THE UNVERSITY OF CHICAGO

Using Topic Modelling to Detect Bias Between News Sources

By

Alexander Tyan

June 2019

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Allyson Ettinger
Preceptor: Joshua Mausolf

Abstract:

How do online news articles' features explain and predict the bias of the article content? In this thesis, I intend to argue that news article features such as topics (along with the source of publication) predict the extent to which news sources converge or diverge from each other in news content. I define this distance in news as *bias.* I propose a novel way to define and operationalize bias to introduce an objective and measurable way for studying this phenomenon. This measure differs from the way prior research treats media bias in that it is applicable to a broader set of research contexts than prior methods and is not reliant on data sources external to the news articles.

Introduction:

How do online news articles' features explain and predict the bias of the article content? In this thesis, I intend to argue that news article features such as topics (along with the source of publication) predict the extent to which news sources converge or diverge from each other in news content. I define this distance between news articles as *bias.* Mathematically, this *bias* is a cosine distance between vector forms of text contents of news articles and their corresponding sources. In proposing this novel way to define and operationalize bias, this research introduces an objective and measurable way for studying the topic. This measure differs from the way prior research treats media bias in that it is applicable to a broader set of national and international contexts, such as studying media bias outside of American political settings.

There are three main motivations behind this research. First, as alluded to above, prior studies of media bias have been largely limited to specific contexts, such as American politics or media coverage of protest movements. Usually, this limits the treatment of bias to specific settings, such as bias measurements along left-right political spectrum (e.g. Niven 2003), as understood in the U.S. Or, for news coverage on protest movements, treatment of bias may be confined to comparisons of media presentation of protests to intended protesters' goals (Smith et al. 2001). Moreover, as we see later in the literature review, research often ties bias measures to some external sources of information, such as texts of Congressional speeches (Gentzkow and Shapiro 2010). This makes bias measurements rely on availability of such external data, which may not always be accessible. While these approaches work in specific settings, they produce a plethora of bias measurements, subject to context and data availability limitations. My research fills this gap by proposing a bias measurement that is less reliant on specific context and is independent from data sources external to news articles and their metadata. This makes my approach more accessible to a researcher studying news content.

Second, from the perspective of concern for democracy and civic engagement, developing a systematic way to measure bias is important for two reasons. First, in the United States there has been for a number of years a growing criticism of media in regards to objectivity (Watts et al. 1999). This has culminated in the perception of “fake news” stories in 2016 presidential election in the United States (Allcott and Gentzkow 2017)[illegible]ond reason is the importance of news media as a political institution. As argued by Timothy Cook (2005), news media is important not only from the perspective of informing the general population, but as a medium of communication between different branches of government. As such, Cook argues that news media should be considered as an important political institution. Despite the importance of the institution for these two reasons, we often lack an approach that may inspire a more nuanced discussion and appreciation of what “news bias” is and what may determine the “bias” extent at any given point in time. This project contributes to previous research by providing a more concrete and measurable way to talk about bias.

[illegible]ird motivation is the importance of the concept of news bias for many academics. A quick survey of academic literature reveals a great number of publications relating to bias in the news. Some examples include the effect of news coverage on financial markets (Engelberg and Parsons 2011; Gurun and Butler 2012) and democratic elections (Allcott and Gentzkow 2017; DellaVigna and Kaplan 2007). Other examples relate to news representations of disadvantaged populations (Peffley, Shields, and Williams 1996). Yet, as Entman (2007b) points out, the definitions of what news bias is are diverse and fractured and there is not an academic consensus on how it should be studied systematically and across disciplines. Therefore, the development of my method contributes to a more generalizable method for understanding news bias.

The research has selected the following four sources for data analysis and testing of hypotheses: BBC (2018), RT (2018), Fox News (2018), and CNN (2018). In terms of provisional arguments, I hypothesize that there is a statistically significant difference in the bias measure between different news sources' articles, conditional on topics. That is, one would expect that for topic A the average cosine distance between texts of sources X and Y should be different than for topic B. Implicit here is the idea that, even for news sources that may be popularly perceived as on different sides of an arbitrary spectrum (e.g. MSNBC vs Fox News along some ideological dimension), it may be that, for particular values of article features (topics), the content between sources is not all that dissimilar most of the time. On the other hand, there may exist other feature values that are particularly predictive of content divergence. For instance, it may be that certain political topics, like Russian government activity during US presidential elections in 2016, predict especially strong divergence between CNN and RT.

Literature Review:

What is news?

An important consideration in choosing the data for this study is determining what "news" is. In part, the selection of sources is based on a larger body of theoretical literature in communications on what should be considered "news". One of the most important lines of work in determining what is news comes from a classic piece by Galtung and Ruge (1965) who present a taxonomy of main characteristics for a news article. The authors identify news values that determine whether a piece of text is "news." These values fall into three categories: impact, audience identification, and pragmatics of media coverage. This work was later expanded upon by another highly cited piece in journalism research by Harcup and O'Neill (2001) and then expanded again in (Harcup and O'Neill 2017). According to these researchers, news stories must satisfy one or more of the characteristics they list (fifteen characteristics or values in 2017 version of the article). Such characteristics include, for instance, "bad news", "good news" (especially bad or good overtones), follow-up (news about stories already in the news), magnitude ("stories perceived as sufficiently significant in the large numbers of people involved or in potential impact, or involving a degree of extreme behaviour or extreme occurrence") and others.[1] My research project in part selected its sources based on such characteristics and news values.

However, my research project excludes editorial and op-ed articles, based on Kahn and Kenney (2002). The authors analyze whether endorsement of a Senatorial candidate in the editorial section influences the news coverage for the same news sources in the news section. They find that it does and news reporting tends to be more favourable if the editorial has endorsed a candidate (they also find that reading such news sources positively affects electoral views of readers for the endorsed candidate). The premise behind their analysis is the idea that there should be a "wall of separation" between the editorial and news sections. While the study is on the effects of traditional print media, I believe it is appropriate to extend this idea to online news publishers, because the layout and section separations of the news sites I choose have historically mimicked that of a print paper (e.g. World News, US News, Sports News) (see Figure 1 for the visual section titles on BBC News and CNN website for an example). Hence, consistent with Khan and Kenney's research premise, I exclude editorial/op-ed pieces from my analysis.

[1] For a full list, see (Harcup and O'Neill 2017, 1482)

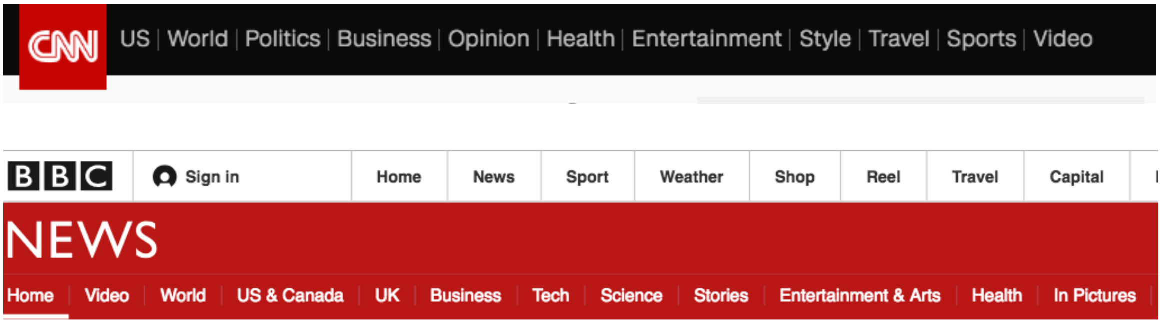

CNN
US | World | Politics | Business | Opinion | Health | Entertainment | Style | Travel | Sports | Video
BBC
Sign in
Home
News
Sport
Weather
Shop
Reel
Travel
Capital
NEWS
Home | Video | World | US & Canada | UK | Business | Tech | Science | Stories | Entertainment & Arts | Health | In Pictures

Figure 1: CNN and BBC News Section Headers

In regards to choosing specific news sources, the selection of sources is based on diversity of English-speaking news sources used in the news analysis in prior literature. For instance, Feldman et al. (2012) analyze climate change coverage on CNN, MSNBC, and Fox News. Another example is Iyengar and Hahn (2009) using Fox News, CNN, and NPR to analyze ideologically-based selectivity in media consumption. So my choice of CNN and Fox News is in line with prior literature choices. I also include BBC as a well-recognized news outlet for international news and RT as a source well-known for its Russian government funding, since bias is a central topic of study in this research. Additionally, since our bias measure is constructed in comparing contents between sources, choosing sources that overlap in their reports on international and US topics should provide us with enough content overlap to detect bias as defined earlier, without sacrificing generalizability of our approach to other contexts.

Bias:

A key theoretical and empirical concept for this project is the notion of bias itself. ~~As mentioned,~~ it is a concept that receives much attention across many academic fields, such as Political Science, Economics, and Communication. However, there does not seem to be a consensus on the definition.

In his theoretical work in Communications, Robert Entman (2007a) identifies three existing meanings of the term. He does so in the context of news shaping political competition and democracy. The first is *distortion bias* for "news that purportedly distorts or falsifies reality" (Entman 2007a, 163). The second is *content bias* as not providing equivalent treatment to both sides of political conflict (Entman 2007a, 163). The third notion is *decision-making* bias, as relating to the editorial decisions of news-makers (Entman 2007a, 163). Entman chooses for his proposed notion of bias to encompass the latter two concepts (he does not elaborate why).

Crucially, ~~the author~~ distinguishes between *bias* and *slant* (Entman 2007a, 165). For Entman, *news slant* pertains to individual news reports favouring one side or the other. In contrast, *bias* is "consistent patterns in the framing of mediated communication that promote the influence of one side in conflicts over the use of government power" (Entman 2007a, 166). So the main difference conceptually is that slant is for individual news articles observations, whereas bias is a more systematic pattern across news reports.[2]

My measure of bias is not nearly as granular and theoretically nuanced as Entman's. For one, I do not distinguish between different kinds of bias and between bias and slant. This stems primarily from the computational methodology I employ to measure bias as a cosine angle distance between news articles. On one hand, this approach does not allow for a conceptual separation of bias types because all "bias" is simply a vector distance. As I write in the discussion section, one could attempt to improve on the current methodology to distinguish bias types and bias vs. slant. On the other hand, the advantage of my more simple approach is in its generalizability, as Entman's notion of bias may be harder to apply depending in different contexts. For instance, Entman's content bias is about not providing equal treatment to different sides of a political contest. Such an assessment, while a useful heuristic, might be difficult to apply in practice as it calls for a value judgement by a researcher. This may be especially difficult if one works with a great number of diverse news articles. My

[2] Entman also proposes measuring *slant* via his Aggregate News Slant Index (ANSI), which uses paragraph level text analysis to determine the slant for those individual paragraphs. The document-level slant is than a type of an aggregate measure across these paragraphs. For many documents, one would then have many corresponding ANSI measures. Taking an average of them would then be a measure of bias according to Entman in so far as a high ANSI average would indicate a more systematic bias of a group of news articles (i.e. a news source in our research).

approach is computationally scalable and does not require any value judgements on the part of the researcher.

In regards to predicting or explaining bias in the news content, a methodologically different approach is Gentzkow and Shapiro's (2010) explanation of media bias (which they call "slant", not to be confused with Entman's notion) as a function of consumer preferences from an econometric perspective. Researchers find this response to market demand accounts for about 20% of variance in slant. This research measures slant/bias as a similarity of the news text to the language of Congressional Republicans and Democrats. A higher measure of similarity between a news report and a Congressional statement then indicates a higher slant.

This measure of bias is then inherently specific to US political context and is in relation to the messages of US politicians at the time of this study. As mentioned, my approach is more generalizable in that it is independent from external data sources, like Congressional speeches, and is applicable to different contexts. For instance, if one were to replicate Gentzkow and Shapiro's study on topics other than the political notion of right and left, it is unclear whether this approach would be as readily transferrable to such contexts.

Another related empirical work in economics is by Groseclose and Milyo (2005). The researchers measure media bias again as tied to the specific left-right political dimension in the US context. The measure compares news sources citing liberal/conservative think tanks and comparing these citation counts to counts by Democrat/Republican Congress members citing those same think tanks. These count comparisons yield a score that measures bias. This research finds a left-leaning bias in US media outlets overall (Fox News' Special Report and Washington Times are an exception). Thus, the study is similar to Gentzkow and Shapiro (2010) in connecting news content bias to Congressional texts. As a result, I would advance a similar critique to such approaches for similar reasons of non-generalizability and dependence on external data sources.

Yet another empirical study on media bias in the US political context is by a communications scholar Niven (2003). The innovation of this research is that the author uses news coverage of US Congress members who switch parties. This allows the researcher to establish a baseline for fare comparison of news coverage to detect potential bias. As argued by Niven (2003), "If the media are biased for liberals and Democrats, then their coverage should be more sympathetic to Republicans who are leaving their party than to defecting Democrats" (316). The study does not find any statistically significant difference between media coverage of Republican/Democrat switchers. The consequence for this study is that it is even more context-specific (Congressional party-switching) and less generalizable.

Rochelle Terman's (2017) empirical study in international relations used the notion of bias as pertaining to western news reports on Muslim women. More specifically, to measure "confirmation bias" in news reports the study used the likelihood of news coverage (determined by article count) on Muslim women abroad. She found, among other things, that western media outlets are much more likely to report on women in predominantly Muslim countries when their rights are violated (controlled for actual statistics on women right violations) than in non-Muslim countries. Here the measure of bias in the news is based on a more objectively measurable standard (count comparisons for news articles). Again, both Terman's is context-dependent and would be harder to generalize for a wider set of use cases by a researcher.

In my research, I define bias as a cosine distance between texts (as vectors; see Methods) of news articles which represent respective news sources. Here I am assuming that these texts exist in an abstract latent geometric space that represents news article content as would be understood by an Englis-speaking reader. Thus, I propose that the distance between these texts is a measure of bias. In this sense, I adopt some conceptual similarity to Entman's research, insofar as I am using a relativistic measure of bias between the articles. However, the distinction with previous research is I do not differentiate between bias and slant, do not tie my measures to specific context (e.g. bias according to political spectrum), nor do I construct my measure of bias in a way that is tied to a more concrete and/or objectively measurable quantity (like occurrence counts or partisan speeches). Instead, my measure uses abstract positioning of texts in relation to each other. In some ways, this study harnesses the advantages of (Groseclose and Milyo 2005) who do not need to classify think tanks as left or right, but use relative positioning of news sources as tied to Congress members' texts. However, unlike (Groseclose and Milyo 2005), my research does not constrain itself to a specific context. As alluded to in the introduction, this has several important contributions to the existing literature.

First, because this measure is not context specific, the methods I use should be more transferrable and generalizable across different text contexts. Entman's and Gentzkow's, Shapiro's, Niven's, and Groseclose's and Milyo's methods are specific to US political context. As a result, it is not clear how one would measure bias if the number of discrete categories (e.g. Republican vs Democrat, Government vs Opposition) is much larger. Nor is it clear how one would measure bias when exogenous information like partisan speeches is not available and the news texts is the only available data. My approach is agnostic to the lack of external data ties to construct bias measurement.

Second, my method allows the research to make no presuppositions about the "truth" against which bias would be measured. It is true that in some context, like in Terman's research, there is an available metric to produce the difference between the expectation and the empirical observation. However, in a lot contexts, especially politically charged ones, asserting a truth to measure bias is itself a normative statement. The problem becomes even more severe when the study explores many dimensions along which data can vary, not just say political spectrum. ~~For~~ it is difficult to come up an operationalizable "truth" across many dimensions. Thus, my approach is more flexible when treating text data across many potential and latent dimensions along which bias could be measured. This way my approach is amenable to be adopted in different contexts where a researcher may decide on the baseline for comparison herself. For instance, she may propose that BBC is a good standard for journalistic integrity and should be used as the baseline from which one would measure bias distance to other news sources.

~~Important to note is~~ I do not claim that all issues reported on in the news do not have an objective truth standard. I do claim, however, that in a lot contexts, operationalizing such benchmarks to measure bias is either practically infeasible or objectionable from the point of view of preserving researcher's impartiality in the matter.

Methods:[3]

While methods are not the main focus of this literature review, it is worthwhile to briefly mention some literature on the relevant methodologies and give a few examples of how these methods have been used in the social science context.

A major part of the analysis will be based on representing text as mathematical vectors which will allow calculation of cosine distances between texts. Several methods exist, but the project applies standard pre-processing steps common to Natural Language Processing (NLP), such as part-of-speech tagging and lemmatization to extract words in parts of speech that are meaningful to text content (e.g. nouns, adjectives) and capture words and frequently occurring word combinations. This processed text base will allow me to use word or lemma vector embeddings, which will be mathematical vector representations of the text upon which I can perform calculations of cosines distances. Processing and vectorizing the text in such a way is outlined in (Chollet 2017).

I will use these representations to engineer features of the text, such as topics. Again, several methods exist, but this project will use Latent Dirichlet Allocation (LDA) described in (Blei, Ng, and Jordan 2003), potentially expanding

[3] Upon consultation with colleagues and the advisor, the methods may be adjusted

to Correlation Explanation (CorEx) as introduced in (Gallagher et al. 2016) or Structural Topic Modelling (STM) as described in (Roberts, Stewart, and Tingley 2014). Thus, I may use such unsupervised or semi-supervised techniques as topic modelling to construct the topic variable. Importantly, the topic space will be constructed for all news articles. Since data collection will happen on a massive scale (we are capturing all news articles from all sources in a given time frame), I expect many topics (i.e. common news subjects) will emerge across many if not all news sources under consideration. This allows comparing distances between news sources conditional on a given news topic.

In regards to prior application of topic modelling in social science literature, Grimmer and Stewart (2013) outline the use of text as data, including probabilistic topic modeling. Using such methods, Grimmer (2016) uses documents topics in Congressional speeches and finds that Republican representatives abandon credit-claiming for policies and shift to criticizing their Democrat counterparts, while Democrat representatives continue to credit-claim and defend their policies in the aftermath of the 2008 presidential election.

Using a Structural Topic Model, a method similar to LDA (Lucas et al. 2015) and NLP techniques I describe earlier, Terman (2017) extracts topics proportions from news articles and finds that, conditional on region and year, news articles on Middle East and North Africa concentrate on women's rights and gender equality disproportionately more than on Western countries, even when controlling for the reality on the ground (i.e. controlling for the fact that women rights in certain regions may indeed be less protected).

Thus, despite the unsupervised nature of probabilistic topic modelling, the methods seem to yield significant insight across a range of studies (see Grimmer and Terman for more examples). I intend to use these methods for constructing topics as an independent variable for news articles.

More generally, the use of text as data, including news text, is not uncommon in the social sciences. For instance, Gentzkow and Shapiro (2010) study media slant of newspaper article content, a measure they tie to congressional Republican and Democrat speeches. The news data they use is from the ProQuest Newsstand database (2018) and NewsLibrary database (2018). Another example of using text as data is Grimmer's (2013) use of congressional speeches to extract expressed priorities of US Congress representatives through text topics; a method similar to what I will describe in the methods section.

Data and Methods:

The data collection process is from January 2018 through May 2019.[4]. The project uses Python 3.6 programming language with the Newspaper3k package

[4] The project is prolonging the data collection period as long as possible, subject to deadline constraints, because capturing more observations increases the likelihood of content overlap in topics, which is helpful for the analysis stage.

(Ou-Yang 2013) to automate scraping of the news article content from these news websites: BBC (2018), RT (2018), Fox News (2018), and CNN (2018). Since ultimately the project needs to represent bias as a mathematical distance between individual news articles, the unit of analysis is at the level of a pairwise cosine distances between any two different news articles. Currently, for each news article, the process captures the article text as a string of characters, as well as article's accompanying metadata, such as the URL address of the article, the date of publication, the media source (e.g. BBC, RT), and other metadata offered by the RSS feeds technology (more on RSS later). I collect only news text data and exclude multimedia content, such as audio and video files, since this project concentrates on the news *text* bias. ~~Hence~~, this is an originally collected data set, not used by previously-published research.

My research necessitates a collection of texts from a wide range of news sources, since distance between texts conditional on different news sources is a central idea for this project. Additionally, unlike previous research, my measure of bias is not tied to external sources or contexts such as US congressional speeches to construct the metric. Thus, I do not collect data outside of the news articles and the associated metadata; the project is constructing the bias measure that relies on the news content and features only. Moreover, unlike Gentzkow and Shapiro (2010), I do not use news databases, making my collection method more economically-effective, which would hopefully make it more accessible to and replicable by a wider research audience.

The scraping of the articles is done through obtaining article URL's on the corresponding sites' Really Simple Syndication (RSS) feeds (Software Garden, Inc. 2004). While it is possible to scrape text content by other means (such as crawling through the sites' URL addresses), the project uses RSS feeds instead for two reasons.

First, it ensures that the news content from each website is captured in its fullness more systematically to reduce the chance of missing articles, since news sources publish all or most of their contents through RSS. This ensures I do not introduce selection bias by omitting articles. Thus, I capture an approximately full population of news articles from these corresponding sources in the specified time frame. When testing alternative methods, such as crawling through the URL links on the websites, the algorithm has hard time catching all article URL's and distinguishing URL's associated with news text versus URL's associated with other media content, such as videos, i.e. content I want to ignore.

Second, RSS feeds used by the chosen websites break down the news articles by broad topics, such as "US news" or "Sports." This is valuable because it may help offer broad validation of topics generated by the topic modelling procedure described in the methods section. Broad topic separation also allows us to explicitly exclude sections such as open-editorials. I avoid such sections because the project concentrates on news reporting and not on opinions that may or may not be representing news source's point of view (see Literature Review). Thus, RSS feeds present a better source for URL coverage and text capture.

I describe data characteristics and perform the analysis here based on the constrained version of the full data set for current computational feasibility. The constraint is articles collected through February 7, 2019. This yields 3327 unique article observations. The breakdown of article counts by data sources is included in Table 1 below.

Source	**Count**
BBC News	797
CNN	939
Fox News	760
RT	831
Total	3327

Table 1: Source Article Counts

To summarize, raw data generated by the collection process has these variables: date of publication, text as character strings (i.e. raw text of the news article), source of the article (e.g. "BBC", "RT"), URL (as the unique article identifier), and RSS Feed broad topic, and other RSS-provided metadata. Once I have the data, we are ready to process it in preparation to analysis.

First, in order to eventually generate topics, i.e. the independent variable, and work with text as data, one must first perform text cleaning and preprocessing (Grimmer and Stewart 2013). Outlining this procedure is important, because pre-processing choices can lead to non-trivial changes in the results during the analysis stage, as shown by Denny and Spirling (2018) in application to research in political science.

My cleaning procedure includes removing any web-publishing artifacts I can identify. For instance, an article may include hyperlinked text chunks like "Read more on our website" in the body of an article. This is not meaningful for the analysis. Hence, I remove such instances from article text. Other examples of distracting text artifacts include "image caption," "BBC website," "BBC Radio Live," "If you like this story, share it with a friend!" etc.[5]

Additionally, I remove other distracting material, such as all new line characters and e-mails and remove punctuations. Then I tokenize each article text, i.e. separate each sentence into words, and then remove stop words, such as "and," "them," "under" which is also a common procedure for treating text as data (Denny and Spirling 2018).[6] Using these tokens (words), for each article, I build a list of bi-, and tri-grams (in addition to single words), which are frequently occurring two and three word combinations for each observations. This is relevant to the research because I anticipate that some topics may include

[5] See the forthcoming code in the GitHub repository to see all text strings I remove
[6] Full list of stop words available at https://gist.github.com/sebleier/554280

multiple-word combinations like "President Trump," for example. Then, I lemmatize all tokens. That is, I conform words that are simply inflected forms to a common (lemma) form.[7] For instance, "better" and "good" should have the same "good" representation in our texts. This is relevant because without lemmatization, our topic modelling risks treating words that are otherwise close in meaning as conceptually different. I also only keep adverbs, nouns, verbs, and adjectives as I anticipate these being the most important in containing the meaning in the body of text.

After cleaning and word-preprocessing steps outlined above, the next step in analysis consists of generating predominant topics for each of the articles, which necessitates building a topic model. Currently, I am using Latent Dirichlet Allocation (LDA)[8] (Blei, Ng, and Jordan 2003), a probabilistic topic modelling technique to infer the most dominant topic for each observation (i.e. article). This allows me to engineer the independent variable, "predominant topic," for each news article.[9]

Since this is an unsupervised algorithm, one does not know a priori the optimal number of topics that the texts would cluster around. One way to decide an appropriate number is the coherence measure; higher coherence measure

[7] I use Spacy Python module for lemmatization and keeping only necessary parts of speech
[8] I may also experiment with an alternative, semi-supervised topic modelling technique, CorEx (Gallagher et al. 2016)
[9] See (Grimmer and Stewart 2013) for an overview of LDA

demonstrates a better topic number (Röder, Both, and Hinneburg 2015). Figure 2 is a result of training 38 different LDA topic models, each associated with a number of topics ranging from 2 to 40.

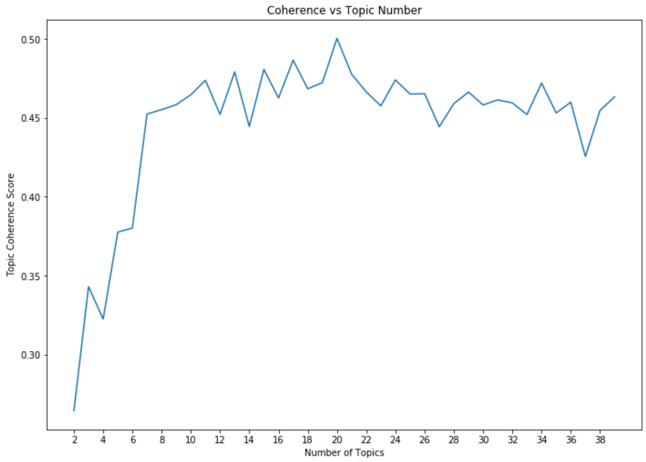


Figure 2: Coherence vs Number of Topics

Note that the coherence measure reaches its maximum and levels off at 20 topics. Thus, I choose the 20-topics LDA model for further analysis.

LDA assumes that each topic is composed of words (lemmas), with each word responsible for a proportion of a topic. Table 2 contains a list of the 2 of the 20 topics[10], with the listing of each top lemmas associated with a given topic and the corresponding weights within the topic. A larger topic table is included in the Appendix. Topic labels in the "Topic" column are researcher's interpretations based on the lemmas within each topic and partial manual reading of news

[10] There is a technical issue printing out all 20 topics' probabilities that is yet to be resolved.

articles in which the topic is dominant. Note that some topics are more human-interpretable than others. Topic labels that are less interpretable are appended with the question mark in the Appendix.

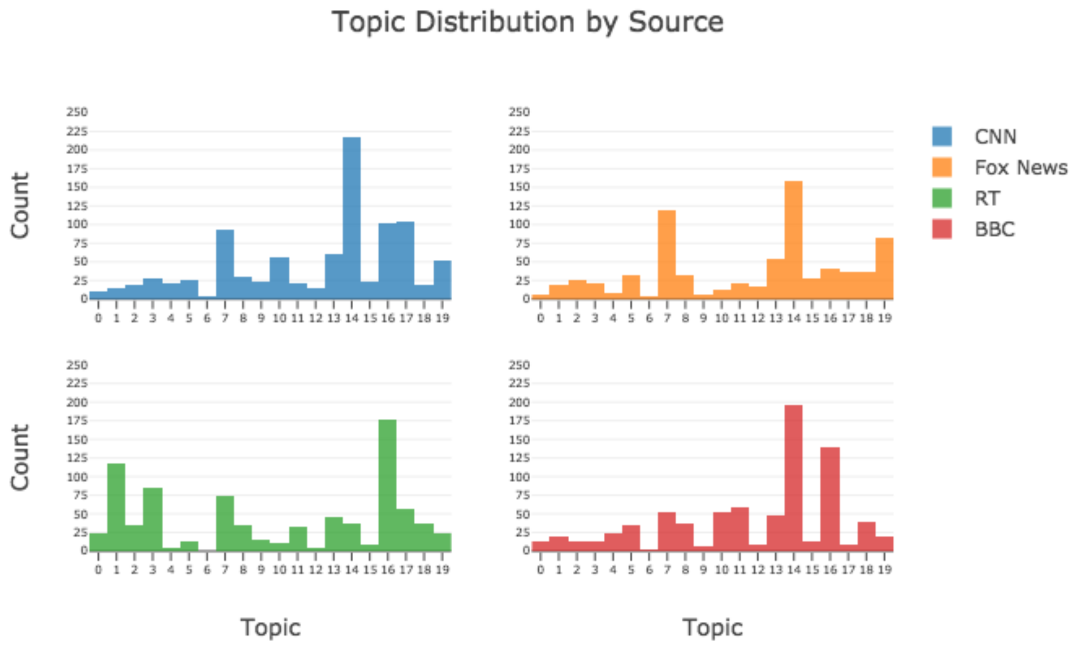
Table 2: Topic Modeling Interpretation Example

Topic	Lemma	Weight	Topic	Lemma	Weight
Healthcare	health	0.021356275	**Virginia Scandal**	woman	0.03458775
	drug	0.017169613		white	0.01401055
	patient	0.014966376		northam	0.013049183
	hospital	0.014058208		racist	0.012259012
	doctor	0.013041544		blackface	0.012191094
	medical	0.012904156		state	0.011485968
	death	0.011811129		abortion	0.010237206
	child	0.011142593		statement	0.009424654
	die	0.009480785		page	0.008444632
	accord	0.009249739		black	0.008390697
	case	0.007865488		law	0.008121205
	marijuana	0.007396302		racism	0.007457194
	state	0.006826073		female	0.007386289
	disease	0.006820308		photo	0.007292737
	care	0.006795365		resign	0.007125538
	virus	0.006617869		allegation	0.007048612
	treatment	0.006525188		call	0.006940115
	physician	0.005527902		man	0.006812302
	report	0.005455626		male	0.006213304
	colorado	0.005275759		fox_news	0.00565377

Once I assign a predominant to each article, we can see the distribution of topic for each source (Figure 3).

Second, to construct the independent variable, cosine distances, I vectorize article texts using word2vec word embeddings, which represent each text string in its vector form as extracted from a vector space that represents latent semantic content of news articles (Mikolov et al. 2013). There are two reasons for using these pre-trained word embeddings. First, to train own word embeddings,

Figure 3: Topic Count Distribution per Source



one would need a very large corpus (of news, in this case) (Chollet 2017). Since the article dataset for the purposes of this analysis has only 3327 observations, I resort to pretrained word embeddings. Second, the embeddings I use were trained on Google News. This is appropriate as the vector space of word embeddings is tied to the language context (Chollet 2017). Since the embeddings were trained on 100 billions words from Google News, I expect that these embeddings would be appropriate to this project's context (Google 2013).

This allows me to construct the cosine difference, the dependent variable. Cosine difference is $1 - (cosine\ similarity)$. For computation purposes, I use cosine similarity, but the statistical results do not differ as statistical tests for difference in cosine distances (i.e. bias) are mathematically equivalent to tests in cosine similarities. The interpretation is that two "similar" news articles would have "little" bias, as defined by this study. The bias measure (cosine difference) between two news articles in their vector representations $\vec{u}$ and $\vec{v}$ is:

$$1 - \frac{\vec{u} * \vec{v}}{||\vec{u}|| * ||\vec{v}||}$$

Where vector norms used in the denominator are L2-norms. Since there are 3327 article observations, I compute a total of $\binom{3327}{2} = 5532801$ pairwise article cosine similarities.

For the purposes of the preliminary analysis, I use Welch's T-test for difference in means in cosine distances between pairs of news sources, conditional on topics. However, for the full version of the analysis, I anticipate experimenting with multiple linear regression and supervised machine learning methods to infer a function that maps article sources and topics onto values of cosine distances (more on that in the Discussion section).

Analysis:

For the purposes of this preliminary analysis, let us look at CNN as the baseline standard for bias measure and Fox News and RT as compared to CNN and constrain the search to topics "Crime" and "US Politics."

Table 3 reports summary statistics on cosine similarities with the data constrained to the conditions in the previous paragraph.

Table 3: CNN to Target Source Cosine Similarity Summary Statistics

topic	target source	count	mean	std	min	25%	50%	75%	max
Crime	Fox News	11160	0.829556933	0.05382656	0.546680808	0.799793571	0.835678518	0.86743091	0.983913898
	RT	6789	0.839204574	0.048338039	0.622709155	0.811861634	0.844044447	0.872817934	0.970934093
US Politics	Fox News	3848	0.883838486	0.053207771	0.616143465	0.859033167	0.89343968	0.919283092	0.986397505
	RT	11856	0.880364607	0.048124528	0.603044808	0.857095465	0.889096975	0.914100915	0.977500319

For instance, on the topic of "Crime," there are 11160 cosine similarities between CNN and Fox News articles, with an average similarity value of 0.82956 and standard deviation of 0.05384, with values ranging from 0.54668 and 0.98391. So, with constraints, there are a total of 33653 cosine similarity observations.

For this analysis, I would hypothesize that, on average, the cosine similarity between CNN and Fox News articles to be different from between Crime and US Politics topics. One would expect that, since US Politics should be a more polarizing topic. On the other hand, when writing on Crime, the two sources should be closer on average, so the mean difference (and similarity) should be significantly lower, or different from the US Politics. Similarly, average similarities between CNN and RT should be different conditional on the topic. So I perform two unequal variance t-tests, one between the two similarity means of CNN and Fox News and one between CNN and RT. More formally, the null hypothesis is that there is no difference between the two cosine similarity means within each target source. Alternative hypothesis is that there is a statistical difference. The assumption of normal distributions of sample values is acceptable due to a large sample size and equal variances are not assumed.

Table 4: Welch's 2 Sample T-Tests

	Mean Difference	Test Statistic	p-value
CNN vs Fox News	-0.05428	-54.40851	0.000
CNN vs RT	-0.04115	-56.03709	0.000

Based on results in Table 4, we can reject the null hypotheses of no difference. There is a statistically significant difference within each set of means, for both t-tests (at alpha significance level of 0.005).

Discussion:

In terms of interpretation of the results, there is a confirmation of the intuition outlined above and the negative direction of the mean difference makes sense (i.e. cosine dissimilarity is lower for the "Crime" topic than for "U.S. Politics"). However, as a counterpoint, one thing to consider is the practical interpretation of the small (albeit statistically significant) mean difference. Would such small mean differences between the two categories for CNN vs Fox News and CNN vs RT (-.05 and -0.04 respectively) have an impact in terms of a reader being able to tell the difference in content? Would a reader be able to tell the difference in similarity by reading articles between two topics? What does a value of -0.05 mean practically for a news reader? This is unclear based on the design of the study and is a key limitation for the project.

Another limitation is that the analysis performed above is quite rudimentary in terms of statistical machinery and is also limited in that it is constrained to a small subset of the data (in both sources and topic categories). Therefore, a natural extension of this analysis will be expanding the data set to all topics and all news sources. This would mean including BBC and expanding to eighteen more topics (though the topic number change as I will likely redo topic modelling on all scraped texts). However, for more feasible statistical analysis, one can still keep the analysis constrained to one baseline source (as CNN in the preliminary analysis). Moreover, one more variable not considered above is the time of publication, a data point that I have also been extracting during data gathering.

The expansion of the analysis may naturally necessitate different statistical methods, because the use of many isolated t-tests may become insufficient.[11] One way to do so is to use supervised machine learning to infer a function that takes in constructed dominant topics and news sources to predict the continuous dependent variable values of cosine similarities. A successful

[11] Though I would need further consultation on what other more sophisticated statistical techniques may be needed.

model should be able to capture this mapping, suggesting that topics and sources may play a nontrivial role in determining news bias.

Another extension of the current methodology is to account for the “dominance” of a topic. Currently, the project only accounts for what the dominant topic is for each article, but not to what extent that article is dominated by the dominant article. For a more nuanced approach, this project may potentially benefit from taking further advantage of topic modelling. Since a topic model represents a statistical distribution of topics and each article is composed of a mixture of topics, perhaps the analysis should account for the extent a topic is dominant for each article and the extent to which lesser topics comprise the article. Further work is needed in incorporating this information into statistical analysis. Additionally, the research lacks any validation on the topic model. Further steps are needed in making sure the models are not simply picking up ‘noise’ in the data.

One challenge is the interpretability of topics coming out of the LDA topic model. This is especially important since the hypotheses are based on the researcher’s interpretations of the topics. When topic interpretation is vague or unclear, as is the case with the “Europe/Religion” or “Miscellaneous” topic seen in the Appendix, it is unclear what hypotheses to posit for further analysis.

Furthermore, this poses additional challenges for the traditional informing of hypotheses by the literature. In a more conventional social science setting, one may incorporate previous theoretical research to inform the formation of hypotheses prior to analysis. In this project, I perform partial analysis by training the topic model, and then formulate loose hypotheses based on general intuition about what one might expect in terms of content divergence conditional on topics. But this approach leaves formal hypothesis formulation at the mercy of unsupervised learning of LDA modelling and the interpretability of topics.

One complementary approach could be to validate topics with the broad topic categories captured through the RSS scraping. Each RSS feed is associated with a particular broad news topic. For instance, CNN feeds might be “US News”, “International News”, “Sports”, etc. Analyzing to what extent interpreted topics from topic modelling successfully fall into these broad categories may be one way to partially validate the topic modelling mechanism.

Another approach to extract topics from text is Named Entity Recognition techniques from NLP (Finkel, Grenager, and Manning 2005). Extracting named entities, such as “Donald Trump” from article text or article titles may approximate extraction of topics in a more interpretable way. It remains to be seen if this is a better approach than topic modelling or is perhaps a good validation mechanism for the topics.

Another important extension needed for the project is validating text-preprocessing. For instance, as Denny and Spirling (2018) suggest in application to text-as-data methods in Political Science, text preprocessing steps such as lemmatization and tokenization of words may have non-trivial consequences for the later analysis. The authors also provide a software tool to measure such sensitivity mathematically. Employing this method can contribute to validating this project's methodology.

Additionally, based on the survey of literature, this research may be subject to several limitations more generally. First, this research does not account for the subjective interpretation of news bias by an audience. So far, the scope of the project excludes consideration of bias from the reader's perspective. Such considerations may be important, because previous research has shown that reader's subjective perspective as well as individuals' characteristics may influence the interpretation of news content.

For instance, Gunther and Schmitt (2004) design an experimental study and find that reader's knowledge that a text content is from a news source (as opposed to a student's school essay) is associated with this reader perceiving the content to be more biased. Furthermore, in a different study Gunther (1992) finds that group identification/membership (e.g. born-again Christians) is associated with whether a reader perceives a news source to be biased (higher involvement in a group may lead to higher perception of bias). African-Americans may also perceive the media to be less biased compared to White Americans (Becker, Kosicki, and Jones 1992). Relatedly but in a different context, Chiang and Knight (2011) find that newspaper endorsements for candidates are more influential on moderate voters. In a US political context, there may also be a general perception of a left-leaning bias in the US media by the population generally (Watts et al. 1999). In other words, individual and audience characteristics seem to matter for studying bias.

Despite this prior literature, this project excludes reader's characteristics and subjective interpretation from the research scope. There are two justifications for this decision. The first and a more logistical reason is resource and time limits. Studying individual subjective interpretations by readers would likely extend this study to interacting with human subjects. This may require financial expenditures to attract potential subjects as well as more extensive precautions to protect the subjects' rights (through, for instance, an IRB approval). Since I perceive that meeting such requirements to a practical and ethically acceptable standard may be difficult within the time period and budget allotted, I exclude this aspect from the study.

The second and more substantive reason is it may make more sense for this study to use human subjects for the validation of methods instead at a later time. In other words, if individual characteristics and subjective interpretation do indeed matter for the definition of bias *and* my methods for operationalizing bias *independently* from human interpretations is valid, then human subjects should be able to subjectively evaluate a distance between two news articles or sources after reading and such subjective measurements should correspond well with my proposed bias metric. Indeed, Iyengar and Hahn (2009) find that individuals select their news sources based on ideological alignment (e.g. Conservatives/Republicans choose Fox News over NPR and CNN).

Appendix

Topic	**Lemma**	**Weight**
Healthcare	health	0.021356275
	drug	0.017169613
	patient	0.014966376
	hospital	0.014058208
	doctor	0.013041544
	medical	0.012904156
	death	0.011811129
	child	0.011142593
	die	0.009480785
	accord	0.009249739
	case	0.007865488
	marijuana	0.007396302
	state	0.006826073
	disease	0.006820308
	care	0.006795365
	virus	0.006617869
	treatment	0.006525188
	physician	0.005527902
	report	0.005455626
	colorado	0.005275759
Virginia Scandal	woman	0.03458775
	white	0.01401055
	northam	0.013049183
	racist	0.012259012
	blackface	0.012191094
	state	0.011485968
	abortion	0.010237206
	statement	0.009424654
	page	0.008444632
	black	0.008390697
	law	0.008121205
	racism	0.007457194
	female	0.007386289
	photo	0.007292737

	resign	0.007125538
	allegation	0.007048612
	call	0.006940115
	man	0.006812302
	male	0.006213304
	fox_new	0.00565377
Internet	company	0.032780383
	facebook	0.02512524
	user	0.02503532
	datum	0.024067476
	use	0.019326303
	app	0.012956737
	apple	0.008173591
	location	0.00787725
	bezos	0.007579328
	device	0.007099449
	account	0.0068419
	information	0.006777159
	content	0.006568655
	online	0.006307459
	ad	0.006261959
	amazon	0.006236916
	platform	0.006192211
	technology	0.006115128
	access	0.006042211
	service	0.006024787
Film Entertainment	film	0.029596165
	star	0.014588127
	show	0.014189566
	actor	0.007148785
	award	0.0070799
	movie	0.006924432
	viewer	0.006377359
	character	0.006114045
	director	0.006041962
	gold	0.006004583
	screen	0.005792552

	singer	0.0055738
	art	0.005544589
	series	0.005529801
	jackson	0.005524482
	tv	0.005436325
	feature	0.005275256
	documentary	0.005188048
	oscar	0.005068018
	audience	0.005006113
Europe/Religion (?)	percent	0.022427117
	france	0.012742646
	migrant	0.012226461
	agent	0.011093353
	italian	0.010313642
	indian	0.009704245
	christian	0.008950181
	brown	0.008346878
	church	0.007707134
	client	0.006923214
	muslim	0.00682775
	pope	0.006379222
	religious	0.006205661
	local	0.006123465
	india	0.006082493
	french	0.005540685
	melenchon	0.00534454
	rap	0.005307347
	rally	0.004812846
	yellow_v	0.004779587
Social Media	post	0.018759077
	show	0.011776809
	write	0.011317118
	twitter	0.011096339
	tweet	0.010678786
	photo	0.010215716
	social_media	0.009319095
	picture	0.008589178

	claim	0.008389646
	appear	0.007832815
	black	0.00766507
	comment	0.007629905
	book	0.007580692
	story	0.007560819
	news	0.006640774
	call	0.00648081
	medium	0.005961065
	share	0.005669491
	message	0.005569206
	white	0.005273922
Crime	say	0.03543368
	police	0.017489148
	report	0.01704276
	investigation	0.010634722
	tell	0.009418309
	case	0.008900867
	man	0.008822541
	court	0.008685392
	charge	0.008662362
	accord	0.008612257
	officer	0.00778858
	statement	0.007241639
	incident	0.007031478
	arrest	0.006884985
	release	0.006214471
	claim	0.006078404
	kill	0.005821042
	authority	0.005523318
	official	0.005438603
	department	0.005374842
Military Technology	new	0.015843192
	use	0.007960635
	system	0.007286145
	could	0.006790177
	build	0.005663232

	large	0.005628427
	make	0.005056441
	say	0.004959554
	need	0.004699855
	test	0.004691333
	missile	0.004619563
	time	0.004604614
	provide	0.004378165
	work	0.004273396
	technology	0.004106806
	project	0.003706465
	create	0.003641293
	part	0.003542974
	small	0.003276763
	year	0.00319213
Miscellaneous (?)	say	0.042327397
	not	0.023021588
	get	0.016169745
	go	0.016026348
	people	0.013929803
	do	0.013088849
	time	0.010979071
	know	0.010479162
	be	0.01028548
	tell	0.010213814
	s	0.010200537
	make	0.00943996
	think	0.009119026
	take	0.009084662
	want	0.008767445
	see	0.008444416
	would	0.008057352
	work	0.007879037
	come	0.006913245
	thing	0.006682945
Government	say	0.040766444

	would	0.018493077
	could	0.008005157
	make	0.007651093
	deal	0.007466237
	take	0.006904717
	plan	0.00637016
	add	0.006093674
	government	0.005916797
	year	0.005739184
	come	0.005267058
	issue	0.005138557
	time	0.00488225
	decision	0.004870927
	change	0.004532074
	thursday	0.004511569
	however	0.004392966
	support	0.004349074
	tell	0.004121621
	leave	0.003997585

Bibliography

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–35. https://doi.org/10.1257/jep.31.2.211.

BBC. 2018. "BBC - Homepage." 2018. http://www.bbc.com/.

Becker, Lb, Gm Kosicki, and F. Jones. 1992. "Racial-Differences in Evaluations of the Mass-Media." *Journalism Quarterly* 69 (1): 124–34. https://doi.org/10.1177/107769909206900110.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chiang, Chun-Fang, and Brian Knight. 2011. "Media Bias and Influence: Evidence from Newspaper Endorsements." *The Review of Economic Studies* 78 (3): 795–820. https://doi.org/10.1093/restud/rdq037.

Chollet, Francois. 2017. *Deep Learning with Python.* 1 edition. Shelter Island, New York: Manning Publications.

CNN. 2018. "CNN - Breaking News, Latest News and Videos." CNN. 2018. https://www.cnn.com.

Cook, Timothy E. 2005. *Governing With the News, Second Edition: The News Media as a Political Institution.* Second edition. Chicago: University of Chicago Press.

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *The Quarterly Journal of Economics* 122 (3): 1187–1234. https://doi.org/10.1162/qjec.122.3.1187.

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26 (2): 168–89. https://doi.org/10.1017/pan.2017.44.

Engelberg, Joseph E., and Christopher A. Parsons. 2011. "The Causal Impact of Media in Financial Markets." *Journal of Finance* 66 (1): 67–97. https://doi.org/10.1111/j.1540-6261.2010.01626.x.

Entman, Robert M. 2007a. "Framing Bias: Media in the Distribution of Power." *Journal of Communication* 57 (1): 163–173.

———. 2007b. "Framing Bias: Media in the Distribution of Power." *Journal of Communication* 57 (1): 163–73. https://doi.org/10.1111/j.1460-2466.2006.00336.x.

Feldman, Lauren, Edward W. Maibach, Connie Roser-Renouf, and Anthony Leiserowitz. 2012. "Climate on Cable: The Nature and Impact of Global

Warming Coverage on Fox News, CNN, and MSNBC." *International Journal of Press-Politics* 17 (1): 3–31. https://doi.org/10.1177/1940161211425410.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 363–370. Association for Computational Linguistics.

Fox News. 2018. "Fox News." Fox News. 2018. https://www.foxnews.com.

Gallagher, Ryan J., Kyle Reing, David Kale, and Greg Ver Steeg. 2016. "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." *arXiv:1611.10277 [Cs, Math, Stat]*, November. http://arxiv.org/abs/1611.10277.

Galtung, Johan, and Mari Holmboe Ruge. 1965. "The Structure of Foreign News." *Journal of Peace Research* 2 (1): 64–91.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78 (1): 35–71.

Google. 2013. "Google Code Archive - Long-Term Storage for Google Code Project Hosting." 2013. https://code.google.com/archive/p/word2vec/.

Grimmer, Justin. 2013. "Appropriators Not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57 (3): 624–42. https://doi.org/10.1111/ajps.12000.

———. 2016. "Measuring Representational Style in the House: The Tea Party, Obama, and Legislators' Changing Expressed Priorities." Computational Social Science: Discovery and Prediction. March 2016. https://doi.org/10.1017/CBO9781316257340.010.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *The Quarterly Journal of Economics* 120 (4): 1191–1237. https://doi.org/10.1162/003355305775097542.

Gunther, Albert C. 1992. "Biased Press or Biased Public? Attitudes Toward Media Coverage of Social Groups." *The Public Opinion Quarterly* 56 (2): 147–67.

Gunther, Albert C., and Kathleen Schmitt. 2004. “Mapping Boundaries of the Hostile Media Effect.” *Journal of Communication* 54 (1): 55–70. https://doi.org/10.1111/j.1460-2466.2004.tb02613.x.

Gurun, Umit G., and Alexander W. Butler. 2012. “Don’t Believe the Hype: Local Media Slant, Local Advertising, and Firm Value.” *The Journal of Finance* 67 (2): 561–98. https://doi.org/10.1111/j.1540-6261.2012.01725.x.

Harcup, Tony, and Deirdre O’Neill. 2001. “What Is News? Galtung and Ruge Revisited.” *Journalism Studies* 2 (2): 261–80. https://doi.org/10.1080/14616700118449.

———. 2017. “What Is News?: News Values Revisited (Again).” *Journalism Studies* 18 (12): 1470–88. https://doi.org/10.1080/1461670X.2016.1150193.

Iyengar, Shanto, and Kyu S. Hahn. 2009. “Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use.” *Journal of Communication* 59 (1): 19-U6. https://doi.org/10.1111/j.1460-2466.2008.01402.x.

Kahn, Kim Fridkin, and Patrick J. Kenney. 2002. “The Slant of the News: How Editorial Endorsements Influence Campaign Coverage and Citizens’ Views of Candidates.” *American Political Science Review* 96 (2): 381–94. https://doi.org/10.1017/S0003055402000230.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. “Computer-Assisted Text Analysis for Comparative Politics | Political Analysis | Cambridge Core.” 2015. https://www-cambridge-org.proxy.uchicago.edu/core/journals/political-analysis/article/computerassisted-text-analysis-for-comparative-politics/CC8B2CF63A8CC36FE00A13F9839F92BB.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv:1301.3781 [Cs]*, January. http://arxiv.org/abs/1301.3781.

NewsLibrary. 2018. “NewsLibrary.com - Newspaper Archive, Clipping Service - Newspapers and Other News Sources.” 2018. http://nl.newsbank.com/nl-search/we/Archives?p_product=NewsLibrary&p_action=keyword&p_theme=newslibrary2&p_queryname=4000&s_home=home&s_sources=location&p_clear_search=&s_search_type=keyword&s_place=&d_refprod=NewsLibrary.

Niven, David. 2003. “Objective Evidence on Media Bias: Newspaper Coverage of Congressional Party Switchers.” *Journalism & Mass Communication Quarterly* 80 (2): 311–326.

Peffley, M., T. Shields, and B. Williams. 1996. "The Intersection of Race and Crime in Television News Stories: An Experimental Study." *Political Communication* 13 (3): 309–27. https://doi.org/10.1080/10584609.1996.9963120.

ProQuest. 2018. "ProQuest | Databases, EBooks and Technology for Research." 2018. http://www.proquest.com/.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2014. "Stm: R Package for Structural Topic Models." *R Package* 1: 12.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. ACM.

RT. 2018. "RT." RT International. 2018. https://www.rt.com.

Smith, Jackie, John D. McCarthy, Clark McPhail, and Boguslaw Augustyn. 2001. "From Protest to Agenda Building: Description Bias in Media Coverage of Protest Events in Washington, D.C." *Social Forces* 79 (4): 1397–1423.

Software Garden, Inc. 2004. "What Is RSS: A Tutorial Introduction to Feeds and Aggregators." July 4, 2004. http://rss.softwaregarden.com/aboutrss.html.

Terman, Rochelle. 2017. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage." *International Studies Quarterly* 61 (3): 489–502. https://doi.org/10.1093/isq/sqx051.

Watts, Mark D., David Domke, Dhavan V. Shah, and David P. Fan. 1999. "Elite Cues and Media Bias in Presidential Campaigns: Explaining Public Perceptions of a Liberal Press." *Communication Research* 26 (2): 144–175.