

## RESEARCH ARTICLE

# Data Driven Model to Investigate Political Bias in Mainstream Media

ERIC NESS<sup>1</sup>, AROOJ FATIMA<sup>2</sup>, AND MAHDI MAKTABDAR OGHAI<sup>2</sup><sup>1</sup>Department of Machine Learning, Performeks LLC, Bethesda, MD 20814, USA<sup>2</sup>School of Computing and Information Science, Anglia Ruskin University, CB1 1PT Cambridge, U.K.

Corresponding author: Arooj Fatima (arooj.fatima@aru.ac.uk)

**ABSTRACT** Media bias refers to the tendency of mainstream media outlets to report news in a way that reflects their own political, social, or ideological beliefs or preferences. Such bias may obfuscate facts, manipulate public beliefs, misinform readers, narrow perspectives and viewpoints, and result in greater polarization and division. To counter this issue, this study presents a model for quantifying media bias, aimed at enabling individuals to make more informed media choices. The proposed media analysis model includes a pipeline that gathers articles from three distinct sources: mainstream media news outlets, known conservative outlets, and known liberal media outlets. The collected articles were subjected to a range of text pre-processing operations and subsequently, curated n-gram and topic lists were generated. Several classification models including BERT, logistic regression, random forest, multinomial, and long short-term memory (LSTM) were created and fine-tuned on polarized news sources and used for analyzing news articles from the mainstream media. Among the various classification models that we investigated in this study, BERT achieved overall higher accuracy across the majority of topics. The analysis of mainstream media on various topics yielded different results, with some being balanced and others leaning left or right, depending on the topic. The research also suggests the effectiveness of using highly polarized news sources for developing models to predict media bias.

**INDEX TERMS** Classification, deep learning, mainstream media, media bias, natural language processing, sentiment analysis.

## I. INTRODUCTION

Media bias is defined as a political or ideological inclination of news that supports certain political actors, policies, or topics [1]. To put it simply, bias is “when not telling the whole story is viewed as inaccuracy” [2]. There has been an increase in media bias in recent years [3], [4], which has negative consequences for democracy around the world [5].

In 2012, a poll by Fairleigh Dickinson University examined how various news sources affected people’s understanding of current events. One major takeaway was that people who listen to National Public Radio (NPR) were the most informed group, and the least informed were conservative Fox News viewers [6]. A study conducted by Cassino, et al. [7] that focused on the analysis of poll responses states that watching Fox News is linked with poorer performance on certain poll

questions compared to those who do not watch any news at all. According to the poll, CNN and MSNBC viewers were in the same category, though viewers of these outlets did slightly better than those followed Fox News [6].

The Pew Research Center (a polling institution in the United States) has been tracking how people’s political attitudes have changed over the last 30 years [8]. One of their recent studies indicated that Americans’ political attitudes remain divided by partisanship far more than any other issue such as age, race and ethnicity, gender, educational attainment, religious affiliation, or other socio-economic factors. [8].

A more recent example shows that after five months of the 2020 Presidential elections, 29% of Republicans still believed that not only President Joseph Biden’s victory was illegitimate, but also that the former President should be reinstated [9]. This belief was so strong that almost a year after the election, 20% of Republicans still believed this

The associate editor coordinating the review of this manuscript and approving it for publication was Nazar Zaki<sup>1</sup>.

falsehood [10]. This was, in part, perhaps caused by ambiguous articles reported by FOX5 in Washington DC stating, Trump telling supporters he expects to be reinstated [11]. The falsehood is predicated on the idea that massive voter fraud occurred. In a \$1.9 billion lawsuit between Dominion Voting Machines and Fox News, Fox News gave significant air time to Trump surrogates and lawyers Rudy Giuliani, Sidney Powell, and others stating that widespread voter fraud occurred [12]. The New York Times reported that Fox News's CEO, Suzanne Scott, believed that allegations of voter fraud were false, but proceeded with these types of controversial news stories [13].

Media bias can affect how people come to understand their world. Highly polarized news may hinder one's ability to interpret current events accurately [14], [15], [16].

Various approaches have been used to analyze media bias: polls, manual analysis [17], and text analysis using machine learning techniques. The thrust of this paper is to use a cross-section of Natural Language Processing and Machine Learning techniques to identify and quantify news bias.

This paper aims to address the following two research questions.

- How can we develop a political bias machine learning model that doesn't rely on third-party datasets to derive its findings?
- How can we determine which machine learning approach is best suited for this problem?

Our research intends to investigate if using known polarized news sources outside the mainstream media as the source material for building your model is a great method to determine bias. We will also study if relying solely on one model may or may not be accurate across all topics.

The rest of the paper is organized as follows: Section II reviews relevant literature and research work concerning media bias. Section III covers our methodology and discusses our 'Media Analysis Pipeline'. Section IV provides details of the models' performance results. Section V reviews the results of our models and presents topic-based and news outlet-based analysis. Section VI discusses the limitations of the project and identifies areas for further research. Finally, conclusions are summarised in Section VII.

## II. LITERATURE REVIEW

In the following section, we look at the history of different approaches to investigate bias in media, from the pre-digital age to Natural Language Processing and more sophisticated Machine Learning techniques. Finally, we outline various limitations and gaps in the literature.

### A. INITIAL ANALYSIS OF BIAS

Prior to the digital era, research in media bias was more in the realm of the humanities which was mostly qualitative. We can trace some of its history back to 1920 with Walter Lippmann's seminal book 'Public Opinion' [18], which focused on media critique, propaganda, and its role in democracy. Not until

Edward S. Herman and Noam Chomsky's groundbreaking work, 'Manufacturing Consent', the research in this area shifted from the humanities to a more thorough and rigorous analysis [19]. For instance, Chomsky and Herman did a quantitative analysis of the New York Times's coverage of the Salvadorian Election in March 1984. They analyzed the number of articles and the percentage of text dealing with various aspects of the election.

### B. NATURAL LANGUAGE PROCESSING APPROACHES

The automated identification of media bias and news articles analysis have recently gained attention in the field of computer science [20] particularly using machine learning and Natural Language Processing (NLP) techniques.

Several studies treat detecting media bias through the lens of a regional issue. And even though this group of research is location or language-specific, the project is often started because of some event, be it political instability or an issue like terrorism. Examples of this type of research would include Al-Gamde and Tenbrink's paper on how the Syrian civil war was covered by Iranian news agency [21] or Al-Sarraj and Lubbad's analysis on how the Palestinian/Israeli Conflict is covered by western media [22]. Sometimes, the analysis is also done to cover regional media bias by grouping articles in countries like Poland [23] or an event such as the election in Brazil [24].

A subset of research relies heavily on labeling documents by analyzing word counts of specific pattern sequences, often referred to as  $n$ -grams. An example of this type of research could be illustrated in Al-Gamde and Tenbrink's work in looking at terrorism in Syria. They looked for sequences of words surrounding terrorism and various other patterns [21]. Ultimately, their conclusions are based on how these patterns are used in aggregate and tracked over time.

Another popular area of research is also to analyze the intersection of political news and social media [25], [26], [27], [28]. As an example, some research suggests that certain news articles are being drowned out as some news goes viral [29], [30], [31]. Other research takes a look at the role of social media acting as gatekeepers for particular news articles [32], [33], [34].

### C. MACHINE LEARNING APPROACHES

A more recent attempt relies on sophisticated statistical and artificial intelligence models to analyze and address media bias. A study by Tran [35] investigates how bias works in several outlets by using tools like Bidirectional Encoder Representation Transformer (BERT). Tran used Aspect-based Sentiment Analysis (ABSA) on a pre-trained BERT base model. ABSA classifies sentiment based on entities in the collected text.

Another study worth mentioning by D'Alonzo and Tegmark [36] investigates media bias by using a combination of machine learning and Principal Component Analysis (PCA) paired with Single Value Decomposition (SVD) to

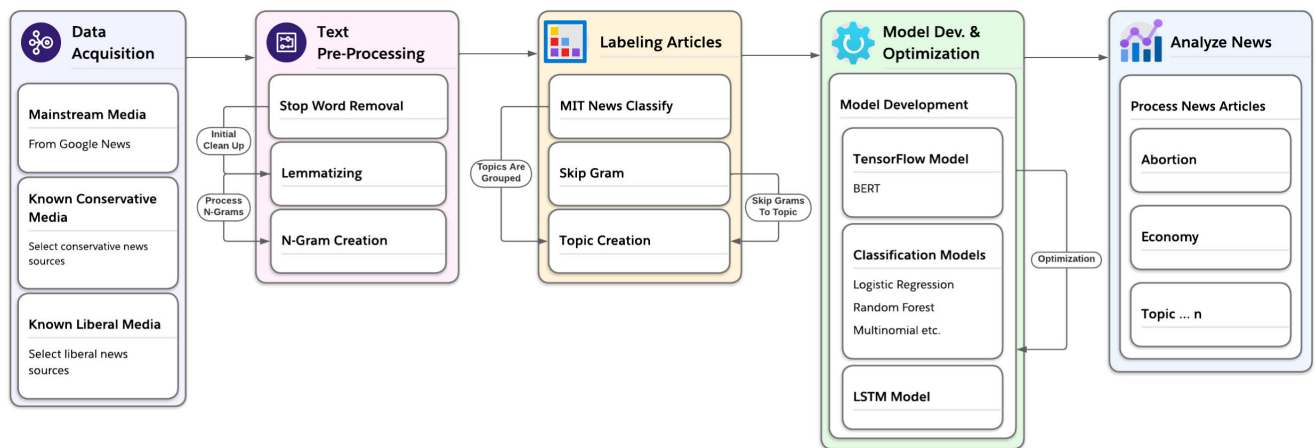


FIGURE 1. Media analysis pipeline.

understand how bias affected different news outlets. Their approach of topic-specific  $n$ -gram lists to create predictive models was the inspiration of our study as well.

#### D. LIMITATIONS AND GAPS

The above mentioned studies have limitations around three distinct areas: news collected for the analysis is limited in terms of size, diversity, and unbiased representation of the real world, in a majority of the cases only a single approach was used in creating models, and finally, relying on third party datasets that were initially devised for purposes other than media bias analysis.

Moreover, the majority of research also relied heavily on outside datasets to do their analysis: Tran [35] used Every-Politician, the Wikipedia API, Political-related Articles Collection datasets, and API's in his analysis. D'Alonzo and Tegmark [36] used the Swiss Policy Research Media Navigator classification to determine how pro-establishment a news outlet is. The data collection regime in these third-party datasets is unclear to us. Relying solely on third-party sources with unpublished strategies for safe guarding media bias may introduce prejudice and drift in data analysis.

Beside the data related issues, the majority of the research do not represent comprehensive analysis of wide range of sentiment analysis and machine learning techniques.

In conclusion, despite significant research in this area, there are some potential gaps to be addressed. Our research aims to address these issues through an established approach to ensure proportional reflection and representation of all view points in the current media climate. It also investigates a large collection of state of the art sentiment analysis and machine learning algorithms to identify the optimal solutions to address media bias.

### III. METHODOLOGY

Figure 1 provides a high-level overview of the proposed data pipeline, which includes five main elements. The first two elements concern article collection and text cleaning.

The third element investigates articles' contents to assign categorical labels and then generates  $n$ -grams. The fourth element runs a multi-step process to build and optimize models. Finally, the fifth element evaluates the models and test their performance in the context of mainstream media.

#### A. DATA ACQUISITION

The articles have been collected via Really Simple Syndication (RSS) feeds from three distinct groups of news sources: mainstream media (MSM), known liberal media (LM), and known conservative media (CM).

Mainstream media is sourced from U.S. political news feed from Google News. Using the Google News feed as the main source, we aim to eliminate any unintentional bias we may have in selecting a news source.

Besides mainstream media outlets, we have included news sources that are widely known as partisan liberal or conservative by the public and literature [37], [38], [39], [40], [41], [42]. The known conservative news outlets are Red State, Hot Air, and The Daily Caller. While known liberal news outlets include Daily Kos, Mother Jones, and Crooks and Liars.

We have built a data collector system that, on average, collected 254 news articles a day. All news articles were collected via RSS feed and stored in a database. The system collected news articles between May 15th and August 19th, 2022. The selection of these dates is arbitrary, with no prejudice or selection criteria. The system collected over 27,000 articles over a period of almost three months. The dataset has been made publicly available through the following link: <https://www.kaggle.com/datasets/newsanalysis/political-bias-in-mainstream-media>

Figure 2 presents the frequency of articles collected on various days. The graph highlights two extreme points i.e. (i) a peak in week 10, when the United States Supreme Court overturned Roe vs. Wade [43] which caused a spike in the news and (ii) a dip in week five which was caused by a technical error, paused data collection for a short period of time. To elaborate, a date parsing error caused articles not

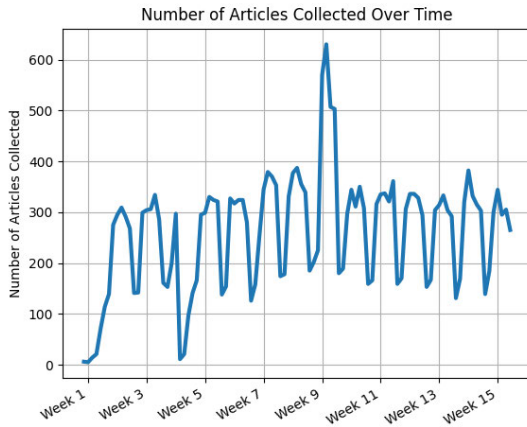


FIGURE 2. Daily article collection.

TABLE 1. Top ten news sources from Google feed.

News Source	Articles
CNN	1,602
The New York Times	813
The Washington Post	661
The Hill	618
The Guardian	404
Business Insider	397
CNBC	362
Bloomberg	326
POLITICO	320
Fox News	263

to be included in the database. However, this glitch has no impact on data sampling, fairness, and frequency.

The article collector system has been built using the Newspaper3k Python library which not only extracts textual data from the RSS feed but also gathers meta-data from the source site i.e. URL identification, images, authors, summary, and some other potentially useful items.

Our article collector system has retrieved over 1000 different news sources from Google News' US political feed [44]. Table 1 lists the top 10 news sources based on the number of articles retrieved per source. A majority of the news comes from less than 10% of all sources.

## B. TEXT PRE-PROCESSING

We use a multi-step process to clean the text. The initial pre-processing includes removing stop words, special characters, erroneous punctuation, e-mails, and quotes. Then the articles go through the process of lemmatizing. The process of lemmatizing text is to take each word down to its root, for example, the root word for “dogs”, “dog’s” and “doggy” will be “dog”. This process removes the duplication of various words that mean the same thing.

The final step for the text pre-processing is to generate  $n$ -grams. A gram in a text-mining context is defined as an individual word.  $n$ -grams by extension are combinations of adjacent grams where  $n$  is the length of the combination. If  $n = 1$ , the gram length in this case would be one word. Where  $n$  is greater than one, the  $n$ -gram is  $n$  subsequent

TABLE 2. A partial list of topics in the system.

Topic Name
ABORTION
DEFENSE AND MILITARY FORCES
ECONOMY
EDUCATION AND SCHOOLS
ELECTIONS
ENVIRONMENT
JAN 6TH
LAW AND LEGISLATION
MEDICINE AND HEALTH
NEWS AND NEWS MEDIA
POLITICS AND GOVERNMENT
SAFETY, HUMAN RIGHTS
SCIENCE AND TECHNOLOGY
UKRAINE

words in the sequence. The  $n$ -grams and their frequencies are extracted and stored in the database. We have used the NLTK library for the text processing.

## C. LABELING ARTICLES

To analyze news data, it was needed to classify articles into categories. We utilize the Massachusetts Institute of Technology (MIT) News Classify library [45] to tag articles with an appropriate topic/category. The model is trained using The New York Times Annotated Corpus [46]. The data includes over 1.4 million cleaned news articles that were manually tagged to a topic out of 538 different topics. MIT News Classify processes the downloaded articles and assigns the appropriate tags. These tags are then used to create categories. We bundle together a group of tags to form a larger coherent topic.

Table 2 displays the list of topics that were created to capture current events happening at the time of article collection. These events include Abortion (regarding a Supreme Court ruling), the War in Ukraine, and the January 6th Committee hearings. Not all topics were used in the final analysis.

To create  $n$ -grams, initially, we use the classical approach by creating monograms, bi-grams, and tri-grams. However, it is very common that as  $n$ -gram increases in size they often contain nonsensical information that is no longer useful. An example of this would be the phrase “seats Georgia January” where the utility in this type of  $n$ -gram would be marginal.

We use Skip-gram model [47] to learn high-quality vector representations of words. It is an efficient process for extracting high-value word relationships which help to determine whether an  $n$ -gram sequence is useful. We calculate the score of phrases using the following formula (see formula 1) proposed by [47]

$$score(w_i, w_j) = \frac{count(w_i, w_j) - \delta}{count(w_i) \times count(w_j)} \quad (1)$$

To create  $n$ -grams using the skip grams model, we employed Gensim's Phrases model [48]. After implementing Gensim's Phrases model, we could see the overall reduction in  $n$ -grams was 69.61%.



**TABLE 3.** *N*-gram counts by topic.

Topic	Conservative	Liberal
Abortion	185	132
Economy	178	121
Election	185	184
Environment	131	159
Jan 6th	156	283
War In Ukraine	362	396

We create curated lists of *n*-grams for each topic. The curation of these lists occurs in the following steps. We look at the general list of *n*-grams that came from known liberal news sources. We sort these *n*-grams by the total number of occurrences, allowing popular *n*-grams to be addressed first. We only select *n*-grams that are related to the topic. We performed this process for all topics. We garner a total of 100 to 300 *n*-grams per topic. The process is repeated for the known conservative news sources (see table 3).

This curated list of *n*-grams is used as the basis for all textual analysis from this point on.

#### D. MODEL DEVELOPMENT AND OPTIMIZATION

We test the performance of various models on the news dataset to obtain improved accuracy. The models include various classification models, BERT (a TensorFlow model) [49] and Long-Short Term Memory networks [50]. The models are built over several iterations, and the testing is done along the way.

##### 1) ABLATION STUDY

The process of model development starts with using various classification models to test whether this is a viable project. The testing process includes the following models: Logistic Regression, Multinomial, Random Forest, SGD Classifier, and Decision Tree.

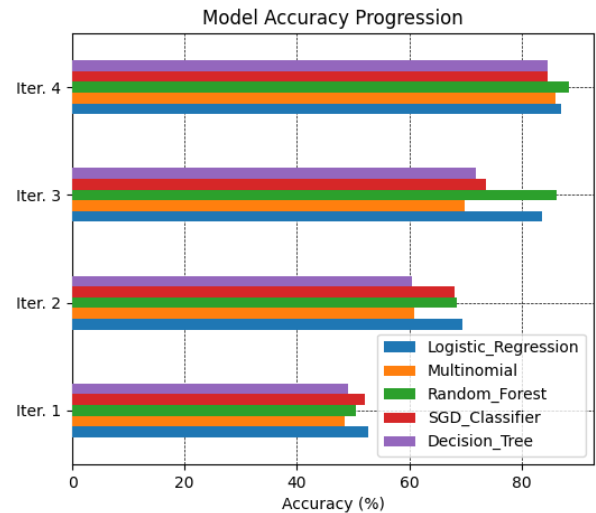
We choose ‘Abortion’ as the candidate topic for this experiment as the United States Supreme Court overturning the law of abortion was the major news event of the time [51]. The ablation study involves four distinct iterations: a baseline with no pre-processing, which involved these articles, a model using an essential list of *n*-grams, another experiment with a more refined list of *n*-grams, and finally, additional model optimization.

##### a: ITERATION 1

We carry out Iteration 1 to establish a baseline of how the models would perform in the absence of text pre-processing and *n*-grams. This test uses 3000 records with a 20/80 split for testing and training. The results are essentially a 50% chance of correctly tagging the text (see Figure 3).

##### b: ITERATION 2

In this iteration, a considerable amount of work goes into selecting all *n*-grams that pertained to ‘abortion’. All news articles tagged with ‘abortion’ or ‘birth control and family planning’ are chosen for *n*-gram analysis. We devise

**FIGURE 3.** Model accuracy over time.**TABLE 4.** Model optimized parameter.

Model	Parameters
LogisticRegression	penalty, C, solver, multi_class
SGDClassifier	penalty, loss, l1_ratio, learning_rate
RandomForestClassifier	criterion, max_features, ccp_alpha
MultinomialNB	alpha
DecisionTreeClassifier	criterion, splitter

a mechanism to cross-correlate *n*-grams that correspond to known liberal media and then known conservative press. The first pass only selects *n*-grams that pertained directly to abortion. We prepared one list of *n*-grams for liberal press and one for conservative media. We run the model again using this list of *n*-grams, achieving a success rate of 69.50% (see Figure 3). We use 968 records in this iteration with a 20/80 split for testing and training.

##### c: ITERATION 3

In iteration 3, we run our application with additional *n*-grams. However, this time, we look at *n*-grams that only appeared in liberal or conservative news sources. Further details on this process are outlined later in this paper. This iteration also includes filtering out articles that only contained one *n*-gram, suggesting that the article was predominantly about another topic. For this iteration, we use 1000 records in a 20/80 split for testing and training.

##### d: ITERATION 4

In this iteration, we fine-tune the parameters to optimize the models’ performance. We outline various parameters we wanted to test in each model (see table 4). For this iteration, we use 1078 records in a 20/80 split for testing and training. This step helps to achieve improved accuracy scoring 88.40% (see Figure 3).

#### 2) TensorFlow MODELS

As the tests from our initial models show a fair amount of promise (see Ablation Study), we feel confident testing

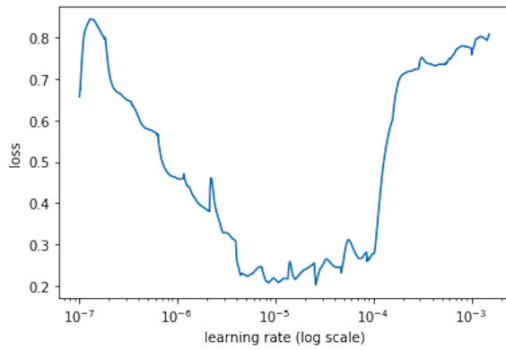


FIGURE 4. BERT learning rate.

TABLE 5. # of Epochs per topic.

Topic	# of Epochs
Abortion	10
Economy	4
Election	2
Environment	6
Jan 6th	9
Ukraine	11

other approaches using more sophisticated models. We run one of the first tests using Google’s BERT (Bidirectional Encoder Representations from Transformers). Initially, the BERT model uses the same articles and format as in Iteration 4 of the above section. The initial run only uses two Epochs without GPUs enabled. The test garners only a 73.12% accuracy.

The BERT model outperforms Logistic Regression, Multinomial, Random Forest, SGD Classifier, and Decision Tree models in this experiment. We detail the results in section TensorFlow Results.

We employ the following process to identify the best-performing model for each topic. The first step is to determine the optimization hyper-parameters. Specifically, we optimize the learning rate using the cyclical learning rate (CLR) and the learning rate range test (LR range test) methods proposed by Smith [52]. Figure 4 shows the values of loss vs. learning rate where we can see the optimal loss value at  $10^{-5}$ . Another hyper-parameter we test is to find the optimal number of epochs we should train the model for. Finding the number of epochs helps the training to make an early stop when validation loss fails to improve [53]. Each topic has been trained for different number of epochs (see Table 5 and Figure 8). During these experiments, we measure and monitor the model loss and accuracy to identify the optimal combination of hyper-parameters. We calculate model loss (see Figure 5) and confusion matrix (Figure 6) to evaluate the overall accuracy of the model. The best-performing model parameters will be saved and used to process and score the remaining mainstream media news articles. We save these scores for further analysis.

### 3) LONG SHORT-TERM MEMORY NETWORK MODELS

To cross-check, we conduct further tests using the same conditions as in the BERT test. However, this time we use a

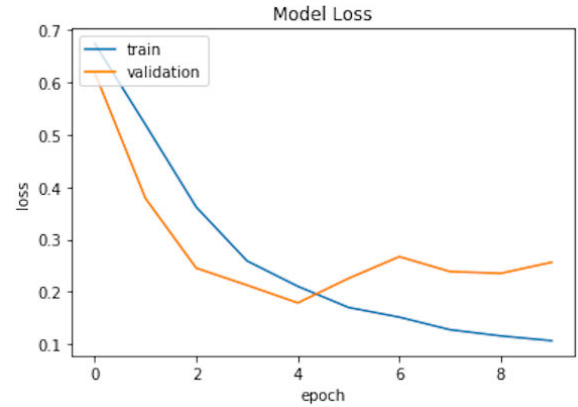


FIGURE 5. BERT model loss.

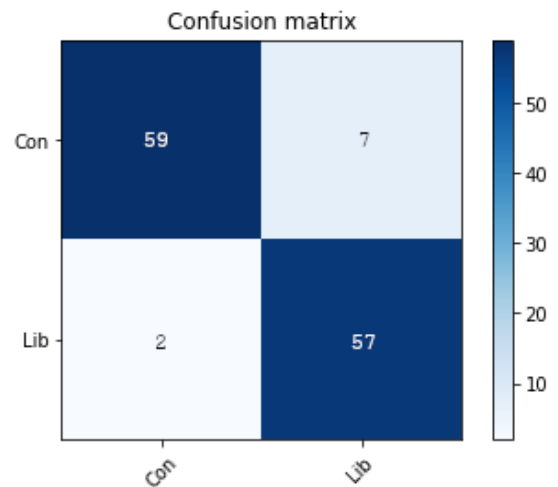


FIGURE 6. BERT confusion matrix.

variant of the Recurrent Neural Network (RNN) model [54] called Long Short-Term Memory (LSTM) network. The history of LSTM starts in 1995, with a published report by Hochreiter and Schmidhuber [55]. One strength of LSTM models over standard RNN approaches is that the latter suffers from a well-known problem called the “vanishing gradient problem”. The issue arises during training when embedding gets longer and the weights get updated, sometimes leading to extremely small values delivering unusable results [56]. The LSTM approach mitigates this issue and is an excellent choice for text classification problems such as this [57]. In the initial run, the LSTM achieves an accuracy of 87.27%. With additional fine-tuning, the model manages to improve the accuracy to 91.32%.

## IV. MODEL EVALUATIONS

In this section, we discuss the performance of all models.

### A. CLASSIFICATION MODEL RESULTS

We use a range of classification models to test the viability of our methodology and set a baseline.

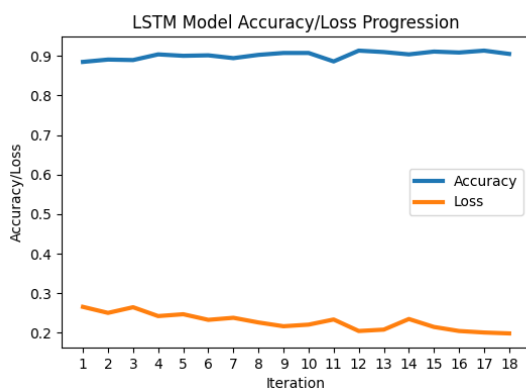
Table 6 presents accuracy scores for our classification models. It can be seen from the results that the Random Forest

**TABLE 6. Classification model accuracy.**

Classification Model	Accuracy
Logistic Regression	87.03%
Multinomial	86.16%
Random Forest	88.48%
SGD Classifier	84.75%
Decision Tree	84.71%

**TABLE 7. LSTM topic accuracy.**

Topic	Accuracy
Abortion	91.32%
Economy	93.40%
Election	92.76%
Environment	94.32%
Jan 6th	92.18%
Ukraine	90.16%

**FIGURE 7. LSTM accuracy/loss rate.**

and Logistic Regression models outperformed the other models with accuracy scores of 88.48% and 87.03% respectively.

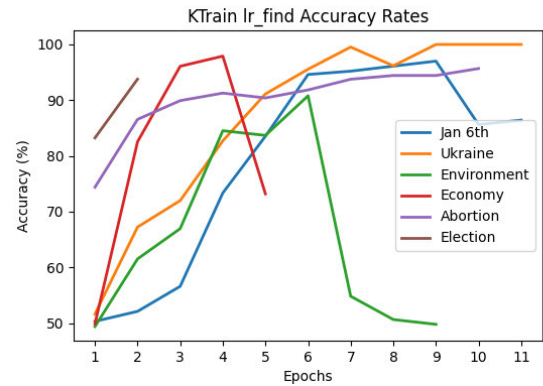
### B. LSTM MODEL RESULTS

Our Long Short-Term Memory Model achieves an initial accuracy of 87.27% while running 2 epochs. Increasing the number of epochs to 18, our model achieves an accuracy of 91.32% for the 'Abortion' topic.

Table 7 lists accuracy scores for all chosen topics after optimizing the model. It is evident from the results that the LSTM model performed well for all the topics achieving an accuracy of more than 90%. Figure 7 shows the accuracy and loss rate for the LSTM model

### C. TensorFlow RESULTS

As we outline in the Methodology (see Section III), we find the optimal learning rate and the number of epochs to achieve optimal accuracy. Figure 8, shows accuracy vs the number of epochs for each different topic. We use the early stopping of training for the topics when the validation loss fails to improve [58], [59]. The average accuracy achieved for all the topics is 95.78%. The lowest accuracy is 90.79% for the topic of 'Environment'. And the highest accuracy achieved is 99.54% for the topic of 'War in Ukraine'.

**FIGURE 8. Dynamic learning rate adjustments per epoch.****TABLE 8. Number of abortion stories.**

# of Stories	Bias
1144	Conservative
751	Liberal
992	Main Stream Media

Further, we calculate learning rates for all topics. Figure 9 provides a visual representation of loss vs learning rate for the topic of Economy. From Figure 9 and 8, we can see that the optimal learning rate for 'Economy' is  $3.1109068 \times 10^{-6}$  for 4 epochs.

Once the model is trained correctly, we save it to be used later in scoring the remaining mainstream news articles. We will detail these results in the following section (Section V).

## V. RESULTS ANALYSIS

In this section, we review our models' results for the mainstream media on various topics. We investigate the performance of the proposed models under two different strategies including topic based analysis and news outlet based analysis.

### A. TOPIC BASED ANALYSIS

Topic based analysis investigates the performance of the models segmented by topic. We explore the following topics: Abortion, January 6th Events, War In Ukraine, Environment, Economy and Elections.

#### 1) ABORTION

One of the major stories that came out at the time of the study was the US Supreme Court decision regarding abortion. The news coverage for the abortion decision was pretty balanced. Table 8 shows the number of stories from three news sources. We have collected 1144 articles from conservative media, 751 articles from liberal media, and 992 news articles from 194 different mainstream media outlets. The top five outlets covering the Abortion decision include CNN, The Washington Post, The New York Times, POLITICO, and The Associated Press.

Overall, this topic turned out 53.66% conservative and 46.34% liberal. Mostly, the model seems to show a balanced coverage. Some notable outliers at the conservative end of

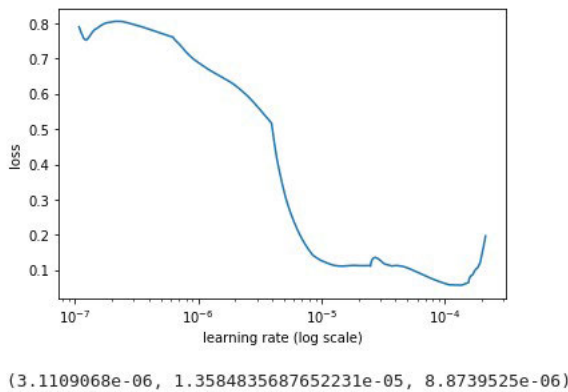


FIGURE 9. Plot and estimate function results for economy.

TABLE 9. Number of Jan 6th stories.

# of Stories	Bias
193	Conservative
186	Liberal
358	Main Stream Media

the spectrum are Fox News (79.67%), the New York Post (80.69%) and Business Insider (80.85%). The highest rated liberal news outlet (with at least ten articles) is CBS News (63.03%).

## 2) JANUARY 6TH EVENTS

Another major news event, for which we have collected articles, is the January 6th Committee investigations. Table 9 shows the number of articles we have collected from three news sources. We have collected 193 articles from known conservative media, 186 from liberal media, and 358 articles from 132 different mainstream news outlets. The top three outlets that covered this topic are CNN, The Washington Post and The New Yorker. On average, coverage for this topic is 50.74% conservative and 49.26% liberal.

This topic seems to be fairly balanced across most major news outlets. Fox News (67.83%) and The Independent (62.40%) are the notable conservative outlets. The Washington Post is on the liberal side with a score of 65.61%.

## 3) WAR IN UKRAINE

One of the biggest global stories this year has been the Russian invasion of Ukraine. We have collected 927 articles from 159 mainstream media outlets. Table 10 shows the number of articles collected from known conservative media, known liberal media, and mainstream media outlets. The top five news outlets from the mainstream media covering this topic are CNN, The Washington Post, The Guardian, The New York Times, and CNBC.

On average, conservative bias can be observed in 68.27% of the reporting, whereas only 31.73% of the articles are on the liberal end. For reporting on the conservative end of the spectrum, the outliers are Fox Business (98.39%) and Fox News (91.30%). Even the coverage of The New York Times ranked 80.66% conservative on this topic. The most liberal

TABLE 10. Number of Ukraine stories.

# of Stories	Bias
842	Conservative
364	Liberal
927	Main Stream Media

TABLE 11. Number of environmental stories.

# of Stories	Bias
227	Conservative
138	Liberal
187	Main Stream Media

TABLE 12. Number of economy stories.

# of Stories	Bias
1548	Conservative
403	Liberal
1424	Main Stream Media

coverage of the war in Ukraine comes from Slate (63.06%), The New Yorker (62.47%), and The Daily Beast (53.77%).

## 4) ENVIRONMENT

We have chosen the topic of the Environment mainly because it is one of the topics where we have relatively balanced data for the articles covered by known conservative and known liberal news outlets. We have collected 227 articles from known conservative media, 138 articles from known liberal media, and 187 articles from 74 different news outlets in the main media (see Table 11). The top five news outlets include CNN, The Guardian, The New York Times, The Associated Press, and The Washington Post.

This is our most polarized model, with coverage being 18.27% conservative and 81.73% liberal. Among liberal news, The Guardian (94.37%), The Associated Press (89.81%), and CNN (87.69%) are the most prominent outlets, respectively. In contrast, the top conservative voice is Fox News, with 54.29%, followed by POLITICO (49.08%) and the New York Times, with (34.25%).

## 5) ECONOMY

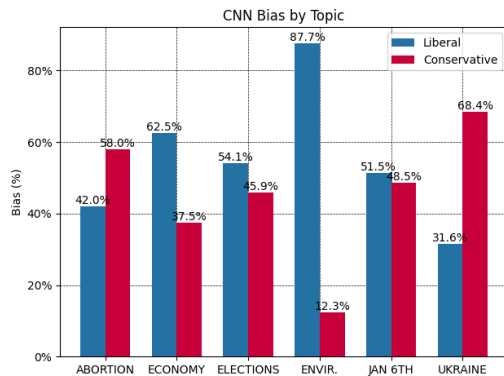
We have chosen the topic of the Economy as it was one of the major topics comprising several sub-topics. We have collected a total of 1548 news articles from known conservative media, 403 articles from known liberal media, and 1,424 articles from 217 different mainstream media outlets (see table 12). Top news sources include CNBC, CNN, The New York Times, The Guardian, and The Washington Post.

The topic for the Economy overall leans slightly towards liberal with 60.70% compared to 39.30% conservative. CBS News (81.25%), The Associated Press - en Español (80.64%), and POLITICO (79.86%) are among the most common liberal outlets. On the conservative end of the spectrum, we have Fox Business (69.88%), Financial Times (61.34%) and Fox News (55.51%)



**TABLE 13.** Number of election stories.

# of Stories	Bias
1675	Conservative
970	Liberal
2800	Main Stream Media

**FIGURE 10.** CNN bias by topic.

## 6) ELECTION

The topic of Election news has been selected because it contains a large number of articles. We have collected 1675 news articles from known conservative media, 970 articles from known liberal media, and 2,800 news articles from 329 different outlets (see table 13). The top news sources are CNN, The New York Times and The Washington Post.

The topic leans slightly towards the liberal end, with 56.84% compared to 43.16% conservative. The most conservative outlets for the Election's topic are Reuters, with a 67.66%, followed by CBS News with 63.59%. On the liberal side of the spectrum, we have The Guardian at 85.16%, followed by Raw Story and The New Yorker with just over 77%.

## B. NEWS OUTLET BASED ANALYSIS

Arguably, the four most influential news outlets in the United States are CNN, The New York Times, The Washington Post, and Fox News [60]. Figures (10, 11, 12, 13) present the cross section of how these news outlets have rated our models across all selected topics.

CNN's coverage across the topics is mostly balanced, with two notable exceptions: highly liberal for the environment and relatively conservative with regard to the war in Ukraine (see Figure. 10).

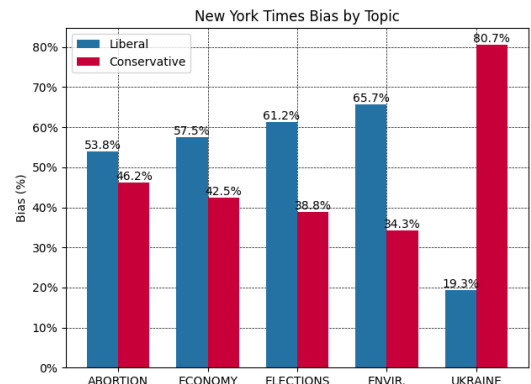
The New York Times is consistently liberal on all topics, except for its coverage of the war in Ukraine (see Figure 11).

The Washington Post shows a more moderate approach to covering topics. It has a conservative slant when it comes to abortion and the war in Ukraine. However, on all other topics, its coverage is decidedly liberal (see Figure 12).

Fox News tends to be conservative on all topics, especially when discussing Ukraine, Abortion, and the Jan 6th Committee (see Figure 13).

## C. RESULTS SUMMARY

Table 14 outlines our chosen topics and presents their corresponding average score for the conservative and liberal sides.

**FIGURE 11.** The New York Times bias by topic.

We present a detailed breakdown of results, for each topic, in Table 15. It shows the number of collected articles from each news outlet by topic along with the corresponding media bias results using the following models including BERT, Logistic Regression, Random Forest, Multinomial, and Long Short-Term Memory (LSTM). Bold values represent statistically significant results.

## VI. DISCUSSION

This section offers a qualitative in-depth explanation and insight into interesting items of note and areas which can be explored further.

We start by reviewing data issues, specifically data and topic limitations, and our solution to ensure that we balance our models on a per-topic basis. We then discuss various observations regarding the content of each topic. Finally, we cover some detailed spot-checking of results that possibly bucked convention to ensure our models produced results in line with the data collected.

## A. DATA AND TOPIC LIMITATIONS

The initial collection of articles from this study included a considerable number of topics [493 individual topics and 18 meta-topics]. However, our experiments are limited to topics with a high number of articles only. We considered topics with a minimum of 180 articles. The articles were collected over a period of 22 weeks during the summer of 2022. Ideally, a project like this would be conducted over several years, covering a wider array of topics and including millions of news articles.

## B. IMBALANCED NUMBER OF ARTICLES

One challenge we faced was balancing the number of articles for both liberal and conservative ends. Conservative news outlets chosen for this project outproduced content by a wide margin (see Table 16).

To ensure that conservative news outlets are not overrepresented, we handle an imbalance in articles by imposing a limit on the number of conservative news articles. We randomly

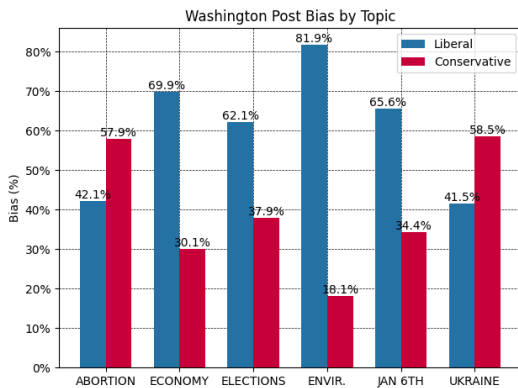


FIGURE 12. The Washington Post bias by topic.

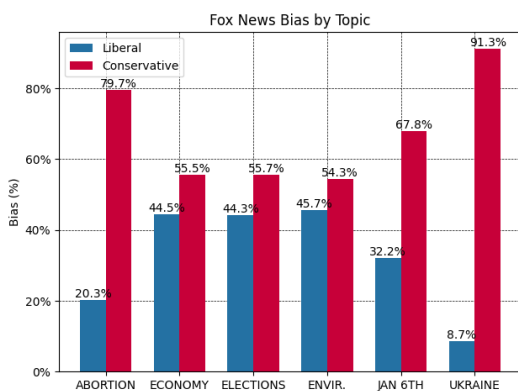


FIGURE 13. Fox News bias by topic.

select a subset of conservative articles to equalize the number of articles in both conservative and liberal media.

This subsampling helps to equalize the number of *n*-grams across topics in both conservative and liberal media. For example, in the initial run of selecting the *n*-grams for the ‘Economy’ topic, conservative *n*-grams outnumbered by 3 to 1 against the liberal side. However, the sub-sampling process resolved this issue.

The selection of sources should be taken into account as it has the potential to introduce bias in the process and, eventually, the results.

### C. TOPIC BASED ANALYSIS

This section provides an analysis for each chosen topic with reference to the *n*-grams.

While selecting *n*-grams for any topic, we choose *n*-grams pertaining to the topic for both known conservative and liberal news outlets. In the first round of selection, we choose *n*-grams that appear in both sets of articles. Then we run a second pass and choose *n*-grams that only appear in conservative or liberal articles. During this process of selecting distinct *n*-grams, we observed some interesting patterns. Most of the *n*-grams in the liberal media have been shown to be correlated with the topic. However, when considering conservative media outlets, there appeared to be varying patterns.

TABLE 14. Bias by topic.

Topic	Con (%)	Lib (%)
ABORTION	53.66%	46.34%
ECONOMY	39.30%	60.70%
ELECTIONS	43.16%	56.84%
ENVIRONMENT	18.27%	81.73%
JAN 6TH	50.74%	49.26%
UKRAINE	68.27%	31.73%

#### 1) ABORTION

When analyzing the topic of ‘Abortion’ from known conservative media outlets, we observed the conflation of the topic of abortion with other topics i.e., interracial marriage, lifers, affirmative action, lefties, federal lands, sexuality, due process, prohibition, far left, pro-lifers, demand abortion, birthing people, pregnant women, pregnancy centers, and stare decisis.

An observation showed that certain *n*-grams in conservative media regarding the abortion topic happened to be individuals’ names, including Nancy Pelosi, Elizabeth Warren, Ilhan Omar, Hillary Clinton, Californians, Gavin Newsom, Chuck Schumer, and Hunter Biden. None of these *n*-grams appeared in liberal media outlets.

We also observed some vocabularies relevant to religion (including *n*-grams such as Pope Francis, Catholics, Catholic Church, Eucharist, Holy Communion, Vatican, Archdiocese, Priests, Cardinal, and sin) and violence (including *n*-grams such as vandalism, arson, firebombing, handcuffed, radicals, homelessness, arsonists, Molotov, graffiti, and ANTIFA).

However, when looking at the known liberal media, only a few *n*-grams strayed off the topic. Notable exceptions would be *n*-grams like gender-affirming, Proud Boys, gerrymandered, Texas Republicans, and telemedicine.

#### 2) JAN 6TH NEWS

Similarly to the abortion topic, we observed a conflation for the ‘Jan. 6th’ topic. We noted that the known conservative outlets often bring up topics outside the scope of the 6th January events. Specifically, it was not uncommon for the following to be mentioned in their text: abortion, Hillary Clinton, Nancy Pelosi, Jamie Raskin, Alexandria Ocasio-Cortez, and Sandy Hook.

For known liberal outlets, they also mentioned other topics in their coverage, including abortion, redistricting, civil war, Patel, far right, dark maga, etc. Also, this is one topic where we observed slightly more expressive words in liberal coverage: wrong, wild, threatened, lies, jeers, etc.

#### 3) WAR IN UKRAINE

The coverage for the topic of ‘War in Ukraine’ contained *n*-grams which seemed to be as expected for both conservative and liberal outlets. During coverage of this topic, conservative media produced articles surrounding gas, oil, and energy. The liberal outlets employed slightly emotional terms for this topic in their coverage, e.g., dangerous, threatening, right-wing, and Mar-a-Lago. However, the model results were predominantly conservative.

**TABLE 15.** Detailed results for the topic and news outlet for each model.

News Outlet	Count	BERT Lib. - Con.	Log Reg Lib. - Con.	Rand Forst Lib. - Con.	Multi ND Lib. - Con.	LSTM Lib. - Con.
<b>Topic: Elections</b>						
CNN	334	54.05% - 45.95%	55.07% - 44.93%	65.31% - 34.69%	59.88% - 40.12%	48.35% - 51.65%
The New York Times	255	61.21% - 38.79%	53.62% - 46.38%	64.72% - 35.28%	57.55% - 42.45%	52.46% - 47.54%
The Washington Post	240	62.12% - 37.88%	59.74% - 40.26%	69.70% - 30.30%	63.10% - 36.90%	51.32% - 48.68%
Business Insider	142	52.66% - 47.34%	59.61% - 40.39%	69.20% - 30.80%	62.62% - 37.38%	46.42% - 53.58%
POLITICO	126	52.72% - 47.28%	54.09% - 45.91%	66.87% - 33.13%	55.89% - 44.11%	59.81% - 40.19%
<b>Toic: Economy</b>						
CNN	169	51.70% - 48.30%	50.31% - 49.69%	76.10% - 23.90%	51.15% - 48.85%	64.43% - 35.57%
CNN	147	62.51% - 37.49%	58.90% - 41.10%	80.73% - 19.27%	66.14% - 33.86%	66.99% - 33.01%
The New York Times	89	57.54% - 42.46%	58.35% - 41.65%	83.20% - 16.80%	66.38% - 33.62%	61.33% - 38.67%
The Guardian	69	70.95% - 29.05%	62.92% - 37.08%	85.42% - 14.58%	71.28% - 28.72%	77.29% - 22.71%
The Washington Post	63	69.91% - 30.09%	57.50% - 42.50%	87.05% - 12.95%	63.38% - 36.62%	80.09% - 19.91%
<b>Topic: Environment</b>						
CNN	28	<b>87.69% - 12.31%</b>	51.97% - 48.03%	61.39% - 38.61%	53.38% - 46.62%	53.30% - 46.70%
The Guardian	19	<b>94.37% - 5.63%</b>	56.70% - 43.30%	63.58% - 36.42%	62.73% - 37.27%	50.25% - 49.75%
The New York Times	14	65.75% - 34.25%	48.61% - 51.39%	60.86% - 39.14%	50.23% - 49.77%	54.33% - 45.67%
The Associated Press	10	<b>89.81% - 10.19%</b>	46.58% - 53.42%	60.30% - 39.70%	45.70% - 54.30%	50.66% - 49.34%
The Washington Post	9	<b>81.87% - 18.13%</b>	50.60% - 49.40%	74.78% - 25.22%	43.62% - 56.38%	49.14% - 50.86%
<b>Topic: Abortion</b>						
CNN	138	<b>41.97% - 58.03%</b>	58.23% - 41.77%	<b>71.20% - 28.80%</b>	63.74% - 36.26%	<b>40.68% - 59.32%</b>
The Washington Post	89	<b>42.15% - 57.85%</b>	60.80% - 39.20%	<b>74.84% - 25.16%</b>	64.80% - 35.20%	<b>34.30% - 65.70%</b>
The New York Times	77	53.85% - 46.15%	56.74% - 43.26%	64.26% - 35.74%	61.10% - 38.90%	<b>36.26% - 63.74%</b>
POLITICO	53	<b>27.05% - 72.95%</b>	56.77% - 43.23%	<b>73.45% - 26.55%</b>	58.46% - 41.54%	<b>39.69% - 60.31%</b>
The Associated Press	43	<b>38.61% - 61.39%</b>	62.90% - 37.10%	<b>72.67% - 27.33%</b>	63.38% - 36.62%	<b>35.94% - 64.06%</b>
<b>Topic: Jan 6th Events</b>						
CNN	39	51.45% - 48.55%	53.93% - 46.07%	56.59% - 43.41%	49.09% - 50.91%	44.97% - 55.03%
The Washington Post	27	65.62% - 34.38%	62.52% - 37.48%	43.11% - 56.89%	61.96% - 38.04%	50.86% - 49.14%
The New Yorker	25	41.93% - 58.07%	48.63% - 51.37%	42.48% - 57.52%	44.34% - 55.66%	47.28% - 52.72%
The Bulwark	10	38.48% - 61.52%	54.87% - 45.13%	50.50% - 49.50%	49.68% - 50.32%	50.26% - 49.74%
Business Insider	9	47.31% - 52.69%	58.35% - 41.65%	47.00% - 53.00%	52.81% - 47.19%	44.13% - 55.87%
<b>Topic: War In Ukraine</b>						
CNN	115	31.63% - 68.37%	43.42% - 56.58%	41.20% - 58.80%	33.37% - 66.63%	29.18% - 70.82%
The Washington Post	73	41.47% - 58.53%	49.13% - 50.87%	46.74% - 53.26%	38.25% - 61.75%	24.92% - 75.08%
The Guardian	51	49.06% - 50.94%	46.56% - 53.44%	43.04% - 56.96%	35.03% - 64.97%	20.91% - 79.09%
CNN	49	28.09% - 71.91%	42.18% - 57.82%	42.00% - 58.00%	31.69% - 68.31%	26.13% - 73.87%
The New York Times	49	19.34% - 80.66%	46.33% - 53.67%	42.96% - 57.04%	37.88% - 62.12%	36.01% - 63.99%

**TABLE 16.** Article counts by political spectrum.

Known Bias	Count
Conservative	11,665
Liberal	4,542
Main Stream Media	12,117

#### 4) ECONOMY

Regarding the topic of 'Economy', some interesting terms and patterns appeared in the selection of  $n$ -grams. Unlike other topics, conservative news outlets employed a much higher degree of technical jargon regarding the economy, including terms such as commodities, demand, indicators, inventory, projections, and reserves. Liberal outlets discussed more generic terms such as workers, families, farmers, and spending.

Another interesting factor was that using a more technical vocabulary did not cause the models to lean toward the conservative end. This may be because mainstream news outlets mostly write to the layperson and use technical jargon sparingly.

#### 5) ENVIRONMENT

The Environment topic was quite balanced in terms of the number of  $n$ -grams. The  $n$ -grams extracted for both liberal

and conservative ends were, by in large, indistinguishable from each other. However, on the whole, the topic leaned heavily liberal in this scenario, making it among our most polarizing issues.

Given that the vocabulary for the  $n$ -grams is essentially the same, a question arises: How come this topic leans liberal? Our current running hypothesis is that even though the language is very similar, the frequency with which they use words differs. For instance, conservatives used the word 'Biden' 416 over all the articles, whereas liberal outlets used the word 'Biden' only 118 times. Or in another case, liberals used the word 'fuel' 195 times, and conservatives only used the word 105 times. We see this pattern play out when we look at the War in Ukraine topic as well.

#### 6) ELECTION

The Election topic was mainly unremarkable. It contained many articles and was reasonably well balanced with respect to  $n$ -gram counts. The only item of note was the TensorFlow model's relatively short epoch run when finding learning rate adjustment. As illustrated in Figure 8, the model was run for only two iterations. It was stopped early because it was making no further improvements. Overall, it ended up rating

the mainstream media as relatively balanced, in line with other topics, so we defer to the model here.

#### D. MODEL BASED ANALYSIS

This section offers a comparative analysis of the models' investigating this study. It specifically looks at results where certain outcomes seem to be in contrast with the rest of the models.

In Table 15, we have highlighted several findings that appeared to be outliers.

For our BERT model, the two topics that returned unusual results compared to the other models were Environment and Abortion. This model returned extremely liberal results when it comes to the topic of Environment. However, it leaned conservative on the topic of Abortion. We believe part of the result is due to the fact that we are using a pre-trained BERT model that may introduce some level of unwanted prejudice.

Our Random Forest model has returned interesting results on the topics of the Economy and Abortion. Most models' results were balanced or slightly liberal. In contrast, the Random Forest Model leaned liberal for both topics. We believe the model might return consistent results with additional fine-tuning of model parameters.

Our LSTM model performed reasonably well in most of the topics. The main issue came under Abortion. Here, the results mirrored the BERT model but were at odds with the classification models.

The two models that produced results that were always consistent with the majority consensus were the Logistic Regression and the Multinomial models.

Regarding the Abortion topic, we manually read several dozen articles to try to assess the discrepancies between some of our models. It is our view that we could read the articles as leaning toward the conservative side. Meaning that in our opinion, the BERT and LSTM assessments are more likely to be correct in this scenario. What is interesting about the Abortion topic is that there are many articles where abortion is not the primary issue. They often mention it in the context of an election and candidates' preferences or in pieces like "Top 10 Things You Need To Know", where it is one of many items being discussed.

A takeaway from these results is that it is essential not to rely on a single model to draw conclusions. A more robust method would be to create a multiple-model consensus approach, as determining political bias is much more complicated than a simple sentiment classification. This proposed technique would be an exciting area for future work.

#### E. PBS NewsHour SPOT CHECK

To check whether our model was working correctly, it was essential to check items that potentially opposed popular conventions. A specific example where this occurred was with PBS NewsHour on the topic of 'Abortion'.

Specifically, PBS NewsHour was rated 71.90% conservative. In the United States, however, PBS NewsHour tends to be publicly viewed as Center-Left. An argument could be

made that because it is publicly funded, a great deal of effort is made to remove any bias, and thus should be considered simply as center.

However, after the rating, a deeper dive was made to examine the nine articles that made up the score. Here are the results regarding the deeper dive.

- 1) 3 stories were rated liberal.
- 2) 2 stories rated as conservative were actually written by Associated Press:
  - a) Story about how 1 million voters are turning Republican
  - b) Biden G7 story stating he appears weak
- 3) Three stories rated conservative contained minimal text (114-word avg including stop words, etc.).
- 4) One story was rated conservative but, in fact, was predominantly liberal as the piece was mainly about half of Americans think former President Donald Trump should face criminal charges; however, only a brief section was about abortion with a conservative slant.

#### VII. CONCLUSION

Media bias is defined as a political or ideological inclination of news that supports certain political actors, policies, or topics. It is important to tackle media bias due to its crucial impact on public perception. This research aims to investigate bias in mainstream media, through a combination of Machine Learning and Natural Language Processing techniques.

This paper lays out a process which channels news articles from the mainstream media and predicts how much political bias it has by comparing it to known quasi-news sources that hold polarized viewpoints. We counter the issue by utilizing a variety of different Natural Language Processing and Machine Learning approaches. We outline the model results as we build and test the various models using our highly polarized news outlets as our training data.

In this research, we considered over 27,000 articles across 18 different topics over the period of 14 weeks. However, to unfold subtle media bias narratives, a project like this would be ideally conducted over several years of data collection covering a wider array of topics and including millions of news articles. This study opens up new avenues of research on tackling media bias and to improve credibility of digital media using Machine Learning and Natural Language Processing methods.

The theoretical contributions made by this paper are as follows: first, we collected known fringe polarised news sources as source material to create our models to determine bias in mainstream media - whereas other approaches in the literature relied on outside datasets to assess bias. Relying solely on outside datasets with unknown data collection strategies may include unwanted bias and prejudice.

Another crucial finding is that relying on a single model can generate unreliable results in certain topics. To counter that, we include multiple models in a consensus mechanism to determine bias, whereas almost all other studies chose



one statistical or machine learning model. This approach improves the robustness of the proposed model and addresses the second research question of this study. In conclusion, detection of bias in mainstream media is a delicate challenge and requires further research in various avenues including data collection and machine learning techniques.

## REFERENCES

- [1] D. Shultziner and Y. Stukalin, "Distorting the news? The mechanisms of partisan media bias and its effects on news production," *Political Behav.*, vol. 43, no. 1, pp. 201–222, Mar. 2021.
- [2] R. M. Eisinger, L. R. Veenstra, and J. P. Koehn, "What media bias? Conservative and liberal labeling in major U.S. newspapers," *Harvard Int. J. Press/Politics*, vol. 12, no. 1, pp. 17–36, Jan. 2007.
- [3] S. D'Alonzo and M. Tegmark, "Machine-learning media bias," *PLoS ONE*, vol. 17, no. 8, Aug. 2022, Art. no. e0271947.
- [4] A. E. Wilson, V. A. Parker, and M. Feinberg, "Polarization in the contemporary political and media landscape," *Current Opinion Behav. Sci.*, vol. 34, pp. 223–228, Aug. 2020.
- [5] J. McCoy, T. Rahman, and M. Somer, "Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities," *Amer. Behav. Scientist*, vol. 62, no. 1, pp. 16–42, Jan. 2018.
- [6] M. Kelley. (May 2012). *STUDY: Watching Only Fox News Makes You Less Informed Than Watching No News At All*. [Online]. Available: <https://www.businessinsider.com/study-watching-fox-news-makes-you-less-informed-than-watching-no-news-at-all-2012-5>
- [7] D. Cassino, "What you know depends on what you watch: Current events knowledge across popular news sources," Public Mind, FDU College Florham, Madison, NJ, USA, Tech. Rep. 03MAY2012, Tech. Rep., 2012. [Online]. Available: <http://publicmind.fdu.edu/2012/confirmed/>
- [8] Pew Research Center. (Dec. 2019). *In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions*. [Online]. Available: <https://www.pewresearch.org/politics/2019/12/17/in-a-politically-polarized-era-sharp-divides-in-both-partisan-coalitions/>
- [9] D. Mastrangelo. (Jun. 2021). *3 in 10 Republicans Believe Trump Will be Reinstated as President: Poll*. The Hill. [Online]. Available: <https://thehill.com/homenews/news/557486-one-third-of-republicans-believe-trump-will-be-reinstated-as-president-poll/>
- [10] A. Stanton. (Dec. 2021). *Nearly 1 in 5 Republicans say Trump Will 'Likely' be Reinstated in 2021*. Newsweek. [Online]. Available: <https://www.newsweek.com/nearly-1-5-republicans-still-say-its-likely-trump-will-reinstated-years-end-1660890>
- [11] (Jun. 2021). *FOX 35 News Staff Report: Trump Telling Supporters he Expects to be Reinstated*. [Online]. Available: <https://www.fox5dc.com/news/report-trump-telling-supporters-he-expects-to-be-reinstated>
- [12] D. Mastrangelo. (Oct. 2022). *Dominion CEO: Fox News 'Knew the Truth' About Voter Fraud Claims*. The Hill. [Online]. Available: <https://thehill.com/homenews/media/3701130-dominion-ceo-fox-news-knew-the-truth-about-voter-fraud-claims/>
- [13] T. Porter. (Oct. 2022). *Fox News CEO had Strong Doubts About Trump's Election-Fraud Claims, NYT Report Says. The Network Pushed Them Anyway*. [Online]. Available: <https://www.businessinsider.com/fox-news-ceo-privately-doubted-trump-election-fraud-claims-nyt-2022-10>
- [14] D. Bernhardt, S. Krasa, and M. Polborn, "Political polarization and the electoral effects of media bias," *J. Public Econ.*, vol. 92, nos. 5–6, pp. 1092–1104, Jun. 2008.
- [15] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: An interdisciplinary literature review," *Int. J. Digit. Libraries*, vol. 20, no. 4, pp. 391–415, Dec. 2019.
- [16] T. Spinde, "An interdisciplinary approach for the automated detection and visualization of media bias in news articles," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2021, pp. 1096–1103.
- [17] A. Panda, D. Siddharth, and J. Pal, "COVID, BLM, and the polarization of U.S. politicians on Twitter," 2020, *arXiv:2008.03263*.
- [18] W. Lippmann, *Greenbook Publications*. Sioux Falls, SD, USA: Greenbook Publications, 2010.
- [19] E. Herman and N. Chomsky, *Manufacturing Consent: The Political Economy of the Mass Media*. New York, NY, USA: Pantheon Books, 1994.
- [20] F. Hamborg, A. Zhukova, and B. Gipp, "Automated identification of media bias by word choice and labeling in news articles," in *Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL)*, Jun. 2019, pp. 196–205.
- [21] A. Al-Gamde and T. Tenbrink, "Media bias: A corpus-based linguistic analysis of online Iranian coverage of the Syrian revolution," *Open Linguistics*, vol. 6, no. 1, pp. 584–600, Nov. 2020.
- [22] W. F. Al-Sarraj and H. M. Lubbad, "Bias detection of Palestinian/Israeli conflict in western media: A sentiment analysis experimental study," in *Proc. Int. Conf. Promising Electron. Technol. (ICPET)*, Oct. 2018, pp. 98–103.
- [23] K. Baraniak and M. Sydow, "News articles similarity for automatic media bias detection in Polish news portals," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, Sep. 2018, pp. 21–24.
- [24] A. Sales, L. Balby, and A. Veloso, "Media bias characterization in Brazilian presidential elections," in *Proc. 30th ACM Conf. Hypertext Social Media*, Sep. 2019, pp. 231–240.
- [25] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, and M. Flammini, "Detecting and tracking political abuse in social media," in *Proc. 5th Int. Conf. Weblogs Social Media (ICWSM)*. Washington, DC, USA: Association for the Advancement of Artificial Intelligence, 2011.
- [26] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, and M. Flammini, "Political polarization on Twitter," in *Proc. 5th Int. Conf. Weblogs Social Media (ICWSM)*. Washington, DC, USA: Association for the Advancement of Artificial Intelligence, 2011.
- [27] D. Gayo-Avello, "'I wanted to predict elections with Twitter and all i got was this lousy paper'—A balanced survey on election prediction using Twitter data," 2012, *arXiv:1204.6441*.
- [28] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welp, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. Int. AAI Conf. Web Social Media*, vol. 4, no. 1, pp. 178–185, May 2010.
- [29] E. Hargittai, "Potential biases in big data: Omitted voices on social media," *Social Sci. Comput. Rev.*, vol. 38, no. 1, pp. 10–24, Feb. 2020.
- [30] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, Jun. 2015.
- [31] B. Suh, L. Hong, P. Pirollo, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network," in *Proc. IEEE 2nd Int. Conf. Social Comput.*, Aug. 2010, pp. 123–276.
- [32] D. Saez-Trumper, C. Castillo, and M. Lalmas, "Social media news communities: Gatekeeping, coverage, and statement bias," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag. (CIKM)*, 2013, pp. 1679–1684.
- [33] I. Wasserman and M. Richmond-Abbott, "Gender and the internet: Causes of variation in access, level, and scope of use," *Social Sci. Quart.*, vol. 86, no. 1, pp. 252–270, 2005.
- [34] E. Pöyry, S. Laaksonen, A. Kekkonen, and J. Pääkkönen, "Anatomy of viral social media events," in *Proc. HICSS*, 2018, pp. 1–17.
- [35] M. Tran, "How biased are American media outlets? A framework for presentation bias regression," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 4359–4364.
- [36] S. D'Alonzo and M. Tegmark, "Machine-learning media bias," 2021, *arXiv:2109.00024*.
- [37] H. Air. (2022). *About*. [Online]. Available: <https://hotair.com/about>
- [38] (2022). *Crooks and Liars Liars About Us*. [Online]. Available: <https://crooksandliars.com/about>
- [39] (2022). *Daily Kos About Us*. [Online]. Available: <https://www.dailykos.com/about-us>
- [40] Mother Jones. (2022). *Influence Watch*. [Online]. Available: <https://www.influencewatch.org/non-profit/mother-jones-foundation-for-national-progress/>
- [41] (2022). *Red State Contact*. [Online]. Available: <https://redstate.com/contact>
- [42] (2022). *Townhall Conservative News, Politics, Opinion, Breaking News Analysis, Political Cartoons and Commentary Townhall*. [Online]. Available: <https://townhall.com/aboutus>
- [43] N. Totenberg and S. McCammon. (Jun. 2022). *Supreme Court Overturns Roe v. Wade, Ending Right to Abortion Upheld for Decades*. NPR. [Online]. Available: <https://www.npr.org/2022/06/24/1102305878/supreme-court-abortion-roe-v-wade-decision-overturn>
- [44] Google. (2022). *Google U.S. Political News Feed*. [Online]. Available: <https://news.google.com/topics/CAAqBwgKMJ3FogkwJeYAg>
- [45] A. Wongprommoon. (May 2021). *Welcome to MIT News Classify's documentation!—MIT News Classify 1.0 Documentation*. [Online]. Available: <https://news-classify.readthedocs.io/en/latest/>
- [46] E. Sandhaus, "The New York times annotated corpus," *Linguistic Data Consortium, Philadelphia*, vol. 6, Dec. 2005, Art. no. e26752.

- [47] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (Oct.2013). *Distributed Representations of Words and Phrases and Their Compositionality*. [Online]. Available: <https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>
- [48] R. Rehurek. (2022). *Gensim: Topic Modelling for Humans*. [Online]. Available: <https://radimrehurek.com/gensim/models/phrases.html>
- [49] GitHub. (Mar. 2019). *Research Google-Research/Bert*. [Online]. Available: <https://github.com/google-research/bert>
- [50] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1–19.
- [51] R. Tumin. (Jun. 2022). *Special Edition: Roe v. Wade Is Overturned*. The New York Times. [Online]. Available: <https://www.nytimes.com/2022/06/24/briefing/roe-v-wade-abortion-supreme-court-guns.html>
- [52] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay," 2018, *arXiv:1803.09820*.
- [53] A. Maiya. (Aug. 2020). *Ktrain: A Lightweight Wrapper for Keras to Help Train Neural Networks*. [Online]. Available: <https://towardsdatascience.com/ktrain-a-lightweight-wrapper-for-keras-to-help-train-neural-networks-82851ba889c>
- [54] M. Saeed. (Sep. 2021). *An Introduction To Recurrent Neural Networks and The Math That Powers Them*. Machine Learning Mastery. [Online]. Available: <https://machinelearningmastery.com/an-introduction-to-recurrent-neural-networks-and-the-math-that-powers-them/>
- [55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997, doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [56] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, *Others Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*, 1st ed. Piscataway, NJ, USA: IEEE Press, Jan. 2001.
- [57] K. Maryada. (2017). *Simple LSTM for Text Classification*. [Online]. Available: <https://www.kaggle.com/code/kredy10/simple-lstm-for-text-classification>
- [58] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks Of The Trade*. Cham, Switzerland: Springer, 1998, pp. 55–69.
- [59] A. Rawat and A. Solanki, "Sequence imputation using machine learning with early stopping mechanism," in *Proc. Int. Conf. Comput. Perform. Eval. (ComPE)*, Jul. 2020, pp. 859–863.
- [60] C. School. (Feb. 2018). *Which News Outlets Have the Most Power? Columbia Business School Research Spotlights Information Inequality in 18 Countries and Identifies Top News Outlets*. Newsroom. [Online]. Available: <https://www8.gsb.columbia.edu/newsroom/newsn/5968/which-news-outlets-have-the-most-power-columbia-business-school-research-spotlights-information-inequality-in-18-countries-and-identifies-top-news-outlets>



**ERIC NESS** received the Bachelor of Science degree in information technology from the University of the District of Columbia and the master's (M.Sc.) degree in computer science from Anglia Ruskin University. He is currently the Chief Technical Officer with Performeks LLC, Washington, DC, USA, specializing in analyzing environmental data for numerous Fortune 500 companies, most of which are in the extractive industries. Previously, he was an U.S. Reporter with press credentials for the U.S. Capitol, Department of Labor, and the Department of Commerce. In other significant projects, he has been a Technical Lead in the USAID Project called HealthSystems 20/20, which analyzed the state of every country's health systems. Another USAID project took him to Rwanda for their national health assessment (NHA) which tracks how money is allocated across a health systems. Over the last decade, he has done considerable work in natural language processing analyzing news. Some projects include monitoring environmental information, tracking HIV/aids, textual analysis regarding corruption in court cases in Indonesia, reputational analysis for a Fortune 500 mining operation, and multiple political candidates in the USA for elections.



**AROJ FATIMA** received the double B.Sc. degree in mathematics and statistics and the M.Sc. and Ph.D. degrees in computer science from Anglia Ruskin University, U.K., in 2016. She is currently a Senior Lecturer with the School of Computing and Information Sciences, Anglia Ruskin University, U.K. Following the B.Sc. degree, she pursued her career as a Data Coordinator with Science Foundation, Pakistan, leading a team to execute automation project for research and travel grants. After finishing the Ph.D. degree, she started her career as an Academic with Anglia Ruskin University. She has experience with knowledge transfer projects (KTP) funding and deployment. She has supervised several projects in machine learning and NLP. During the Ph.D. degree and her career, she managed to publish several articles in various international journals and conferences. Her research interests include natural language processing, semantic data technologies, and machine learning.



**MAHDI MAKTABDAR OGHAS** received the B.E. degree in software engineering from Islamic Azad University, Iran, and the M.Sc. and Ph.D. degrees in computer science from University Technology Malaysia (UTM), in 2016. He is currently a Senior Lecturer with the School of Computing and Information Sciences, Anglia Ruskin University, U.K. His primary research focus at UTM was computer vision and machine learning in specific accurate skin detection for medical and security applications. In 2016, he started his career as a Postdoctoral Researcher with UTM and joined a research project sponsored by Cyber Security Malaysia and the Ministry of Higher Education Malaysia, aimed to promote safety and security in cyberspace using artificial intelligence and machine learning techniques. Later, he joined Kingston University London ROVIT research team to work on the H2020 MONICA project, aimed to promote the crowd safety and security in large-scale outdoor events using video analytics, artificial intelligence, and computer vision techniques. In 2019, he progressed his career to work as a Lecturer with the School of Computing and Information Science, Anglia Ruskin University. As a result of his research career, he managed to publish several articles in various international journals and conferences. His primary research interests include deep learning and convolutional neural networks, machine learning, crowd analysis, and medical image processing.

...