)*. Association for Computing Machinery, New York, NY, USA, 253–257. https://doi.org/10.1145/3342220.3343658

[13] Ralf Krestel, Alex Wall, and Wolfgang Nejdl. 2012. Treehugger or petrolhead?: identifying bias by comparing online news articles with political speeches. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 547–548.

[14] Michael Levandowsky and David Winter. 1971. Distance between sets. *Nature* 234, 5323 (1971), 34–35.

[15] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.

[16] Alan H Lipkus. 1999. A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry* 26, 1-3 (1999), 263–265.

[17] Media Bias Fact Check. 2019. *The Most Comprehensive Media Bias Resource*. Accessed May, 2020 from https://mediabiasfactcheck.com/.

[18] Corrado Monti, Alessandro Rozza, Giovanni Zappella, Matteo Zignani, Adam Arvidsson, and Elanor Colleoni. 2013. Modelling Political Disaffection from Twitter Data. In *Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*.

[19] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying Framing Bias in Online News. *ACM Transactions on Social Computing* 1, 2 (2018), 5.

[20] Adithya Rao and Nemanja Spasojevic. 2016. Actionable and Political Text Classification using Word Embeddings and LSTM. In *Proceedings of the 5th International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*.

[21] Filipe N Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth International AAAI Conference on Web and Social Media*.

[22] Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 3 (1948), 379–423.