A Link-based Approach to Detect Media Bias in News Websites

Victoria Patricia Aires
Institute of Computing – Federal University of Amazonas (UFAM)
Manaus, AM
victoria.aires@icomp.ufam.edu.br

Fabiola G. Nakamura
Institute of Computing – Federal University of Amazonas (UFAM)
Manaus, AM
fabiola@icomp.ufam.edu.br

Eduardo F. Nakamura
Institute of Computing – Federal University of Amazonas (UFAM)
Manaus, AM
nakamura@icomp.ufam.edu.br

ABSTRACT

News websites are currently one of the main sources of information. Like traditional media, these sources can have a bias in how they report news. This media bias can influence how people perceive events, political decisions, or discussions. In this paper, we describe a link-based approach to identify news websites with the same political orientation, i.e., characterize the bias of news websites, using network analysis techniques. After constructing a graph from a few seeds with previously known bias, we show that a community detection algorithm can identify groups formed by sources with the same political orientation.

CCS CONCEPTS

• **Information systems → Web applications**; *Data extraction and integration.*

KEYWORDS

media bias detection; news analysis; network analysis

ACM Reference Format:
Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. 2019. A Link-based Approach to Detect Media Bias in News Websites. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316460

1 INTRODUCTION

Nowadays, due to the popularization of the Web and social media, news websites have become one of the main sources of information [5]. However, like in traditional media, these portals have a bias in how they report news. By using aspects such as selective omission and choice of words, each source conveys a different impression of a fact [8]. This may impact the way the audience perceives events, political decisions and discussions regarding several topics. Studies have shown, for example, that media bias can influence the outcome of elections [4]. Detecting media bias is very hard for humans because of the inherent subjectivity involved [14]. Therefore, developing automatic methods for this purpose is an interesting research direction.

In this context, studies have been developed to automate the process of detecting media bias in news websites, most of them focusing on analyzing the textual content of the webpages [5, 11, 12]. However, news webpages form a link structure that can be explored, possibly revealing relationships between news websites. The link structure has been successfully studied in problems such as web spam and search engine rankings [7, 10], revealing that similar webpages form groups on the Web.

In this paper, we present a preliminary study to characterize the bias of news websites, i.e., identify portals with the same political orientation, using a link-based approach. The focus is on politics because the bias of news websites was previously classified by fact checking websites such as Media Bias Fact Check [3]. In our approach, we construct a graph with data crawled from a few seeds whose bias is known. Then, we apply network analysis tools such as community detection algorithm and centrality measures. We analyze the resulted graph, showing that websites with the same orientation form groups on the Web through links established between them.

The remainder of the paper is organized as follows. Section 2 includes works that focused on media bias detection on websites. In Section 3, we describe the steps of our approach. In Section 4, we report our experiments and results. Finally, Section 5 contains our conclusions and opportunities for future work.

2 RELATED WORK

In this section, we describe studies that address problems such as bias detection and political orientation detection in online news.

Efron [6] introduced a method for estimating political orientation of hypertext documents using cocitation information employing a probabilistic model and evaluating the likelihood of cocitation between the documents and a few seeds, i.e., documents of known orientation. Results showed that the model outperformed lexically based classifiers such as naïve Bayes and SVM.

Dallmann et al. [5] performed an analysis to identify political orientation of German online newspapers. They proposed measures that indicate a bias towards a political party, including mentions to a party and the sentiment associated with them. The results showed that the analyzed newspapers have bias towards specific parties, which is consistent with the public perception.

Morstatter et al. [11] focused on identifying framing bias, where specific aspects of a story are reinforced over others. They studied the ability of a machine learning classifier to detect frames and polarity in sentences of a news corpus. The results showed that simple linguistic features and the use of *n*-grams performed best in finding frames in text.

Niculae et al. [12] proposed an unsupervised framework based on quoting patterns for analyzing how the media selects what to cover in an article. They applied this framework to a dataset of political news and presidential speeches. They showed that the main dimension of bias align with the ideological spectrum and

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316460

source type, exposing differences in how different sources portray reality.

These works are mostly focused on the text of the articles, analyzing aspects such as vocabulary. In this paper, we intend to use a link-based approach to identify news websites with the same political orientation. Our main innovation is using network analysis tools such as community detection to characterize groups with the same political bias, something unexplored by previous works.

3 METHODOLOGY

In this section, we describe our approach to detect the bias of a news website. Below, we discuss each step of the proposed approach.

3.1 Selecting News Websites

The first step is selecting websites to be seeds in a web crawling process. We need to know in advance the bias of the seeds. Thus, we used Media Bias Fact Check (MBFC) [3]. MBFC is a fact checking site that rates websites based on ideological bias and credibility of factual reporting. Their methodology is subjective, but based on a numeric scoring system to assign labels. The possible labels for political bias are: Left, Left Center, Center, Right Center and Right. They are assigned on domain level, so every article originating from the same source will have the same label. For each class of political bias, we choose four websites to represent them as our seeds. These websites are listed in Table 1.

Table 1: Websites selected as our seeds and their respective bias, as determined by Media Bias Fact Check (MBFC) [3].

Bias	Website
Left Bias	https://www.cnn.com/
	https://www.huffingtonpost.com/
	http://nymag.com/
	https://theintercept.com/
Left Center Bias	https://www.bbc.com/
	https://www.latimes.com/
	https://www.nytimes.com/
	https://www.washingtonpost.com/
Center	https://www.reuters.com
	http://apnews.com
	https://www.politico.com/
	https://www.desmoinesregister.com/
Right Center Bias	https://www.forbes.com/
	https://nypost.com/
	https://www.thetimes.co.uk/
	https://www.wsj.com/
Right Bias	https://www.foxnews.com/
	https://observer.com/
	https://dailycaller.com/
	https://www.thesun.co.uk/

3.2 Crawling Pages and Creating the Graph

After selecting the seeds, the second step is to implement the web crawler to collect the data required to construct the graph. We implemented the web crawler using Python 3.7 and Scrapy, a web crawling framework [13]. The websites listed in Table 1 were given as input to the crawler. As visiting these URLs, it identifies all the hyperlinks in the page, and visits these links next. The process continues recursively until the crawler reaches 4 levels deep from the starting page in the Web graph, or the list of links to visit ends.

When visiting a webpage, we use the links to create the graph. We add in a CSV file an entry representing the source and the destiny of an edge. The source is the domain of the webpage we are visiting, and the destiny is the domain of the webpage it links to. However, we don't add an entry where the source and the destiny are the same website, filtering internal links. Another filter is removing links to social networks and advertisements of major brands.

At the end of the process, the CSV file is populated with several entries representing connections between websites. In this file, each domain is a node, and each entry is an edge. Repeated entries increase the edge weight, resulting in a directed and weighted graph.

3.3 Analyzing the Graph

The third step is to analyze the graph previously generated. Here, we used Gephi, a network analysis software [1]. The graph, represented in the CSV file, is given as input to Gephi. Before the analysis, we applied some filters. First, we merged entities from the same newspaper. For example, Page Six is the entertainment column of New York Post. In our analysis, we consider them as part of the same publication. Second, we filtered nodes with degree lesser than 10. After this filter, a new graph is generated without the removed nodes and their links.

Subsequently, we applied a community detection algorithm called Louvain method. The algorithm identifies groups in the graph that are more connected between them than with the rest of the graph [2] and it is based only on network topology, i.e., considering only the link structure of the network. We also applied HITS (Hyperlink-induced Topic Search), an algorithm that measures the centrality of the nodes in the network using scores called hubs and authorities [10], to identify the most influential nodes in each community. We focused on hub scores, which indicates the nodes that make links with higher relevance.

4 EXPERIMENTS & RESULTS

In this section, we report the results obtained after applying our methodology. After the web crawling process, the resulted graph had 3274 nodes and 4782 edges. Below, we describe two experiments. In the first, we applied the filters we described in Section 3. However, we noticed that the graph contains some websites that weren't labeled by MBFC. So, in the second experiment we analyze the graph composed only by websites whose labels are known. After describing both experiments, we discuss our results.

4.1 Experiment 1: Applying Basic Filters

In this experiment, we applied the filters we described in our methodology. This resulted in a new graph containing 98 nodes and 743 edges, illustrated in Figure 1. Five communities were detected, represented by colors in the figure. Below, we describe the communities, considering unlabeled websites. Table 2a shows the number of websites of each bias that occurred in each community.

Table 2: Distribution of websites and biases in our experiments.

(a) Experiment 1.

	Left	Left Center	Center	Right Center	Right	Unlabeled
Community 1	11	17	7	4	1	9
Community 2	2	9	5	3	4	9
Community 3	-	2	-	-	-	1
Community 4	-	1	-	2	1	7
Community 5	-	-	-	-	-	3

(b) Experiment 2.

	Left	Left Center	Center	Right Center	Right
Community 1	9	8	5	2	1
Community 2	4	15	4	2	2
Community 3	-	4	-	3	2
Community 4	-	1	3	-	-
Community 5	-	1	-	2	1

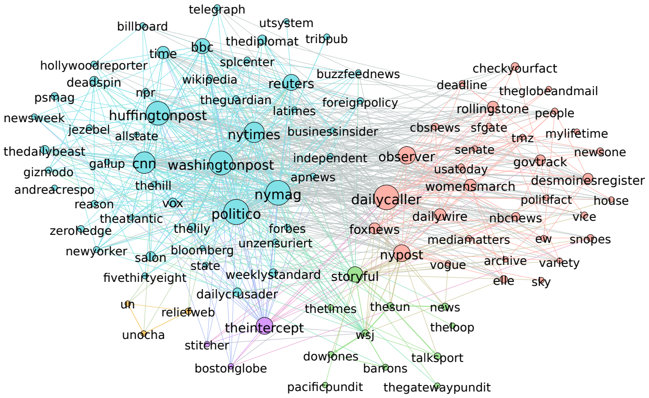

Figure 1: The resulted graph after applying our approach. The colors represent the communities detected and the size of the nodes corresponds to their hub scores.

Community 1 (blue). The main nodes, according to hub scores, are Politico, New York Magazine and Washington Post. These portals have, respectively, Center, Left and Left Center bias. The community is mainly composed by left-wing websites (Left and Left Center bias). Between the unlabeled websites, we highlight Daily Crusader, a junk news website, i.e., a portal that publishes low quality news using serious journalism to support their claims [9]. Storyful linked to this website in an article about questionable content. Daily Crusader is in this community because it links to credible sources such as New York Times, Wikipedia and Washington Post, which are also in this group.

Community 2 (red). In this community, the main nodes are The Daily Caller, New York Observer and New York Post. They have Right and Right Center bias, respectively. The community has almost the same number of left- and right-wing portals. However, as we mentioned earlier, right-wing nodes have most influence in this community. We believe the mixing occurred due to the tendency of these portals to criticize other of the opposite bias. Between the unlabeled websites, we highlight Women's March. The Daily Caller cited this website to criticize the movement and some members. On the other hand, Huffington Post had some articles communicating about the movement in a positive way. This shows how the sources report facts accordingly to their orientation. It also shows how right-wing portals criticize sources related to the left-wing.

Community 3 (purple). This was a small community, containing only The Intercept and The Boston Globe, two Left Center websites; and an unlabeled website that is an advertising of a small company. The presence of this website probably influenced the Louvain algorithm, which is greedy and based only on links, leading to an isolated community.

Community 4 (green). The main nodes in this community are Storyful, news.com.au and Wall Street Journal. The first is unlabeled; the second has Left Center bias, and the latter has Right Center bias. Most of the websites in this community are right-wing portals. Between the unlabeled, there are two websites classified by MBFC as Questionable Sources, i.e., sources that exhibits extreme bias, promotion of conspiracies and/or fake news [3]. These websites were linked by Storyful, which is in this community, in two articles about unreliable portals, acknowledging their suspicious nature.

Community 5 (yellow). Another small community, it is composed by unlabeled websites only. However, it's interesting to observe that they are humanitarian websites, such as United Nations and ReliefWeb. They have links only between them, showing their independence of news portals and therefore justifying the community.

4.2 Experiment 2: Filtering Unlabeled Websites

In this experiment, we wanted to check the impact, regarding to community detection, of removing unlabeled websites. The removal resulted in a new graph with 69 nodes and 575 edges. Five communities were detected, as in Experiment 1. Figure 2 illustrates the graph and the colors indicate the communities, and Table 2b shows the number of websites of each bias in each community. Similarly to Section 4.1, we describe the communities below.

Community 1 (blue). In this community, the main nodes are New York Magazine, Politico and Washington Post, respectively of Left, Center and Left Center biases. It contains mostly left-wing websites, with few right-wing sources, demonstrating a good grouping. We highlight that The Intercept and The Boston Globe, which had their own community in the previous experiment, were grouped with other left-wing websites after filtering unlabeled websites.

Community 2 (red). The main nodes in this community are The Daily Caller, Huffington Post and New York Times. The first has Right bias and the two latter, Left and Left Center bias, respectively. This community has majorly left-wing websites, and the hub scores also indicates that these are the most influential nodes. Again, the criticism of right-wing websites (linking to sources of the opposite bias) may have influenced in this grouping.

Community 3 (pink). Here, the main nodes are New York Observer, New York Post and Fox News, two of Right Center bias and one of

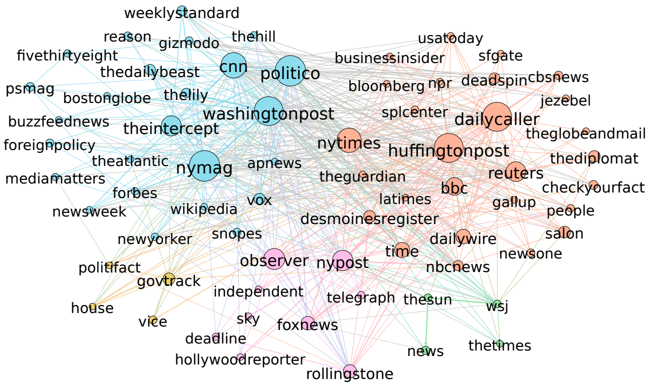

Figure 2: The graph after filtering unlabeled websites. The colors represent the communities detected and the size of the nodes corresponds to their hub scores.

Right. The community is mostly composed of right-wing websites. The websites of Left Center bias are mostly entertainment news portals, which can be interesting to the audience of the right-wing websites, justifying the links and, therefore, the community.

Community 4 (green). This community is very similar to Community 4 of Experiment 1, with the same nodes, excluding unlabeled websites. Like the other, this community has mainly right-wing websites and the main nodes are news.com.au, The Sun and Wall Street Journal.

Community 5 (yellow). This is a small community, however, it's interesting to observe that it's composed mainly of Center websites. The main nodes are GovTrack, House.gov and Politifact. The first two are related to the United States Congress, and the latter is a fact checking websites dedicated to U.S. politics. So, this community grouped the least biased sources that are related to American politics.

In both experiments, five communities were detected. In the first experiment we observed a mix in the communities, with most of the right-wing websites grouped with left-wing sources. In the second experiment, after removing unlabeled websites from the analysis, the detected communities were more discriminative. Each one of them had a majority belonging to one of the bias classes, namely: Left and Left Center (Community 1 and Community 2); Right and Right Center (Community 3 and Community 4); and Center (Community 5). We believe that these results indicate that a link-based approach can be a promising technique to characterize the bias of websites, leading to an interesting characterization that reflected what would be expected based on human judgement.

5 CONCLUSION & FUTURE WORK

In this paper, we presented a link-based approach to detect the bias of news websites. With only a few seeds of each bias class, we could construct a graph representing the connections between news portals. In this graph, we applied a community detection algorithm, based only on network topology, to identify the groups and check if they are composed of websites with the same political orientation.

The results showed that the communities are related to the bias of the websites, indicating that websites with similar biases tend to establish connections between them. We also noted that the communities reflected the common sense regarding the portals that were grouped together. Thus, a link-based approach can be effective to identify if a determined source is aligned with others.

However, since the results described are part of a preliminary study, there is room for improvements. In our experiments, some communities mixed sources of different bias due to links between them, established mainly by right-wing portals criticizing left-wing portals. Based on this, we intend to expand the study developing, in a future work, an approach to analyze the sentiment associated to the article or to the citation. This can be useful to determine if the websites have a similar point of view, therefore improving the link-based method.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and it is part of Process 88887.150732/2017-00 PGCI - Call 02/2015, held in cooperation by the Department of Computer Science and Engineering (CSE) at New York University (NYU) and by the Instituto de Computação (IComp) at the Universidade Federal do Amazonas (UFAM).

REFERENCES

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Media Bias Fact Check. 2019. The Most Comprehensive Media Bias Resource. Accessed January 24, 2019 from https://mediabiasfactcheck.com/.

[4] Chun-Fang Chiang and Brian Knight. 2011. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies* 78, 3 (2011), 795–820.

[5] 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.

[6] 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.

[7] Dennis Fetterly, Mark Manasse, and Marc Najork. 2004. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*. ACM, 1–6.

[8] Matthew Gentzkow and Jesse M Shapiro. 2006. Media bias and reputation. *Journal of Political Economy* 114, 2 (2006), 280–316.

[9] C Jensen. 2001. Junk Food News 1877-2000. *Phillips, P.(2001) Censored* 2001 (2001), 251–264.

[10] Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.

[11] 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.

[12] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 798–808.

[13] Scrapy. 2019. A Fast and Powerful Scraping and Web Crawling Framework. Accessed January 19, 2019 from https://scrapy.org/.

[14] Sevgi Yigit-Sert, Ismail Sengor Altingovde, and Özgür Ulusoy. 2016. Towards detecting media bias by utilizing user comments. In *Proceedings of the 8th ACM Conference on Web Science*. ACM, 374–375.