Controversy Score Calculation for News Articles

Paul Kim, Ziyu Fan*, Lance Fernando*, Jacques Sham*, Crystal Sun*, Yixin Sun*,
Brian Wright*, Xi Yang*, Nicholas Ross, Diane Myung-kyung Woodbridge
{ykim57,zfan18,ljfernando,jjsham,xsun45,ysun84,bawright3,xyang68,ncross,dwoodbridge}@usfca.edu
Data Science Program
University of San Francisco

***Abstract*—From the 2016 election continuing to the present, a strong focus has developed on the US news media landscape. From issues such as fake news to ideological wars between news sources, helping readers become more discerning news consumers remains a priority for technologists and journalists alike. This paper is a summary of our contribution to the continuing discussion of public empowerment towards making more informed news consumption decisions. In this paper we introduce NewsPhi, a news-feed application designed to provide topic-based contextual information on a changing corpus of daily news. NewsPhi's main contribution is a controversy score of news topics - a score that summarizes how controversial a certain news topic is according to the larger news media environment. We find that this approach to building products for news readers is both unique and in line with ongoing bias and opinion formation research in the social sciences. Rather than seeking to correct or standardize news opinions, NewsPhi provides important contextual information of opinions on existing and new topics while also lightly encouraging further exploration.**

***Index Terms*—Natural language processing, Unsupervised learning, Information Science, Information retrieval**

I. Introduction

Following the 2016 U.S. presidential election, the meta-narrative surrounding US news media environments became incredibly robust. Bad faith actors, domestic and international, gained political influence through the employment of news that is "intentionally and verifiably fake and could mislead readers" [1] - a phenomenon otherwise known as "fake news". It has been speculated that fake news, in conjunction with the clever usage of social media tactics, may have influenced the most recent presidential election - and, perhaps more importantly, how parts of the public interact with the news at large. The discovery and enlightening of these disinformation tactics has led to critical concerns regarding the US news media and political landscapes; namely, a growing distrust of mainstream media outlets and an unwillingness to seriously consider or care for factual legitimacy [1]. These issues go hand in hand and often augment each other. A distrust for mainstream media (consisted of for the large part vetted, rigorous news organizations) leads to a reliance on alternative news sources; alternative news sources claim that their outsider status gives access to certain truths that mainstream media sources will not disclose out of either self-interest or ideologically conformist behavior.

Our interest in this crucial point in news history is not to address fake news itself, though fake news is an important starting point for understanding how political polarization looks in 2019. We are concerned with the persistent ideological split that seemingly strengthened after the election, admittedly partially as a result of fake news and fake news-related consequences.

Various studies have presented that fake news do have a strong polarizing effect. Guess, Nyhan, and Reifler found that Trump supporters are "disproportionately more likely to consume pro-Trump fake news and less likely to consume pro-Clinton fake news relative to Clinton supporters, supporting a selective exposure account" [2]. The same authors in 2018 found that, similar to in 2016, fake news consumption was concentrated among the top 10% of Americans that consumed the most conservative media. They also found that, overall, fake news readership decreased about 75% (About 1 in 4 Americans read a fake news article in fall 2016 compared to about 7% in fall 2018) [3]. Further, these authors found that though a majority of the public were able to distinguish between authentic and fake news stories, a significant minority believed fake/hyperpartisan news were accurate. This percentage was higher when the partisanship of the news matched that of the consumer (17 - 48% vs. 10 - 22% for when the partisanship did not match), hearkening the previous concerns of extreme partisanship - but only for a minority population. Fake news indeed is the modern beginning point of how we have come to understand what now appears to be an unbridgeable gap between ideological groups; however, it remains to be proven that these extreme ideological groups are very large, diverse, or even equal in size. One news-focused approach has been to attempt to bridge this gap by exposing people to differing opinions from different sides of the left-right spectrum. We took a couple of issues with this approach.

Our first issue concerns bias. Bias is not so easy to overcome through this sort of meek exposure. Guess, Nyhan, and Reifler's research indicates that partisans are more likely to believe in partisan conspiracy theories and that fake news consumption was mostly limited to a small outlier group [3]. Confirmation bias is a strong (and often rational, according to Susan and Jack Gorman [4]) instinct to overcome. Oswald and Grosjean acknowledge the effects of confirmation bias when referring to affirmation of "motivationally supported" hypotheses [5]. Our approach then is not to combat fake news by providing a quantitative score on how credible a news

* These authors contributed equally.

source is. Consumer applications do not appear to show much promise in swaying those heavily set in opinions, rational or not. There exist attempts to judge how reliable a news source is - our concern is not to sway people from their biases regarding news reliability or truthfulness, given the existing social science research. Fake news are highly unique and are constructed to benefit from existing social media infrastructure [6], thusly endemic to the information sharing ecosystem logic itself. It is difficult to introduce a new paradigm of news consumption or to disrupt those existing, and to combat fake news effectively, it is likely more important for social media sites to continue ongoing processes to mitigate disinformation (efforts which have taken and are currently taking place in bot detection, content moderation, and more - [7]). Thus, we do not try to prescribe news objectivity. Rather, we are interested in informing readers how the news media feels in aggregate on specific topics. We are interested in providing topic-based context, especially for news topics that are potentially new to consumers. This mission is born out of a twofold reasoning: first, a genuine interest and passion for news education, and second, further research on political opinion formation and changing.

Our approach to the current political landscape is informed by Carsey and Layman's 2006 study on changing party affiliations [8]. There are two crucial elements to whether Carsey and Layman would consider an individual to possibly change parties: that individual's knowledge of party difference and issue-specific granularity. According to Carsey and Layman's work, individuals that find certain issues very important are more likely to change party identification according to each party's stance on those issues. Providing topic-based context aggregated over a large number of publications of different political stances is, then, key. We are able to effectively display party difference through the combination of reader knowledge/perception of a publication and our "topic controversy score" (to be discussed soon). If we consider topics as issues (as there were a significant number of topics that could be considered policy issues in our modeling), we are able to provide issue-level granularity to understanding of politics.

Our second issue with the exposure approach we consider equally important but less immediately necessary is the complexity of assigning political bias. The allocation of news sources into "liberal" and "conservative" or "left" and "right" buckets does not provide distinction between vastly differing political stances within these general buckets. For example: the Left is a broad political group that contains liberal internationalism, anarchism, globalism, and other incongruous/less congruous ideologies. There is a commonality in research and applications we reviewed that emphasized or reinforced a binary understanding of political positioning. Though we acknowledge that this is a useful and necessary delineation, we wanted to introduce a further level of nuance to news topics. Our goal is to help readers understand that a certain publication may feel a certain way about a certain topic that goes against the general opinion of its relative position on the left-right spectrum, allowing for a multidimensional and richer understanding of the various pieces of the news media. As we are developing a customer product, there is an opportunity available for us to paradigmatically influence how the public understands politics, political structure, and the news media.

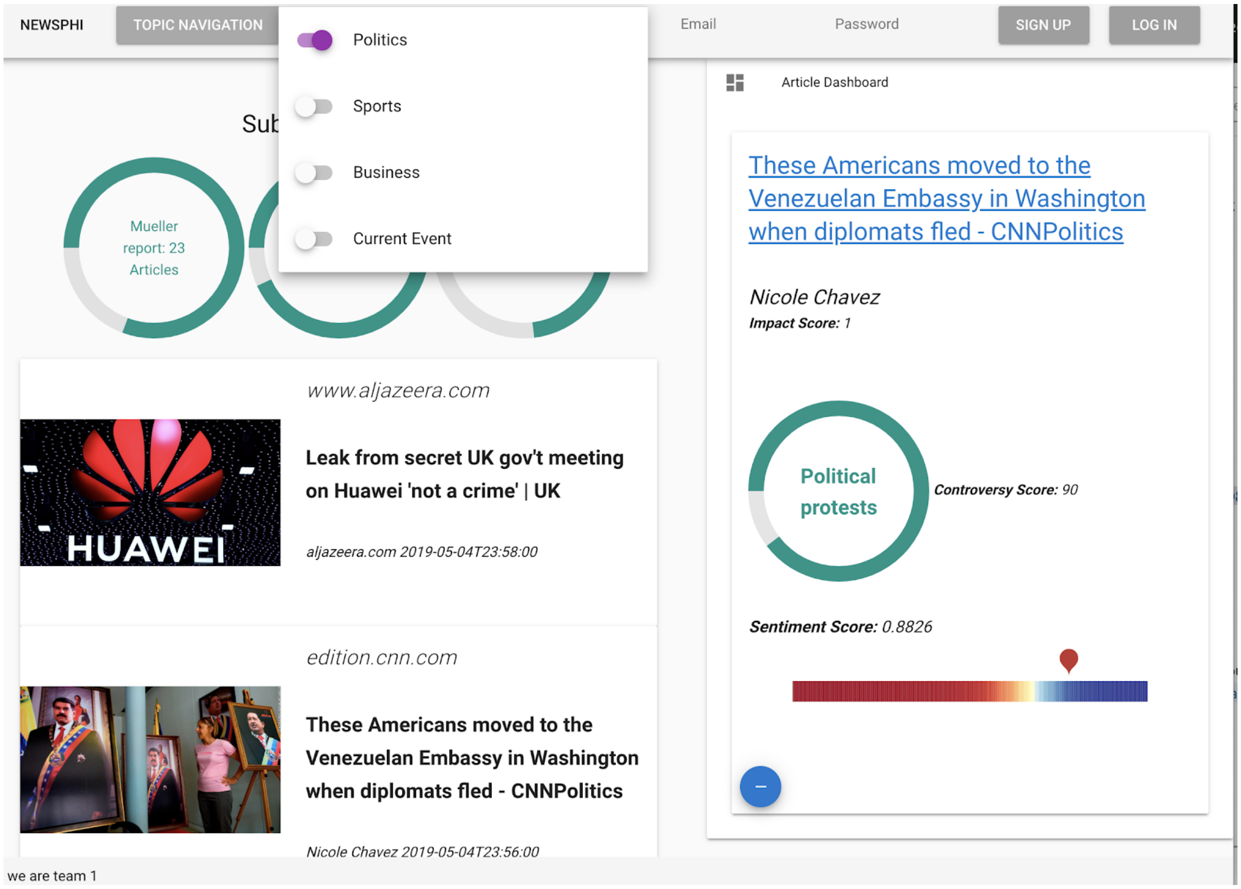
Guided by these dual goals, we created a web application called NewsPhi. NewsPhi is an interactive news-feed designed to help consumers better understand the overall context of the news that they are reading. The main feature of NewsPhi is the topic controversy score. Unlike scoring bias or veracity of news sites, our goal is to give an aggregated score of the level of agreement/disagreement (thus, controversy) on a given number of news topics per day. The driving idea is a summation of the above. It is useful and insightful for the public to understand the full spectrum of opinion on news topics, especially as new topics crop up. If we consider topics as policies, there is an opportunity for a reduction in partisanship among non-outlier groups of news consumers. Finally, if we can successfully delineate topics, we can help the public gain a more holistic understanding of news publications. This approach requires a multitude of moving parts: normal extract, transform and load (ETL) processes, topic modelling, sentiment analysis, and the setting up of servers and user interface for web hosting. Explicitly, our product is an interactive news-feed that allows for filtering by topics. It gives topic-level aggregate analysis - the previously mentioned controversy score - as well as a short article-level analysis - the sentiment score of that specific article in comparison to the whole. Figure 1 is what the front page of NewsPhi looks like.

II. Related Work

There are a number of existing applications that support educational efforts surrounding modern news consumption. We acknowledge these existing efforts in the following similarities and dissimilarities:

A. News credibility and bias

Many technological efforts in the news space following the discovery of fake news were made towards identifying and quantifying objectivity. OwlFactor assigns a credibility score based on four categories: "Site quality", "Author expertise", "References quality", and "Tone" [9]. NewsPhi operates similarly to OwlFactor in that it relies on machine learning to assign attributes on an article basis. However, combating fake news, again, is not explicitly the aim of NewsPhi, nor is it in NewsPhi's interests to quantitatively decide which sites are considered to be of a certain quality. Additionally, OwlFactor uses categorical methods to assign biases to news sources, relying initially on binary bias ratings from Allsides [10] and Media Bias Fact Check [11], then moving towards a machine learning approach to identifying "slant", guided by a paper by Gentzkow and Shapiro [12]. Allsides employs a crowd-sourced aggregation of blind bias ratings in addition to third party research (if available) in order to assign one of five bias ratings: Left, Lean Left, Center, Lean Right, and Right. Media Bias Fact Check uses a volunteer team of five to qualitatively assess



www.aljazeera.com

Leak from secret UK gov't meeting on Huawei 'not a crime' | UK

aljazeera.com 2019-05-04T23:58:00

edition.cnn.com

These Americans moved to the Venezuelan Embassy in Washington when diplomats fled - CNNPolitics

Nicole Chavez 2019-05-04T23:56:00

These Americans moved to the Venezuelan Embassy in Washington when diplomats fled - CNNPolitics

Nicole Chavez

Impact Score: 1

Political protests

Controversy Score: 90

Sentiment Score: 0.8826

we are team 1

Fig. 1: **An example of the NewsPhi interface.** At the top is the drop down menu, which filters the articles based on broader topics. Individual articles can be seen on the left. Upon click, the right sidebar populates with the article hyperlink, the topic, and the controversy score of the topic. The donut chart indicates how "controversial" this topic is. Below the donut chart is the sentiment score of that specific article, along with a visual of where the specific article sentiment lies on the spectrum of topic-wise article sentiment. This spectrum of topic-wise article sentiment displays how, in general, the aggregate of our news sources feels about the topic. It should be noted that the controversy score for this particular topic, "political protests", has been updated.

a source's bias into the same categories (with the addition of Extreme Left and Extreme Right). OwlFactor is currently transitioning into a machine learning approach of assigning bias based off of natural language processing methods outlined by Gentzkow and Shapiro [12]. Gentzkow and Shapiro identify "media slant" by comparing phrase frequencies in newspapers to phrase frequencies in the 2005 Congressional Records.

AllSides itself also provides a news-feed in addition to their bias methodology [13]. The same categories described above are employed in this feed, dividing news articles by their assigned bias categories. Each article has a display that shows where on this left-right spectrum its home publication lies. AllSides also lists topics for news articles, to be explored further in the following subsection.

A couple reasons our approach avoided categorizing news sources into biases were the difficulty and subjectivity of identifying and labeling these biases as well as the suspect assumption that consumers will be swayed by exposure to opposing viewpoints in this unstructured approach. Regarding the former concern: As Andrew Heywood argues, the left-right spectrum is restrictive as it does not allow for understanding of division within the left and right, respectively [14]. For example, a publication could be considered fiscally left but socially right. This publication would have left views for certain topics but right views for others. This nuance is hard to parse when using the previously mentioned political categories. Our goal by not presenting a bias assessment is to complicate the common understanding of bias and its relationship to institution and thusly introduce a more holistic approach to judging news media. As for the latter concern of the efficacy of exposure: as discussed in the introduction, we believe that a topic-based (issue-specific) approach, rather than an ideologically-focused approach will be most successful in possibly reducing partisanship. Similar to OwlFactor and All-sides, NewsPhi is concerned with how opinions and ideologies are formed; however, our approach attempts to allow for more nuance while being less prescriptive with our user base.

B. *Topic Modelling*

A central feature to NewsPhi is its topic modelling. OwlFactor and Allsides similarly create topics out of their news articles, sharing a common desire for consumers to explore news based on groupings other than just bias. OwlFactor relies on machine learning to generate the exact topics that are shown to consumers. One drawback, as expected with such a difficult task as constant topic modelling with ever-changing corpi and topics, is that an automated approach gives, in some cases, nonsensical or vague results. For example: a topic on August 1, 2019 on OwlFactor was "Foreign Minister". Though certainly more useful than the lack of a topic, this is not functionally useful (which foreign minister, what happened, etc.). This said: in general, the topics on OwlFactor change dynamically as new articles are pulled and they are fairly granular and mostly make sense. However, as we considered consistent and sensical granularity for topics crucial for the NewsPhi product, we relied on domain expertise to generate topics instead. This has its advantages and disadvantages later detailed in the evaluation section. Regardless, NewsPhi adds to this topic-allocated news-feed by providing aggregated and specific analysis per topic and article.

AllSides introduces a feature called the Headline Roundup, where it pulls 1-3 common headlines throughout the day and presents a summary of the topic. It further provides reporting trends among the left, the center, and the right for these topics. This kind of topic level analysis is akin to the work that NewsPhi accomplishes. Besides the Headline Roundup topics, there exist other discrete topics that, if clicked upon, filter news articles accordingly. The topics on Allsides indeed reach a level of granularity that matches those of NewsPhi's. However, Allsides does not provide topic-level analysis beyond the few daily Headline Roundup topics and a few other selected topics. NewsPhi strikes a healthy medium between OwlFactor and Allsides, assigning topics by domain experts and automating the aggregated contextual analysis. Further, NewsPhi places each individual article within the broader topic-level context, as opposed to Allsides that places articles within ideological context. Thus, through NewsPhi, one can develop specific understanding of all topics for each specific publication that we provide, as well as understanding the distribution of opinions on this topic across the broader news media environment.

III. System Overview

A. *System Workflow*

The process of creating NewsPhi begins, broadly, with creating a news-feed. The first step is data collection. As NewsPhi is designed to be a consumer facing app, news need to cycle in on a consistent basis. We built a script to pull news from a list of 26 unique sources listed in Table I from various points on the economic-political axes using the webhose.io API [15]. Considering the driving motivation of this paper as exploring the American political landscape, we chose news sites that mostly wrote about politics and current events. The data pulled using webhose.io contains fields such as: headline, byline, publication, publication date/time, URL, and the entire article itself. The data were roughly 43,000 articles pulled from a list of 26 sources over the 30-day period immediately prior to May 1, 2019. After de-duplication and data quality processes, our total corpus amounted to 37,898 articles. These articles ranged from reports to profiles to reviews to opinion pieces (and more).

TABLE I: Sources by the count of articles

Source Name	Count of Articles
Forbes	5241
CNBC	5158
Fox News	3580
Wall Street Journal	3048
Business Insider	2758
NPR	2346
CNN	2275
CNN International	2082
NBC	1978
Daily Wire	1574
Politico	1218
Al Jazeera	1111
National Review	910
The Verge	891
The Daily Beast	829
BBC	684
NewsDay	471
The Economist	352
Salon	269
Bloomberg	267
MSNBC	210
PBS	207
TMZ	200
Foreign Policy	197
ProPublica	41
Huffington Post	1

This data were stored between Amazon Web Service (AWS) Redshift[16] and S3 [17] - the articles themselves were stored on S3 while the metadata were stored on Redshift. Our web server was built on Flask and our user interface (UI) was built using Vue.js. We performed topic modeling using a Latent Dirichlet Allocation algorithm with the headline and the first paragraph (known as the lede) as our inputs. As this is an unsupervised learning problem, there was no ground truth to which we could compare our clustering results. Therefore, evaluation steps were mostly heuristic. After the topic modelling was complete, we performed the controversy score analysis using an aggregation of sentiment per topic. On each article is displayed the news topic, its controversy score, the sentiment score of the specific article currently selected, and a visualization of the spectrum of sentiments expressed in the aggregate of our sources on that topic.

B. *Algorithms*

1) Latent Dirichlet Allocation: The model we employed is a Latent Dirichlet Allocation model (LDA). LDA, developed by Blei, Ng, and Jordan in 2002, can be described as a "generative probabilistic model of a corpus" [18]. LDA can be understood as a probabilistic approach to processing large numbers of documents in such a way that relationships can be drawn in-between them. LDA is, then, suitable for the task of topic modeling.

Within the text modeling context, we can take what is known as a “bag of words” approach - where the order of words in a document is considered to be unimportant - as well as an analogous “bag of documents” approach. Blei et al. refer to this as exchangeability, and their approach is built on this exchangeability assumption that leads to mixture distributions for words/word units. Topics are chosen from a multinomial distribution using probability vectors drawn from a Dirichlet distribution as the parameter. This is slightly intuitive as topics are discrete and the Dirichlet distribution is conjugate to the multinomial (along with some other advantages to using a Dirichlet distribution). That is to say, in Bayesian terms, our prior distribution is the Dirichlet and our posterior distribution is the multinomial.

The first step in understanding LDA is to contextualize the Dirichlet and multinomial distributions within topic modeling. The Dirichlet distribution is the multivariate version of the beta distribution. Dirichlet distributions have one parameter, α, that is a vector comprised of positive real numbers α_i. α can be understood as a constant, α_0, multiplied by what is called a base measure, $\langle \alpha_i' \rangle$. α_0 is the sum of α_i, and $\langle \alpha_i' \rangle$ can be expressed as the means for each topic i, and can thusly be represented as $\frac{\alpha_i}{\alpha_0}$. The formulas for the mean and variance of the Dirichlet distribution are as follows:

$$Mean = \frac{\alpha_i}{\alpha_0} = \alpha_i' \tag{1}$$

$$Variance = \frac{\alpha_i'(1-\alpha_i')}{\alpha_0 + 1} \tag{2}$$

As α_0 gets larger, the variance decreases and the draws from the Dirichlet distribution will be close to α_i'. As α_0 gets smaller, however, the variance increases, resulting in extreme distributions. The multinomial distribution uses the probability vector outputted by the Dirichlet distribution to draw outcomes. If the probability vector is sparse, then the number of possible outcomes becomes small.

Using the above knowledge, one can devise an algorithmic process for writing a corpus of documents. Consider the following hypothetical generative process for a corpus of D documents each of length N, given a user-inputted K as number of topics:

First, we select the topics, θ_d, for each document $d \in D$:

$\theta_d \sim Dir(\alpha)$ where α is sparse (so that each document is only possibly about a small number of topics and we are able to focus on a small number of highly relevant topics and words).

Next, we create the topics themselves, ϕ_k, for $k \in K$: $\phi_k \sim Dir(\beta)$ where β is sparse (so that only relevant words are considered).

Next, we assign the topic, $z_{d,n}$, that the nth word in each document belongs to: $z_{d,n} \sim Multinomial(\theta)$

Finally, we generate the words, $W_{d,n}$, themselves: $W_{d,n} \sim Multinomial(\phi_{z_{d,n}})$.

LDA assumes this simplified corpus generative model, and though it is a naive understanding of how articles are written, this structure proves useful if reverse engineered to provide the topics themselves [19]. The task at hand is then to compute the posterior distribution of our hidden variables. As expected, this is intractable and requires approximation. Our implementation of LDA uses Gibbs sampling [20] for this inference. Gibbs sampling is a specific implementation of MCMC methods [21], which are high volume sampling processes used in order to approximate these intractable posterior distributions. Gibbs sampling approximates these distributions by sampling and then updating probabilities according to which of the possible outcomes were drawn. For example: when assigning words to a topics, we start with equal probabilities of assignment for all words to all topics. After this initial allocation, we pass through and sample each topic for each word again, according to the underlying distribution. It is helpful to have the sparse parameter for the prior distribution so that we have only a few possible words per topic, and analogously a few possible topics per document. Through this sampling process, we eventually reach a state of convergence and the topics become stable.

Our approach to topic modelling used MALLET [20] with a combination of the headline and the lede as our input data. MALLET, short for MAchine Learning for LanguagE Toolkit, is a package that can be used in a variety of Natural Language Processing (NLP) contexts, including topic modelling using LDA with Gibbs Sampling. In order to use MALLET for topic modelling, we passed in a customized list of what is commonly referred to as “stop words” particular to our news data. In text processing, “stop words” are words that one chooses not to consider due to their likely unimportance. Common stop words include generic direct/indirect articles, prepositions, and pronouns such as ”the”, ”where”, etc. A couple of modeling iterations proved useful in determining words to add to our customized list of stop words in order to effectively topic model for the news. We found it useful to add all common words related to time (days of the week, months of the year, etc.). We also found it useful to add the names of all the publications we included in our source list, as webhose.io often included these names in the article data. Additionally, we added bland adjectives and other generic words such as ’good’, ’man’, and ’woman’. These words, though important for other NLP tasks such as predictive text modelling, resulted in relatively useless keywords in our model. A crucial decision was to not add proper nouns (capitalized nouns that refer to specific nouns). Though this made our data at times less clean, especially considering how webhose.io sends data (inclusion of source, author, etc. in some article data), proper nouns were important for achieving the level of granularity desired. For example, in many of our modeling iterations, the word ”Brexit” was a keyword. If not for this word itself, the Brexit topic would likely be both weaker and harder to interpret.

C. Implementation of LDA using MALLET

The main hyperparameters of our model include the number of topics to return (K in the above generative model) and which n-gram to consider as word units. n-grams can be understood as linguistic units; for example, a unigram is a singular word, a bigram is a pair of consecutive words, a trigram is a trio of consecutive words, so on and so forth. An example of a useful bigram could be something like "climate change".

The introduction of n-grams higher than unigrams complicates the previous "bag-of-words" assumption; however, LDA (and MALLET specifically) allows for these larger linguistic units. For NewsPhi, an iterative process resulted in a K of 100, allowing for n-grams up to $n = 2$ (considering both unigrams and bigrams as word units).

As standard with LDA, what is learned from our data is not the topic itself explicitly but rather a list of keywords with corresponding probabilities as possible topics. The following is an example of training data that was given as an input to our model:

> *"Democrats excoriate Attorney General Bill Barr's handling of Mueller report - CNNPolitics (CNN) In the hours after receiving the long-awaited report from special counsel Robert Mueller , some Democrats on Capitol Hill called not to impeach the President but for the removal of Attorney General William Barr."*[22]

This short text contains the headline and the lede for this article. Both parts contain important information about the main actors/subjects. Both parts contain important and distinct proper nouns. The amount of data to include per article was an exercise in balance- too little and topics would likely be too vague, too much and topics would likely be weaker (and training time would increase). Judging off of the contents of just the headline and the lede, effective topic modeling can take place from a heuristic level.

A good possible topic for the above article might be "Mueller Report" or "Bill Barr". After topic modelling took place with our training corpus, we had the following as a unique topic:

$$(62, 0.045 * \text{“}Trump\text{”} + 0.031 * \text{“}report\text{”} + 0.021 * \text{“}Mueller\_report\text{”} + 0.020 * \text{“}Mueller\text{”} + 0.018 * \text{“}Attorney\_General\text{”} + 0.017 * \text{“}special\_counsel\text{”} + 0.016 * \text{“}Robert\_Mueller\text{”} + 0.016 * \text{“}Barr\text{”} + 0.014 * \text{“}William\_Barr\text{”} + 0.013 * \text{“}release\text{”}),$$

Where 62 indicates the cluster number and the words are the keywords off of which one could infer a topic. Each of these keywords have a corresponding probability that that word will be used in an article that is in that cluster. We could reasonably deduce that articles within cluster 62 were about the Mueller report, and within these articles, we should see mentions of Bill Barr, Congress, Trump, an investigation of some sort, etc. The article above did indeed fall into cluster 62, which meant that our topic modelling was successful for at least this example.

The process of assigning topics to clusters remained manual. This process was also a methodological choice. As previously discussed, procuring sensical and specific topics from data purely using machine learning methods is difficult. Manual work is seen as a distinct advantage for NewsPhi. A human domain expert will be able to conceptualize a topic from a group of related words that, put together in any given permutation, may not make a coherent or fully encompassing topic. For example:

$$(61, 0.079 * \text{“}Disney\text{”} + 0.051 * \text{“}Star\_Wars\text{”} + 0.015 * \text{“}film\text{”} + 0.014 * \text{“}series\text{”} + 0.010 * \text{“}streaming\_service\text{”} + 0.010 * \text{“}movie\text{”} + 0.009 * \text{“}trailer\text{”} + 0.008 * \text{“}Iger\text{”} + 0.008 * \text{“}Vogue\text{”} + 0.008 * \text{“}original\text{”})$$

In cluster 61, we have a multitude of articles having to do something with Disney. Heuristically, the topic seems to match; all but one of the terms are Disney and media related. A possible topic is indeed Disney - however, a human expert was able to pull more from this list of keywords with the knowledge that the Disney released news on April 11 regarding the pushed-back launch of Disney+, its new streaming service. In this sense, a newsfeed that tries to produce topics benefits from a human labeler to create more specific, time-relevant topics. The obvious tradeoff is the decrease in automation ability, but this approach provides more precise labels.

D. Controversy Score

The controversy score methodology is rather simple. After the articles are grouped into their respective topics, we call the vaderSentiment Python package [23] to give a sentiment score of all of the articles within each topic. Sentiment scores are the result of sentiment analysis, which is an area of text analysis that is concerned with calculating how positive or negative the sentiment of a text is. The resulting score is usually on a scale of -1 to 1, with -1 indicating an extreme negative sentiment and 1 indicating an extreme positive sentiment. After the articles are grouped, we take the standard deviation of the sentiment score of each topic and then normalize the results. The formula for this is as follows:

$$s = \sqrt{\sum \frac{(X - M)^2}{n - 1}} \tag{3}$$

where X is the sentiment score for each article in the topic, M is the mean sentiment score of the topic, and n is the number of articles in the topic.

The normalization process is a min-max scaling, which takes the resulting list of standard deviations and scales them from 0 to 1. This formula is as follows:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4}$$

where X_{sc} is the sentiment score scaled, X is the sentiment score for each topic, and X_{max} and X_{min} are the maximum and minimum sentiment scores, respectively.

Standard deviation measures how far data points generally fall from the mean. The intuition is that the more controversial

TABLE II: The number of notable topics correctly modeled and the percentage of articles in less useful topics, per K

K	Number of notable topics correctly modeled (max 28)	Percentage of articles in vague/non-newsworthy topics (%)
60	15	9.70
70	16	9.29
80	22	9.10
90	23	8.30
100	24	6.89

a topic is, the higher the standard deviation of sentiment for that topic should be. A higher standard deviation in our case would indicate many articles that have very different opinions about the topic. Thus, we used this standard deviation of sentiment as our controversy score. If we see a corpus of articles with generally uniformly neutral, positive, or negative feelings towards a topic, we expect a lower controversy score. On the other hand, if we see a corpus of articles with very different, highly opinionated feelings towards a topic, we expect a higher controversy score.

IV. Analysis and Evaluation

The main form of evaluation we employed was heuristic on two levels: How specific and useful is the topic modelling so that the topic modeller can make meaningful, coherent topics; and, how sensical were the final controversy rankings?

The former was a process through which we modified the list of stop words, the n-grams to consider, and the K number of topics to model. These parameters and hyperparameters were systematically changed until the list of keywords was sensible enough. Eventually, a stable list of stop words was reached. In addition, we found that considering up to bigrams sufficed, as trigrams tended to introduce too much nonsensical information. Once we settled on the list of stop words and which n-gram to consider, we began iterating through K values to consider. This evaluation process was a combination of qualitative and quantitative methods.

First, we determined a list of notable and/or ongoing events that were easy to glean from our article data. This was a list of 28 topics that included events such as Julian Assange's expulsion from Ethiopian asylum on April 11, the 2020 US presidential elections, the Notre Dame cathedral fire, and the Game of Thrones final season, among other important and distinct cultural and political events. For each K number of topics considered, we tallied how many of these 28 topics could be formed out of the results of the model.

Second, as we labeled the results, we kept track of the clusters that were heuristically too vague or not newsworthy enough to form useful topics. We then calculated the percentage of articles that were in these clusters. Through this iterative process, we hoped to learn which K value would result in the lowest percentage of articles that are assigned such topics.

As this evaluation process requires a fair amount of manual work, we iterated through 5 numbers of topics: 60 through 100 by increments of 10. The results of this iteration can be seen in Table II. We decided to model 100 topics through this evaluation process.

TABLE III: The ten most controversial topics in April 2019

Topic	Controversy Score
Border crisis	0.900
Game of Thrones	0.897
Racial justice in America	0.888
US immigration	0.860
Notre Dame fire	0.859
US-Russia relations	0.857
Politics in the Middle East	0.857
Recent Supreme Court rulings	0.845
Climate change	0.844
US Navy in the Pacific	0.843

TABLE IV: The ten least controversial topics in April 2019

Topic	Controversy Score
Artificial intelligence	0.443
Amazon	0.416
Skincare	0.415
Travel	0.378
Fashion	0.350
Social media marketing and branding	0.348
Credit card points	0.272
Smart homes	0.245
Investment funds	0.201
Consumer reports	0.186

Once modeled, we calculated the controversy scores for each topic and checked that each topic seemed to be in the correct general area. The ten most and least controversial topics in April 2019, according to our methodology, are in Table III and Table IV respectively.

The main surprise from these lists is the Notre Dame fire as the fifth most controversial topic, as there is nothing overtly controversial or political about the fire. However, after reviewing several articles in the cluster, we found that three historically black churches in Louisiana were burned down a couple days prior due to the result of a racist hate crime. After the Notre Dame fire took place, these churches gained a fair amount of publicity due to well organized social media campaigns that raised funding for their rebuilding. The connection between these three churches and the Notre Dame is that the organizers believed that the Notre Dame would easily receive sufficient funding for reconstruction but these churches would not be able to similarly do so [22]. The Notre Dame fire thusly became a launching point to raise awareness for these churches. As a result, a significant number of articles with a wider range of sentiment were written about the Notre Dame fire and the Louisiana churches hence resulting in a higher controversy score.

Overall, the list of topics were generally granular and informative. For the utilized data, we had 8 similar topics that were merged together for a resulting list of 92 distinct topics across the cultural, political, and economic fields. The order of controversy of these topics generally made sense. Almost all of the political topics were in the upper half, while the lower half was comprised of more of the lifestyle, sports, and business topics.

V. Conclusion

Our motivation behind NewsPhi was to create an app that educates people in a non-prescriptive way, following the research of experts in political science and psychology. This foundation we formed in the social sciences was important to us, as conjecturing about the news space could lead to simplified or misinformed assumptions. A prime example of the assumptions we wanted to avoid is the over-emphasis of the effects of fake news. Due to the works of Grinberg et al. [24] and Guess et al.[25], we came to understand that fake news do not pose as serious a problem as they may at first seem. Building on top of this, we used research from political scientists such as Carsey and Layman [8] to determine how party loyalty is formed. We also used research on confirmation bias and extreme news consumption to form a hypothesis on how people react to news that is contrary to what they believe in. Based on this research, we presented an implementation of a topic-organized, less-ideologically prescriptive news-feed that attempts to de-emphasize conventional understandings of political bias and ideology formation while serving an educational purpose and possibly partially bridging the ideological gap. We used a combination of unsupervised learning including Latent Dirichlet Allocation (LDA), a homegrown controversy score formula, and human labeling to create a consumer-facing product with a complete web-server, database, and UI. The results are promising and contribute to the existing contextual newsfeed efforts. To sum, we believe that NewsPhi is just the beginning of more work that should be done in exploring less-prescriptive yet still analytical newsfeeds.

NewsPhi has multiple directions in which it could move further. There may exist other indicators of approval or disapproval of a topic other than just sentiment. Across the topic-level, sentiment certainly is useful. Regardless of whether an article uses strong language in such a manner that either agrees or disagrees with its actual stance, that language likely either describes that article's stance or that of other writers or interviewed/quoted subjects. In aggregate, this sentiment score should be at least a medium indicator of the actual overall opinion on a topic. However, further work should explore additional indicators of opinion. Additionally, though topic assignment is greatly aided by our topic modelling approach, the labeling process is still manual. Allsides appears to prevent this issue by fitting incoming articles into existing categories and focusing on a small number of articles daily to break out into more specific topics (the Headline Roundup feature). OwlFactor appears to use a different NLP approach that at times serves less interpretable topics for incoming articles. Though NewsPhi brings a mixture of the two, further work should focus on automating topic generation while still keeping the same level of granularity accomplished by hand labeling. Lastly, we believe further breakdown of articles by author would be useful. The challenge of tracking authors can be seen in OwlFactor's implementation. Author expertise is one of four elements of a news story's credibility score on OwlFactor; however, oftentimes the author is missing or does not have a strong history. This problem does not exist solely in OwlFactor's implementation, as the problem of author-level contextuality seems broadly complex. Authors may in fact be guest writers, as is the case with many opinion writers. Authors also may share names. If implemented correctly, author reputation could be useful in forming a more holistic view of the news media in general, past the publication-level understanding that currently persists.

References

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[2] A. Guess, B. Nyhan, and J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign," *European Research Council*, vol. 9, 2018.

[3] A. Guess, B. Lyons, J. M. Montgomery, B. Nyhan, and J. Reifler, "Fake news, facebook ads, and misperceptions: Assessing information quality in the 2018 u.s. midterm election campaign," *Public Report*, 2018.

[4] S. E. Gorman and J. M. Gorman, *Denying to the grave: Why we ignore the facts that will save us*. Oxford University Press, 2016.

[5] M. E. Oswald and S. Grosjean, R. Pohl, Ed. Psychology Press, 2004.

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[7] A. Mosseri, "Working to stop misinformation and false news," *Newsroom. fb. com*, 2017.

[8] T. M. Carsey and G. C. Layman, "Changing sides or changing minds? party identification and policy preferences in the american electorate," *American Journal of Political Science*, vol. 50, no. 2, pp. 464–477, 2006.

[9] CivikOwl Inc. (2019) Owlfactor : News on your terms. [Online]. Available: https://www.owlfactor.com/news

[10] Allsides.com. (2019) How allsides rates media bias: Our methods. [Online]. Available: https://www.allsides.com/media-bias/media-bias-rating-methods

[11] Media Bias Fact Check, LLC. Methodology - media bias/fact check. [Online]. Available: https://mediabiasfactcheck.com/methodology/

[12] M. Gentzkow and J. M. Shapiro, "What drives media slant? evidence from u.s. daily newspapers," *Econometrica*, vol. 78, no. 1, pp. 36–37, 2006.

[13] Allsides.com. (2019) Allsides - balanced news via media bias ratings for an unbiased news perspective. https://www.allsides.com/unbiased-balanced-news.

[14] A. Heywood, *Political ideologies: An introduction*. Macmillan International Higher Education, 2017.

[15] webhose.io. (2019) Tap into web content at scale - crawled web data - webhose. [Online]. Available: https://webhose.io/

[16] Amazon Web Services. (2019) Amazon redshift. [Online]. Available: https://aws.amazon.com/redshift/

[17] Amazon Web Service . (2019) Amazon s3. [Online]. Available: https://aws.amazon.com/s3/

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 994–996, 2003.

[19] J. Boyd-Graber, Y. Hu, and D. Minmo, "Applications of topic models," *Foundations and Trends in Information Retrieval*.

[20] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[21] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

[22] A. Rogers. (2019, apr) Democrats aim fury at attorney general bill barr's handling of mueller report. [Online]. Available: https://www.cnn.com/2019/04/19/politics/democrats-barr-fury/index.html

[23] E. Hutto, C.J. Gilbert. (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. [Online]. Available: https://github.com/cjhutto/vaderSentiment#citation-information

[24] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 u.s. presidential election," *Science*, vol. 363, pp. 374–378, 2016.

[25] A. Guess, B. Nyhan, and J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign," *European Research Council*, vol. 9, 2016.