### Analysis of Media Writing Style Bias through Text-Embedding Networks\*

Iain J. Cruickshank, <sup>1</sup> Jessica Zhu, <sup>1, 2</sup> Nathaniel D. Bastian <sup>1</sup>

<sup>1</sup> Army Cyber Institute, United States Military Academy, West Point, NY
<sup>2</sup> University of Maryland, Baltimore, MD
iain.cruickshank@westpoint.edu, jeszhu@umd.edu, nathaniel.bastian@westpoint.edu

#### **Abstract**

With the rise of phenomena like 'fake news' and the growth of heavily-biased media ecosystems, there has been increased attention on understanding and evaluating media bias. Of particular note in the evaluation of media bias is writing style bias, which includes lexical bias and framing bias. We propose a novel approach to evaluating writing style bias that utilizes natural language similarity estimation and a networkbased representation of the shared content between articles to perform bias characterization. Our proposed method presents a new means of evaluating writing style bias that does not rely on human experts or knowledge of a media producer's publication procedures. The results of experimentation on realworld vaccine mandate data demonstrates the utility of the technique and how the standard bias labeling procedures of only having one bias label for a media producer is insufficient to truly characterize the bias of that media producer.

#### Introduction

The rise of fake news and heavily biased media ecosystems has contributed to many societal ills across the world. For example, biased media ecosystems have led to increased polarization in society and destructive phenomena like 'truth decay' (Kavanagh and Rich 2018). As such, there is distinct importance to better understanding and evaluating media bias. When it comes to evaluating media bias, most citizens rely on third-party bias estimations, such as Adfontes or mediabiasfactheck.org. While these websites and the people behind them often do provide valid bias labels, they also require subjective interpretations of the media producers, as well as culturally specific evaluation procedures which may not translate to other media ecosystems (e.g., the methods used to evaluate bias in the United States may not translate to other nations, like India, for example), and can only scale at the rate that humans can evaluate media websites. As such, there has been increasing interest to find more computational approaches to evaluating media bias.

Of particular note in the evaluation of media bias is writing style bias, which includes lexical bias and framing bias. Various computational approaches exist for detecting writing style bias in media texts, with Natural Lan-

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

guage Processing (NLP) techniques being the most common. However, most approaches are unsupervised due to the lack of labeled data sets and the difficulty in obtaining non-skewed labels. Despite this, recent studies have proposed supervised machine learning models for detecting informational bias, which has shown that additional context surrounding informationally-biased sentences aids in detection. Network-based approaches for analyzing writing style bias have also been proposed, including Centering Resonance Analysis and graph-theoretic approaches. However, most of these methods construct graphs at the individual text level rather than between different texts. It is also important to consider content redistribution in news production via news agencies, where media sources frequently share textual content. This phenomenon must be taken into account when comparing the writing style biases of different media sources.

In this paper, we propose a novel approach for detecting writing style bias in media texts that takes into account the shared content between different media sources. Our approach applies natural language similarity estimation and a network-based representation of the shared content between articles to perform bias characterization, and we evaluate its performance on a data set of news articles. We show that our proposed method can produce meaningful characterizations of writing style bias across any number of articles and media producers. We do so without having any knowledge of how those media domains produce their articles, which is frequently required for bias estimation by human subjective means. As part of our analysis, we also find that media domains' exhibited writing style biases can vary substantially based upon the event being reported on, which makes the standard bias labeling procedures of only having one bias label for a domain insufficient to truly characterize the bias of a media domain.

#### **Related Research**

Since the rise of phenomena like 'fake news' and the growth of heavily-biased media ecosystems, there has been renewed attention on understanding media bias. Hamborg et al. recently highlighted the various ways by which media can be biased and the computational approaches that exist (or do not exist) for the different forms of media bias (Hamborg, Donnay, and Gipp 2019). An important result of these re-

cent bias studies is that there are often distinct elements of topical bias, or what one chooses to talk about, as well as writing-style bias, or how one chooses to talk about that topic, present in any given media space. Of particular note is the word choice — or lexical — and framing biases in the text; a text can be biased by both the words, or phrases, it chooses to use as well as the context of keywords in the text (Hamborg, Donnay, and Gipp 2019; D'Alonzo and Tegmark 2022). A recent sub-class of framing is Information Bias, which is the conveyance of side information about the main event in the text in order to frame that main event in a certain way for the reader (Fan et al. 2019; Guo and Zhu 2022). As an example of the effect of framing, Kahneman and Tversky (1984) performed a study whereby they presented the same information, but in two different ways (one with natural numbers and one with probabilistic estimates) and found distinct differences in respondents' perceptions of the information. As an example of writing style bias created through word choice, an author can substitute words with similar syntactic meanings, but distinctly different connotations when describing something (D'Alonzo and Tegmark 2022). For example, one could say "the government is ditching vaccine mandates" versus "the government is removing vaccine mandates", where the words "remove" and "ditch" serve the same function in the sentences, but give very different connotations to the reader.

To date, most approaches to dealing with word choice and framing biases — or writing style bias — utilize NLP techniques, particularly in the form of unsupervised, text sentiment labeling (Hamborg, Donnay, and Gipp 2019; Cox and Acharya 2021; Hamborg 2022; Hamborg et al. 2021; Semeraro et al. 2022). Occasionally, this analysis is supplemented by topic modeling for event detection and clustering topics on the same article together (Julinda, Boden, and Akbik 2014; Best et al. 2005; Hamborg 2022; Jiménez Muñoz 2022). The preponderance of unsupervised approaches is largely due to two main reasons. First, there are few goldstandard, labeled writing style bias data sets (Hamborg, Donnay, and Gipp 2019; Shahid et al. 2020). Second, labeling bias is very difficult for humans to do, given inherent personal biases, and is also context-dependent (e.g., having a certain political bias will be relative to contexts like the country of the media source) (Hamborg, Donnay, and Gipp 2019; Shahid et al. 2020; Sridharan et al. 2022). It is important to note, however, that there has been a number of recent works that have published data sets and proposed supervised machine learning models around the concept of informational bias (Fan et al. 2019; van den Berg and Markert 2020; Chen et al. 2020; Liu et al. 2022; Guo and Zhu 2022). Of particular note is that these studies have generally found that the inclusion of additional context (surrounding sentence, topics of the article, etc.) around informationallybiased sentences within an article aids in the detection of those informationally-biased sentences.

In addition to the preponderance of NLP-based approaches, there are also network-based approaches to writing style bias. Most notably, Centering Resonance Analysis constructs networks from a text by creating concept chunks, relating those chunks to each other by their nearness in the

text, and then using network analysis techniques to analyze the bias (Corman et al. 2002). Graph-theoretic approaches for prioritizing social media posts for fact-checking treat media bias (and other forms of misinformation) as a virus using epidemic spread modeling (Smith and Bastian 2022). Finally, some recent work has proposed using sentence-to-sentence networks for bias prediction at the sentence level in media articles (Guo and Zhu 2022). Yet, the proposed graph-centric approaches typically work within, constructing graphs of texts at the individual level, rather than between different texts (Corman et al. 2002; Semeraro et al. 2022). Overall, writing style bias remains an active area of research, with the overwhelming majority of techniques utilizing unsupervised NLP techniques, particularly combining sentiment labels with human analysis.

An important consideration in the analysis of writing style bias in media texts is the phenomenon of content redistribution in news production via news agencies. Media sources frequently share textual content between themselves; authorized copyediting is a fundamental component in the production of news (Hamborg, Donnay, and Gipp 2019; Boumans et al. 2018). Oftentimes this text reuse will take the form of less well-known news media organizations recycling content from major news agencies, like the Associated Press, and publishing that text (Hamborg, Donnay, and Gipp 2019; Boumans et al. 2018; Kim, Candan, and Tatemura 2009; Sanderson 1997). This type of phenomenon has been investigated using techniques from Semantic Text Similarity analysis (e.g., plagiarism detection) and is typically used to winnow down a collection of texts to just those that are unique (i.e., low-overlap in textual similarity to any other texts in the corpus) (Hamborg, Donnay, and Gipp 2019; Agirre et al. 2016). This phenomenon of text re-use, or indiscriminate text use, can also result in "nut-picking" whereby certain words or phrases which have a distinct valence or sentiment in regards to a topic are used by media sources that generally have the opposite sentiment toward that topic (e.g., the use of the typically negatively connotated word of "socialism" in a U.S. media article that actually supports a "socialism") (D'Alonzo and Tegmark 2022). The presence of such text use can challenge methods based on sentiment or valence to characterize writing-style bias. The common phenomenon of text reuse in media publication implies that when trying to compare the writing style biases of different media sources, one must do this analysis in the context of possibly large amounts of shared text between all of the media sources.

#### Methodology

Given the difficulties with creating a labeled data set for analyzing writing style media bias (Hamborg, Donnay, and Gipp 2019; Shahid et al. 2020), we opted to approach this problem as an unsupervised machine learning problem. While it is possible to use proxy labels, like the domain bias label for the website from which a text emanates from Cruickshank and Carley (2021), these are at best weak labels of the actual writing style bias present in the text. This latter point will become more evident in the analyses conducted further on in our paper. Additionally, the use of sentiment as a proxy for the writing style bias of a text presents its own

issues that are not easily reconciled without additional human analysis (Hamborg, Donnay, and Gipp 2019; Hamborg 2022; D'Alonzo and Tegmark 2022). Finally, we present a method that is capable of working with data that potentially has large amounts of text reuse between news sources, as is common in media (Boumans et al. 2018).

Thus, for the methodology in this paper, we use short-text embeddings and sentiment combined with carefully defined metrics over texts in order to construct networks for evaluating the writing style biases between texts and domains. Our proposed methodology for analyzing the writing style bias between articles and domains consists of four main steps: preprocessing and segregating the articles, creating sentence-level embeddings and the sentiment of each of the sentences for the articles, measuring article similarity by their sentence-level embeddings and sentiment differences, and then analyzing the articles and their domains by the networks created from the latter similarities <sup>1</sup>. Figure 1 displays our proposed methodology.

#### **Data Collection and Cleaning**

The first step in our proposed methodology is to collect and preprocess the textual data. We investigated the news surrounding military vaccine mandates in the United States. We chose to investigate this particular news story, as it was a major government policy with significant attention from partisan political sources, and thus a good topic for investigating media bias. We began by collecting the stories on military vaccine mandates that were shared over Twitter. By collecting news in this fashion, we can collect the articles that are most relevant to online discussions of the topic as well as acquire a greater variety of news sources than if we had collected only from a pre-determined list of news websites. We collected Twitter data from Twitter's search API<sup>2</sup> using the search phrase "military vaccine mandate" from 1 February 2022 to 5 November 2022. In total, there were 1.3 million tweets of which 17.66% had links to external websites within them. We then extracted the Uniform Resource Locators (URLs) from websites shared in the Tweets, deduplicated those URLs — leaving 30,210 unique URLs — and scraped the textual content of the URLs primarily using the Python package NewsPlease (Hamborg et al. 2017). This left 19,177 websites with textual content. From there, we labeled each successful article scraped from the URLs by their domain's political bias label (Cruickshank and Carley 2021), which left 7,065 articles with a known domain bias label.

Having obtained a selection of relevant news articles about a topic, we then preprocessed the articles. The first step of the preprocessing was to segregate the articles into distinct events. To do this, we adopted the topic modeling approach used in other works (Julinda, Boden, and Akbik 2014; Best et al. 2005), which uses a technique like Latent Dirichlet Allocation (Blei 2012) to segment the articles into distinct topics. It should be noted that other methods could be used, to include a Term Frequency Inverse Docu-

ment Frequency (TF-IDF) and anchor approach used in (Liu et al. 2022). We then further refined the segmentation of articles by down-selecting the articles within topics to those that only pertained to the primary event of the given topic (e.g., U.S. Air Force Academy refusing commissions to graduating cadets that refused the COVID vaccine, versus military members refusing the COVID vaccine). While this last step is not necessary to the overall methodology, we conducted it to ensure that the effects of other forms of bias (e.g., the bias of choosing what to report on (Hamborg, Donnay, and Gipp 2019)) are minimized relative to the writing style bias that we primarily wish to model and analyze. Table 1 summarizes the collected data set.

Primary Topic	Air Force Cadets Refused	Navy SEALs Refuse	Religious Exemptions to Vaccine	
	Commission	Vaccine	Mandate	
Number of Articles	45	52	72	
Number of Unique Domains	41	37	51	

Table 1: Summary of data set by events

Finally, we preprocessed the text within the articles by first removing junk sentences (e.g., sentences related to advertisement or subscription, "clink the link to subscribe..."), breaking each of the texts into sentences, and adding the article title as a sentence. We broke the texts down to the sentence level because previous work has demonstrated the sentence as both being an important mesostructure of textual meaning and one that is often shared between news sources (Kim, Candan, and Tatemura 2009; Boumans et al. 2018; Liu et al. 2022).

#### **Measuring Sentence Similarity**

Having obtained cleaned sentences for each of the articles, we then measure the similarity between the sentences of articles in order to compare those articles. We adopt a two-step procedure for this comparison: step one is to embed the sentences and compute their sentiment scores, and step two is to match sentences by their embeddings and sentiment scores. In the first step, we embed each of the sentences into a vector space using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model<sup>3</sup>. Embedding at the sentence level, rather than the word or document level, allows for better identification of particular biases, like informational bias, as well as better means of comparing biases between documents (Fan et al. 2019; van den Berg and Markert 2020; Liu et al. 2022; Guo and Zhu 2022). We compute the sentiment score for each of the sentences using Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert 2014). Differences in sentiment can help to distinguish between subtle word choice or punctuation

<sup>&</sup>lt;sup>1</sup>code and data available at: https://github.com/ijcruic/analysis-of-media-writing-style-bias-through-text-embedding-networks

<sup>&</sup>lt;sup>2</sup>https://developer.twitter.com/en/docs/twitter-api

<sup>&</sup>lt;sup>3</sup>https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

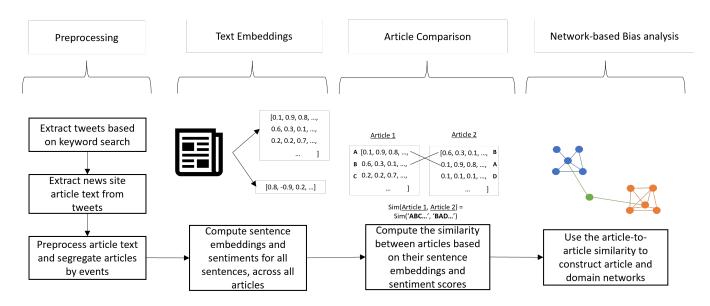


Figure 1: Proposed method for analyzing writing style media bias

manipulations for bias. See Appendix A for more details on the use of embeddings and sentiment and how they compare to other means of text comparison.

In the second step, with an embedding vector and sentiment score for each sentence, we compare sentences across articles for their semantic similarity and sentiment differences. In particular, we use embeddings to identify semantically similar sentences between articles, where sentences with only minor formatting or paraphrasing differences are mapped to nearby points in the same embedding space, but would not have been considered the same by exact text matching. Thus, the use of embeddings versus exact text matching allows for a more flexible model that can better handle real-world textual data. More specifically, we consider those sentences with a cosine similarity score of their embeddings above a certain threshold as matches. For two sentences,  $s_1$  and  $s_2$ , they are a match if,

$$cos(emb(s_1), emb(s_2)) > \tau_1 \tag{1}$$

where cos(.,.) is the cosine similarity between the embeddings of the sentences,  $emb(s_1)$ , and  $\tau_1$  is the threshold at which two sentences are semantically similar enough to be considered a match. For this study we set  $\tau_1 = 0.7$ .

We further compare differences in sentiment between semantically similar sentences to separate between those sentences which may be semantically close but have subtle differences in word choice or punctuation that give rise to a different connotation, and hence a different sentence. For two sentences,  $s_1$  and  $s_2$ , they are considered sentimentally unrelated, and therefore not a match if,

$$|sent(s_1) - sent(s_2)| > \tau_2 \tag{2}$$

where sent(.) is the sentiment score of a sentence and  $\tau_2$  is how different the sentiment between two sentences can be before they are considered different sentences. For this study we set  $\tau_2=0.1$ .

#### **Measuring Article Similarity**

Having obtained similarity measurements between the sentences of two articles, we can now compute the similarity between articles themselves. Since media articles often have high text reuse and the incorporation of reused text in an article can be done in subtle ways for writing style bias — like changing the order of sentences or adding/omitting certain sentences to a reused text — we measure the similarity between the articles in a two-step procedure. In the first step, we give a unique character to all of the unique sentences, by their embeddings. To determine uniqueness across articles we consider those sentences which have a similarity, sim(.,.), greater than 0 to be matched, unique sentences. Each of these unique sentences is then assigned a unique character. For the unique characters, we use a large subset of around 1,000 characters of the UTF-8 set of characters, which has over a million possible unique characters, to ensure the alphabet can support sets of long articles with no character reuse. With the sentences mapped to characters, we represent each article as a string of those characters. In this way, each article's string preserves the order of the sentences in the article, as well as what sentences are shared between that article and the other article. We can then measure the similarity between two articles by the edit, or Levenshtein, distance (Zhang, Hu, and Bian 2017) between their string representations. We also note that our informationtheoretic, embedding-based method of comparing articles also allows for an unordered comparison between the sentences, whereby one can use something like the Overlap Coefficient (Vijaymeena and Kavitha 2016), with the sentences that match between articles and those that do not match, to also compute article-to-article similarity.

#### **Creating Article-level and Domain-Level Networks**

Having obtained article-to-article similarities, we now construct article-level and domain-level networks to analyze the

writing-style bias. With the article similarities obtained from the previous step, we construct a weighted network, where each node is an article and each link is weighted by the similarity between the endpoint articles. Due to how the similarity was measured between articles, this network can then be used to analyze writing-style bias, by clustering the network, as well as to understand key source articles, by using standard network analysis techniques (as has been done in other studies on semantic text similarity (Hamborg, Donnay, and Gipp 2019; Kim, Candan, and Tatemura 2009)).

From the article-to-article network, we can induce a domain-to-domain network by matrix algebra:

$$D = A^{\circ \frac{1}{2}} S A^{\circ \frac{1}{2}T} \tag{3}$$

where D is a domain-to-domain network, A is the domain-to-article bipartite network, and S is the previously described article-to-article network.  $\circ \frac{1}{2}$  is the Hadamard, or elementwise, square root function. This domain-to-domain network can be similarly analyzed for writing style bias at the domain level.

#### **Results and Discussion**

In this section, we analyze the results of the writing style bias networks created on the collected data set. In the first part of the results, we analyze the networks produced by our proposed methodology, by analyzing their topology. In the second part, we analyze the clusters in the network relative to known domain biases (Cruickshank and Carley 2021) using known cluster evaluation metrics (e.g., Adjusted Rand Index (Hubert and Arabie 1985)).

#### **Network Results**

After applying our method on the data set, we had six different networks: two networks for each event. For each event (e.g., Air Force Cadets being denied a commission), we compute an article-to-article similarity network, S, and a domain-to-domain similarity network D. The six networks are visualized in Figure 2.

From observation of the similarity networks, a couple of key results emerge. First, we observe that the topologies between the article and domain networks are similar within events. This makes sense from the data as for any event, most domains only have a single article. So the articles often represent a proxy for the domains reporting the event. Second, the networks often have a core-periphery structure, where there are a few articles or domains, nodes, that have high similarity to one another (i.e., the core) and several nodes that have some similarity to certain core articles (i.e., the periphery), but not to other nodes on the periphery. This structural pattern is especially true in the Air Force Cadets and Religious Exemptions events. As these networks are a representation of a media ecosystem, this result makes sense as there is often a high amount of text reuse in media ecosystems (Boumans et al. 2018). Thus, the cores of these networks represent the main content describing the events, while the periphery contains less of that main content and, possibly, includes other related content (i.e., informational bias). Indeed, from inspection of the domains comprising

the core structures and the most central nodes by degree centrality in each of the networks, many, but not all, of these domains are more mainstream news sources or agencies (e.g., ABC, Washington Times, NBC, AP, etc.).

From the networks, we also observe that there are differences in the media reporting on each of the events. For example, the Air Force Cadets event has a strong core-periphery structure, while the Navy SEALs event has a much weaker core network structure. The Religious Exemptions event has a strong core, but also a larger periphery structure with the semblance of smaller cores in the periphery. Upon manual inspection of the articles from each of these events, this pattern matches the nature of the reporting on the events. For the Air Force Cadets event, most of the reporting centers around the official statements from the U.S. Air Force Academy on denying commissions, while the Navy SEALs articles use various elements of information, including personal interviews, mentions of other U.S. Navy vaccinerelated court cases, in addition to official statements. In fact, both the Navy SEALs and Religious Exemption events contain both isolate articles and domains. These articles and their respective domains report on aspects of each event that no other source does. Some example sentences from the periphery of each of the data sets are given in Table 2. From this analysis, we demonstrate that the network structure captures differences in how the various media actors portray events and, by doing so, represents the framing and informational biases present in media reporting.

# Comparison of Network Clusters to Known Domain Biases

From the networks, we also observe that there are often strong links between articles and domains with different known biases. To understand better the link between common bias labelings and the networks, we first clustered the networks using a common network clustering algorithm (i.e., Louvain (Blondel et al. 2008)). We then compared these clusters to publicly-available domain bias labels compiled from online sources like mediabiasfactcheck.org and allsides.com. There are six bias ratings ranging from far right to far left (i.e., far left, left, left-center, center, right-center, right, and far right). The following table, Table 3, displays the results of comparing the found clusters to the domain bias labels.

From Table 3, it can be clearly observed that the network clusters do not have much relation to the known domain bias labels. Many of the ARI scores are less than 0.1 indicating little relation between these quantities. The only exception is the domain network for the Religious Exemptions event, which has an ARI of 0.124. While this still is not an indicator of a strong relationship between the network clusters and the known domain biases, it does match the topology of the network. The Religious Exemptions domain similarity network has a core of mostly center-biased domains and then connectivity in the periphery between right-biased domains. We then computed the modularity of the known biases over the networks to get a measure of whether domains or articles tend to use content from similarly biased articles or domains or not. The low modularity scores for each of the

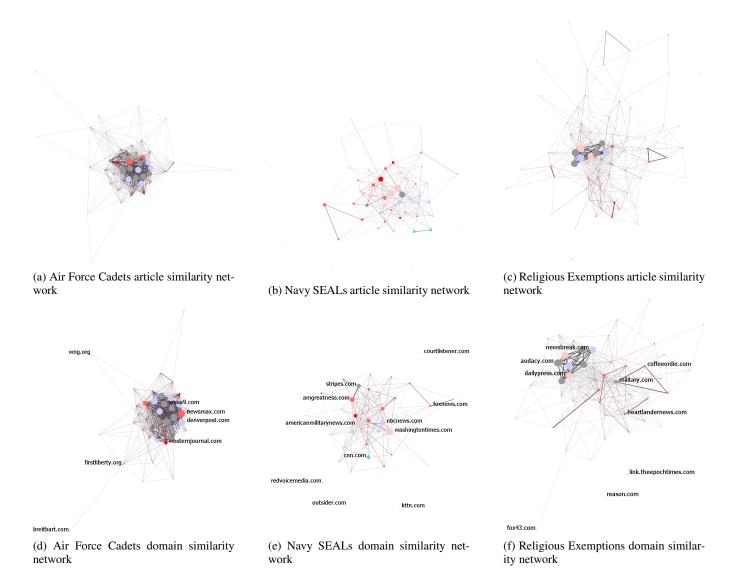


Figure 2: Article similarity networks, S, and domain similarity networks, D, for each of the events in the military vaccine mandates news data set. Nodes are sized by their degree centrality and colored by their domain bias ratings (red=right, gray=center, blue=left). There are six bias ratings ranging from far right to far left. Edges are sized by their weight, which represents the similarity, between 0 and 1, between the nodes at the endpoints of the edge.

networks indicate that domains and articles do not show a preference for using material from other articles or domains that share the same bias. Finally, we observe that there are differences between the events in terms of how much their clusters match known domain biases; the Religious Exemptions event tends to agree more with known domain biases, while the Air Force Cadets event bear little relation to them. This result may be due to how polarizing an event or topic is and how diverse the information coming from that event is. For example, much of the Air Force Cadet's event stems from official statements from the U.S. Air Force and was not as polarizing as an event like the court cases around the religious exemptions to the COVID-19 vaccine, as this event mentions the very divisive topic of religion.

We lastly analyzed the clusters across all of the events for their relation to known domain biases. To do so, we used cluster ensembling (i.e., Locally Weighted Bipartite Graph Partitioning Algorithm (Huang, Wang, and Lai 2017)) to cluster the clusters from each of the events across all of the domains present. For those domains that did not produce an article for a given event, and hence were not present in that event's clusters, we gave a new cluster label to them. The ARI between the ensemble clusters and known domain bias labels was 0.09155. While still not a high ARI, it is more positive than most of the event ARIs, indicating that as we aggregate across sub-events, the clusters more closely resemble the known bias labels. This result mirrors that of other studies, which found there are nuances in tradi-

Event	Core Sentence Example	Periphery Sentence Example	
Air Force Cadets	Three cadets at the U.S. Air Force Academy who have refused the COVID-19 vaccine will not be commissioned as military officers but will graduate with bachelor's degrees, the academy said Saturday.	I have been very vocal against the United States Air Force Academy (my alma mater) for its corrupt teaching of cultural Marxism and its Covid 'vaccine' mandates.	
Navy SEALs	Justices Clarence Thomas, Samuel Alito and Neil Gorsuch dissented.	The decorated Navy veteran appeared to recite the speech from memory, emphasizing certain phrases that are alarmingly relevant to the tyrannical times we're living in now.	
Religious Exemptions	The Air Force became the second military service to approve religious exemptions to the mandatory COVID-19 vaccine, granting requests from nine airmen to avoid the shots, officials said Tuesday	Health experts said Novavax's COVID-19 vaccine could help address religious arguments made for exemptions.	

Table 2: Sentences characteristic of Core articles and Periphery Articles from the networks of each of the three events.

	Air Force	Navy	Religious	
	Cadets	SEALs	Exemptions	
ARI between Clusters and Bias Ratings at the Article Level	0.062673	0.00213	0.02405	
ARI between Clusters and Bias Ratings at the Domain Level	-0.00704	0.05441	0.12442	
Modularity of Bias Ratings at the Article Level	-0.01091	-0.0141	0.00186	
Modularity of Bias Ratings at the Domain Level	-0.02142	0.08939	0.08742	

Table 3: Comparison of network clusters to known Domainlevel bias labels. The network clusters and Bias labels are compared by Adjusted Rand Index (ARI) (Hubert and Arabie 1985). We also show the modularity of the bias labels on the networks, indicating how much domains or articles share content within their respective biases versus between biases.

tional bias labels between topics, but that over a larger group of topics, these biases become more stable (D'Alonzo and Tegmark 2022).

#### **Sensitivity Analysis**

We conducted an analysis of the two user-set parameters of the proposed method. More specifically, we wanted to see how much varying the semantic threshold (i.e.  $\tau_1$ ) and the sentiment threshold (i.e.  $\tau_2$ ) affected the results. To analyze the sensitivity of the method to these two parameters, we constructed article-to-article networks, A, across all three data sets while varying the values of  $\tau_1$  and  $tau_2$  between 0.1 and 0.99. We then measured the difference in the con-

structed networks by the normalized manhattan distance between the network adjacency matrices, which is given by the equation:

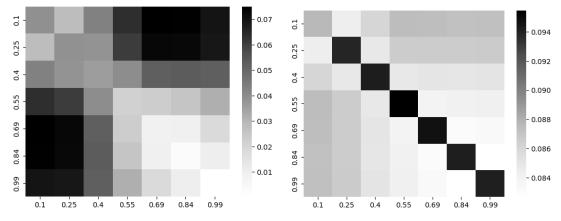
$$d(A_i, A_j) = \frac{\sum |A_i - A_j|}{N(N-1)}$$
 (4)

where N is the number of nodes, which makes N(N-1) the total number of possible links in an undirected network. This measure indicates how different two networks are when the parameters involved (i.e.,  $\tau_1$  and  $\tau_2$ ) in their construction are varied. Having computed the normalized Manhattan distances between all of the possible networks, we then averaged these across all of the data sets where  $\tau_1$  and  $\tau_2$  match and across all of the parameter values for each parameter's set of values (e.g., all of the  $\tau_2$  values for each combination of  $\tau_1$  values and vice versa). The following figure, Figure 3 displays the average distances between networks for each of the  $\tau_1$  and  $\tau_2$  combinations.

From Figure 3, it can be observed that the article-to-article networks produced are relatively stable for values of  $\tau_1$  greater than 0.6 and for  $\tau_2$  for values less than 0.3 or greater than 0.6. Thus, the method is relatively insensitive to reasonably high semantic thresholds  $(\tau_1)$ , where the match between sentences should be at least 0.6 out of 1.0 for them to be considered a match and relatively insensitive to values of sentiment that are reasonably close to each other (e.g., not differing by more than 0.3). Based on these results, this validates our setting of  $\tau_1=0.7$  and  $\tau_2=0.1$ , as these are within a stable range of results and represent having sentence pairs that are reasonably close semantically and only vary a little in sentiment to be considered as matches.

#### Conclusion

In this article, we proposed a new methodology for analyzing writing-style bias in media publications. Our methodology uses computational techniques to assess text overlap and identify bias, specifically at the sentence level, by applying natural language embeddings and a metric to evaluate text overlap between articles. The methodology does



(a) Sensitivity of  $\tau_1$  (semantic sentence match threshold) (b) Sensitivity of  $\tau_2$  (sentiment sentence match threshold)

Figure 3: Normalized manhattan distances between networks produced using varying parameters. Each cell represents the average distance between networks constructed by the parameter value in the row and the parameter value in the column, averaged across all three data sets and all varying levels of the other parameter.

not rely on expert opinions or knowledge of media producer publishing practices but instead on what the domains produce. Our validation shows that text reuse and framing vary by the event, indicating that a single label for media producer bias is insufficient. Our study has some limitations, as we only focused on three events related to U.S. military COVID-19 vaccine mandates, and more research is needed to explore the generalizability of our findings across different topics and events. Additionally, we believe that there is an opportunity for future research to improve the methodology for determining whether two sentences are discussing the same thing in the same way.

#### **Ethical Statement**

All articles collected for this study were done so under the provisions of Section 1078 of the U.S. Copyright Act and ensured that our collection action fell under the fair use category. The Tweets were collected in accordance with Twitter's terms of service at the time of collection.

## Acknowledgements

#### References

Agirre, E.; Banea, C.; Cer, D.; Diab, M.; Gonzalez Agirre, A.; Mihalcea, R.; Rigau Claramunt, G.; and Wiebe, J. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511. ACL (Association for Computational Linguistics).

Best, C.; van der Goot, E.; Blackler, K.; Garcia, T.; and Horby, D. 2005. Europe media monitor. Technical Report EUR221 73 EN, European Commission.

Blei, D. M. 2012. Probabilistic topic models. Communications of the ACM, 55(4): 77-84.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10): P10008.

Boumans, J.; Trilling, D.; Vliegenthart, R.; and Boomgaarden, H. 2018. The Agency Makes the (Online) News World go Round: The Impact of News Agency Content on Print and Online News. International Journal of Communication, 12(0).

Chen, W.-F.; Al-Khatib, K.; Stein, B.; and Wachsmuth, H. 2020. Detecting media bias in news articles using gaussian bias distributions. arXiv preprint arXiv:2010.10649.

Corman, S. R.; Kuhn, T.; McPhee, R. D.; and Dooley, K. J. 2002. Studying complex discursive systems. Centering resonance analysis of communication. Human communication research, 28(2): 157-206.

Cox, G.; and Acharya, A. 2021. Sentiment Analysis and NLP models for Identifying Biases of Online News Stations. Cruickshank, I. J.; and Carley, K. M. 2021. Clustering analysis of website usage on twitter during the covid-19 pandemic. In Annual International Conference on Information Management and Big Data, 384–399. Springer.

D'Alonzo, S.; and Tegmark, M. 2022. Machine-learning media bias. Plos one, 17(8): e0271947.

Fan, L.; White, M.; Sharma, E.; Su, R.; Choubey, P. K.; Huang, R.; and Wang, L. 2019. In plain sight: Media bias through the lens of factual reporting. arXiv preprint arXiv:1909.02670.

Guo, S.; and Zhu, K. Q. 2022. Modeling Multi-level Context for Informational Bias Detection by Contrastive Learning and Sentential Graph Network. arXiv preprint arXiv:2201.10376.

Hamborg, F. 2022. Towards Automated Frame Analysis: Natural Language Processing Techniques to Reveal Media Bias in News Articles. Ph.D. thesis.

- Hamborg, F.; Donnay, K.; and Gipp, B. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4): 391–415.
- Hamborg, F.; Heinser, K.; Zhukova, A.; Donnay, K.; and Gipp, B. 2021. Newsalyze: Effective Communication of Person-Targeting Biases in News Articles. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 130–139. IEEE.
- Hamborg, F.; Meuschke, N.; Breitinger, C.; and Gipp, B. 2017. news-please: A Generic News Crawler and Extractor. In *Proceedings of the 15th International Symposium of Information Science*, 218–223.
- Huang, D.; Wang, C.-D.; and Lai, J.-H. 2017. Locally weighted ensemble clustering. *IEEE transactions on cybernetics*, 48(5): 1460–1473.
- Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of classification*, 2: 193–218.
- Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.
- Jiménez Muñoz, H. 2022. *Towards an automatic bias detection system in journalism*. B.S. thesis, Universitat Politècnica de Catalunya.
- Julinda, S.; Boden, C.; and Akbik, A. 2014. Extracting a repository of events and event references from news clusters. In *Proceedings of the first aha!-workshop on information discovery in text*, 14–18.
- Kahneman, D.; and Tversky, A. 1984. Choices, values, and frames. *American psychologist*, 39(4): 341.
- Kavanagh, J.; and Rich, M. D. 2018. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. Santa Monica, CA: RAND Corporation.
- Kim, J. W.; Candan, K. S.; and Tatemura, J. 2009. Efficient overlap and content reuse detection in blogs and online news articles. In *Proceedings of the 18th international conference on World wide web*, 81–90.
- Liu, Y.; Zhang, X. F.; Wegsman, D.; Beauchamp, N.; and Wang, L. 2022. POLITICS: pretraining with same-story article comparison for ideology prediction and stance detection. *arXiv* preprint arXiv:2205.00619.
- Sanderson, M. 1997. Duplicate detection in the Reuters collection. "Technical Report (TR-1997-5) of the Department of Computing Science at the University of Glasgow G12 800, UK".
- Semeraro, A.; Vilella, S.; Ruffo, G.; and Stella, M. 2022. Emotional profiling and cognitive networks unravel how mainstream and alternative press framed AstraZeneca, Pfizer and COVID-19 vaccination campaigns. *Scientific reports*, 12(1): 1–12.
- Shahid, U.; Di Eugenio, B.; Rojecki, A.; and Zheleva, E. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, 120–125.

- Smith, J.; and Bastian, N. 2022. A ranked solution for social media fact checking using epidemic spread modeling. *Information Sciences*, 589: 550–563.
- Sridharan, A.; et al. 2022. An Automated News Bias Classifier Using Caenorhabditis Elegans Inspired Recursive Feedback Network Architecture. *arXiv preprint arXiv:2207.12724*.
- van den Berg, E.; and Markert, K. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6315–6326.
- Vijaymeena, M.; and Kavitha, K. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2): 19–28.
- Zhang, S.; Hu, Y.; and Bian, G. 2017. Research on string similarity algorithm based on Levenshtein Distance. In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2247–2251. IEEE.

# Appendix A: Example of Text Manipulations and Similarity Measurements

In this section, we demonstrate a number of approximate string comparison techniques, and how those techniques produce results around a number of ways sentences can be changed or manipulated, potentially for bias purposes. To demonstrate the techniques, we performed a number of manipulations on the base sentence of "Three cadets at the U.S. Air Force Academy who have refused the COVID-19 vaccine will not be commissioned as military officers but will graduate with bachelor\'s degrees, the academy said Saturday.". The following Table 4 summarizes the alterations to the base sentence.

We then analyzed the many means of comparing texts using the base sentence and the altered sentences from Table 4. Specifically, we looked at exact text matching, approximate text matching by both the Jaccard Index and by Levenshtein distance, by the cosine similarity between whole-sentence embeddings and between pre-trained word embeddings, and by differences in VADER sentiment. For the approximate text matching by Jaccard Index, approximate text matching by Jaccard distance, and the cosine similarity between pre-trained word embeddings, we preprocessed the text following common convention with these techniques by removing common English stopwords, punctuation, and pronouns. The results of these comparisons are displayed in Table 5.

From this example, one can observe that exact text matching is insufficient to handle comparing sentences generally. In terms of the approximate matching techniques, no particular one is clearly superior to the others when it comes to still matching minorly edited or paraphrased sentences and not matching those sentences which have been manipulated in ways that affect the bias or connotation of the sentence. The only exception is the ability of deep learning models to handle foreign languages; it's possible for a sentence to be in a foreign language and still be detected as similar to the base

Text Alteration	New Sentence Text		
	Three cadets at the U.S. Air Force Academy who have refused the COVID-19		
formatting typo	vaccine will not be commissioned as military officers, but will graduate with		
	bachelor\'s degrees,\n the academy said Saturday.		
	Three cadets at the U.S. Air Force Academy who have refused the COVID-19		
inserted punctuation	vaccine will not be commissioned as military officers, but will graduate with		
_	bachelor\'s degrees, the academy said Saturday.		
	Cadets at the U.S. Air Force Academy who have refused the COVID-19 vaccine		
missing unimportant word	will not be commissioned as military officers, but will graduate with bachelor\'s		
	degrees, the academy said Saturday.		
	Three cadets at the U.S. Air Force Academy who have \"refused\" the COVID-19		
framing with quotes	vaccine will not be commissioned as military officers but will graduate with		
	bachelor\'s degrees, the academy said Saturday.		
	The Air Force Academy said Saturday that the three cadets who refused the		
rephrasing	COVID-19 vaccine will not be commissioned as military officers but will		
	graduate with bachelor\'s degrees.		
heavy rephrasing	The Air Force Academy is requiring cadets to vaccinate against COVID-19		
heavy repiliasing	to commission.		
	Three cadets at the U.S. Air Force Academy who declined the COVID-19		
simple word change	vaccine will not be commissioned as military officers, but will graduate with		
	bachelor\'s degrees,\the academy said Saturday.		
	Three cadets at the U.S. Air Force Academy who have refused the COVID-19		
complex word change	vaccine will be denied commissions as military officers but will graduate with		
	bachelor\'s degrees, the academy said Saturday.		
	Three cadets at the U.S. Air Force Academy who have refused the COVID-19		
omission	vaccine will not be commissioned as military officers, the academy said		
	Saturday.		
	Three cadets at the U.S. Air Force Academy who have refused the COVID-19		
unimportant addition	vaccine will not be commissioned as military officers but will graduate with		
	bachelor\'s degrees, a spokesman from the academy said Saturday.		
unrelated	The U.S. Air Force Academy is located in Colorado Springs.		
	Three cadets from the U.S. Air Force Academy, along with other		
more similar unrelated	commissioned military officers, presented at the graduate		
	academy colloquium on COVID-19 vaccination among bachelors this Saturday.		
	Tres cadetes de la Academia de la Fuerza Aérea de Estados Unidos que rechazaron		
foreign language	la vacuna COVID-19 no serán comisionados como oficiales militares, pero		
	se graduarán con una licenciatura, dijo la academia el sábado.		

Table 4: Different types of text alterations which can happen both in a copy-editing scenario and when trying to inject bias into a sentence.

sentence by suitable, multilingual deep neural network embedding models. We can also observe that differences in sentiment can capture subtle changes in punctuation and word choice, which can be done to alter the connotation and bias of a sentence. Thus, from these results, we propose that one must compare both the semantics, in a flexible way, and the sentiment in order to best capture sentence-level bias alterations.

Text Alteration Technique	Exact Match	Jaccard Approximate Match	Levenshtein Fuzzy Approximate Match	Sentence Embeddings Approximate Match	Cosine Similarity Between Embeddings	Difference in VADER Sentiment
formatting typo	FALSE	0.94	0.99	1	0.99	0
inserted punctuation	FALSE	1	0.99	1	0.99	0
missing unimportant word	FALSE	1	0.99	1	0.95	0
framing with quotes	FALSE	1	0.99	1	0.98	0.15
rephrasing	FALSE	0.94	0.82	0.98	0.97	0
heavy rephrasing	FALSE	0.33	0.58	0.87	0.66	0.15
simple word change	FALSE	0.83	0.95	0.99	0.99	0.15
complex word change	FALSE	0.94	0.96	0.99	0.98	0.22
omission	FALSE	0.81	0.88	0.96	0.94	0.14
unimportant addition	FALSE	0.94	0.91	0.99	0.98	0
unrelated	FALSE	0.21	0.59	0.71	0.36	0.15
more similar unrelated	FALSE	0.63	0.61	0.95	0.83	0.15
foreign language	FALSE	0.02	0.47	0.06	0.97	0.14

Table 5: Results of different text comparison techniques across different text alteration techniques.