Detecting Biases in Newspapers using Natural Language Processing

Syeda Jannatul Ferdous, Purobi Paromita, Puspita Das, Sabbir Bin Abdul Latif,
Farah Binta Haque, Adib Muhamma Amit and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{syeda.jannatul.ferdous, purobi.paromita, puspita.das, sabbir.bin.abdullatif, farah.binta.haque,
adib.muhammad.amit}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—In this information-driven world, media bias on public perception is now a growing concern. So, this paper uses Natural Language Processing to address the difficulty of finding biases in news articles. Here, we examine an enhanced expert-labeled dataset of news article sentences using a combination of sentiment analysis and traditional machine learning approaches such as Naive Bayes and Logistic Regression. We have followed a systematic approach from data collection and preprocessing to model training and evaluation. This study highlights NLP's potential in media analysis and sets the foundation for future research in this field, highlighting its importance in supporting transparent journalism and a knowledgeable society.**

I. Introduction

In today's age of information, the news media plays an important role in shaping public opinion and guiding social interactions. To guarantee that the public is presented with factual information free from political bias and ideological influence, unbiased reporting is essential for a strong, democratic society. Unfortunately, this goal is being severely compromised by biased reporting, which is frequently motivated by political, economic, or ideological agendas. The influence of media bias significantly affects both individual and public perceptions of news reports, which impacts political choices [4]. News that has been biased can deepen societal divisions, spread misinformation, and influence public opinion. In an era when information is abundant, the proficiency to detect bias in news content is more important than ever.

News media bias can appear in different ways, including minor variations in language to explicit biased political speech. It also can be presented by constructing narratives, selecting particular issues, establishing many sources, or even ignoring important details. Finding such bias is a difficult task that involves several different aspects. Conventional approaches mostly rely on human assessment, which is very labor-intensive and has the possibility of biases and limits. The rise of digital news media and the rapid increase in the amount of news content need a strategy that can handle large amounts of information and consistently evaluate biases.

With the use of Natural Language Processing, this paper aims to address this problem with the detection method of newspaper biases. NLP is a field that combines computer science, artificial intelligence, and linguistics and provides a vast number of strong tools for analyzing and understanding human language at large. This study aims to examine and showcase the successful utilization of several NLP techniques for identifying biases in news articles. This project aims to utilize sentiment analysis, and machine learning classification along with advanced text processing techniques on an expert-labeled dataset of newspapers from various sources. The objective of the paper is to identify patterns and signs of bias, providing an improved and scalable tool for understanding media bias.

II. Existing Work

The study of media bias detection has a lengthy and rich history, in which researchers implemented several techniques to identify and analyze it. However, these techniques frequently depend on individual established norms that might not contain the complete range of biases. Recent studies have made progress in detecting news biases using Natural Language Processing. Gangula, Duggenpudi, and Mamidi (2019) proposed a new headline-focused methodology. Their headline attention network emphasizes headlines in bias detection to mirror how headlines influence readers [3]. This unique way of understanding headlines may ignore the complex nature of bias in the full articles. Headlines can be powerful, but they may not reflect the article's full biases. Our research fills this gap by analyzing all news material, beyond headlines, to understand bias and context.

(Cox & Acharya, 2021) presents a study that was carried out to identify occurrences of bias in reporting that were found in articles created by four major news organizations. They have used VADER from the Natural Language Toolkit (NLTK) for sentiment analysis. This method is different since it makes use of VADER, a technique for bias detection that hasn't been thoroughly studied in the literature up to this point [2]. However, this paper may not detect complete bias because of its complex nature, and the only method it employs is sentiment analysis, which may cause it to overlook actual biases in the words or sentences of articles.

III. Dataset

In our study, to identify bias in newspapers we will utilize Natural Language Processing (NLP) methods focusing on the ”BABE” (Bias Annotations By Experts) dataset. Spinde et al. Devised a technique to detect media bias by utilizing supervision with BABE [5]. This dataset is designed to offer a complete overview of articles from diverse newspaper websites, containing a wide range of political and ideological viewpoints.

A. Features of the Dataset

The dataset used in this analysis comprises several expert annotated texts, with 3700 articles. This diverse collection have a range of sample sizes for analysis. It includes texts from various news sources ensuring a variety of political ideologies viewpoints.

Each entry in the dataset contains the labeled text of the article indicating whether it is biased or non-biased. This labeling helps to analyze and train the model with the text. The dataset also includes metadata about each article's publishing outlet, type of the article, etc which is crucial for evaluating biases on specific to each source.

The articles in this dataset are classified as politics, environment, economy, and many more which allows us to examine biases within these areas or topics.

One notable aspect of this dataset is its label of articles as either 'Biassed' or 'Non-biased' by experts. These labels perform as a foundation, for supervised machine-learning models since they were assigned based on standards developed by media specialists.

B. Implementation in our Study

Correlation Analysis: Our goal is to explore the level of correlation between the articles' linguistic features and the biases labeled on them.

Sentiment Analysis: We can find out more about how emotional tone could be related to a sense of bias by considering the sentiment of the texts.

Machine Learning Models: The dataset will be utilized to train and evaluate different machine learning models, such as Naive Bayes, and Logistic Regression, to automatically identify bias in news articles.

The dataset's diverse and extensive content makes it an optimal selection for this research on newspaper bias, offering a solid basis for both quantitative and qualitative analysis.

IV. Methodology

The section will explain the whole process of our work starting from dataset organization to algorithm implementations using NLP. Here, we have applied the necessary checks to make the dataset ready to be trained with our machine-learning models.

The steps followed are shown in the “Fig: 1”.

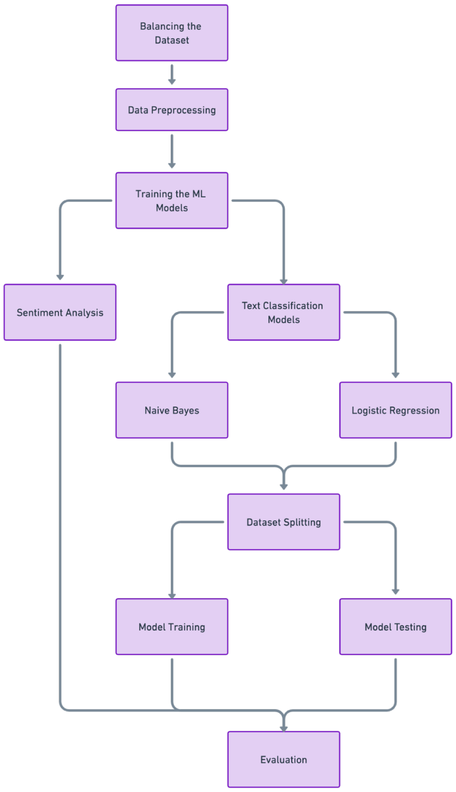

Fig. 1. Methodology Flowchart

A. Balancing the Dataset

For the study, we've used a dataset named 'BABE' which includes news articles from diverse political ideologies. But, because the dataset was labeled manually, we needed to verify whether the dataset is balanced or not. For example, if we have 300 biased labels and 3400 labels that are unbiased for training, our algorithm is unlikely with correctly recognize the bias. The dataset consists 50.7% unbiased labels and 49.3% biased labels shown in the “Fig: 2”.

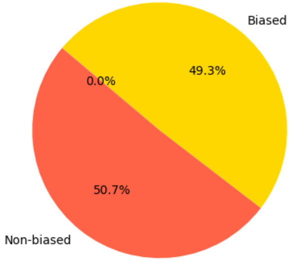

Fig. 2. Portion of Articles Labeled Biased vs Non-biased

B. Data Preprocessing

Data preprocessing is an essential and crucial step in the preparation of data for analysis. Data cleaning and organization are performed to improve the quality and usability of raw data. Inaccurate processing of data can lead to misleading results and biased consequences. To preprocess the data, we undertook the following procedures:

- Removing null value
- Removing line break
- Removing extra spaces
- All texts converted to lowercase
- Removing punctuation
- Removing stopwords

C. Training the ML Models

1) Sentiment Analysis: Sentiment Analysis is a method used to identify and categorize opinions in text, determining the writer's attitude towards a topic or the overall tonal polarity. It helps understand the emotional tone of news articles, assessing how different news outlets present topics with varying degrees of positivity, negativity, or neutrality. The TextBlob library is used for this study.

   Process:

   - Text Processing: The text of each article is first analyzed, typically dividing it into smaller portions like words or phrases.
   - Polarity Score Assignment: TextBlob then assesses the sentiment of these pieces, giving a polarity score between -1 and 1. A score of -1 indicates a highly negative sentiment, 0 denotes neutrality, and 1 means a highly positive sentiment.
   - Contextual Analysis: The library considers not only individual words but also their context and composition, giving a more advanced evaluation of sentiment.

   To find bias in news articles, sentiment analysis is very important. Comparative research shows patterns and trends in things, which might match up with political beliefs [1]. Labeling articles as positive or negative isn't the only thing this method does; it also finds the deeper emotional themes in articles. It helps us understand how different newspapers shape their narratives, showing the many levels of bias and sentiment. Given that the majority of sentiments in our dataset are centered around zero shown in "Fig: 3". Which suggests a neutral sentiment. So, our following research on bias in newspapers has the potential to provide a balanced insight.

2) Text Classification Models: Text classification models use machine learning methods to put text into predefined groups, such as "Biassed" or "Non-biased." This helps us find biases in news articles and learn more about how media is distributed.

   - **Naive Bayes Classifier:** It is a probabilistic classifier that predicts based on Bayes' Theorem. Naive Bayes is simple, efficient, and excellent at text classification, especially for huge datasets. The simple implementation makes it useful for basic baseline models. The formula of Naive Bayes is:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \tag{1}$$

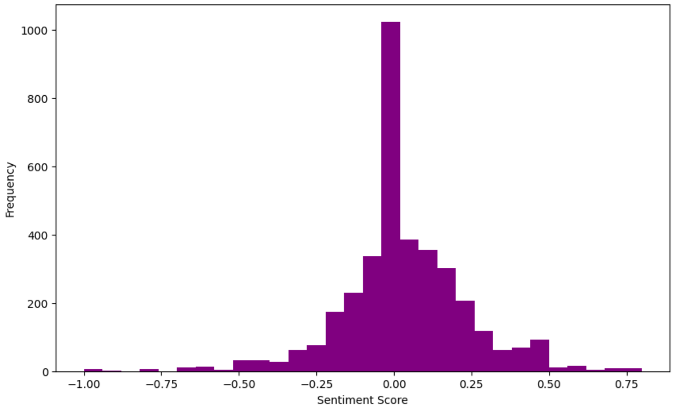

Fig. 3. Distribution of Sentiment Scores

where $P(C|X)$ is the probability of class C given the features X.

- **Logistic Regression:** A statistical model that predicts a binary outcome (such as 'Biassed' or 'Non-biased') from one or more independent factors. Logistic Regression is useful when the connection between the input features and the output label is both linear and probabilistic. It can handle high-dimensional data, making it ideal for text analysis. We used TF-IDF for Logistic Regression. The formula of Logistic Regression is:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{2}$$

where $P(Y = 1|X)$ is the probability of the dependent variable Y being 1 given the features X.

  - TF-IDF Vectorization: TF-IDF Vectorization is a statistical metric used to assess the relevance of words to documents in text mining and information retrieval. It helps identify important terms in each document and improve feature set analysis by converting text data into numerical form for machine learning algorithms.
  - How TF-IDF Works:
    - Term Frequency (TF): A document's term frequency is calculated by dividing its number of appearances by its total terms. The formula of Term Frequency (TF):

$$\text{TF}(t, d) = \frac{\text{Term } t \text{ frequency in document } d}{\text{Total terms in document } d} \tag{3}$$

      Higher frequency indicates greater importance.
    - Inverse Document Frequency (IDF): This calculates word significance across documents. It is the logarithm of the total number of documents divided by the term-containing docu-

ments. Inverse Document Frequency (IDF):

$$\text{IDF}(t, D) = \log\left(\frac{\text{Total documents}}{\text{Documents with term } t}\right) \tag{4}$$

Terms in many articles have less significance than terms in fewer articles.

∗ TF-IDF Calculation: The TF-IDF value for each term is obtained by multiplying its TF score with its IDF score.
Formula:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{5}$$

In this work, classification techniques help interpret complex articles. These models objectively categorize these articles to analyze and understand media biases.

Process:
The dataset was split into two parts: training and testing set (70% and 30%). The training set was used to learn patterns and correlations. During the testing phase, the model was evaluated based on data that had not been seen before. This was done to ensure that the model could generalize and generate correct predictions without overfitting.

a) Model Training::

- Data Feeding: The models are fed a training set that after vectorized and processed through TF-IDF.
- Learning: The models learn how to relate text attributes like word frequency and phrases to desired outputs like 'Biassed' or 'Non-biased'.
- Iteration: The model adjusts its internal parameters to enhance its predictions across numerous iterations.

b) Model Testing::

- Evaluation on New Data: After the training process, the models are evaluated on the testing set.
- Performance Measurement: Model predictions are compared to actual results to assess accuracy and other metrics.

c) Evaluation Metrics::

- Accuracy: Calculates the percentage of total correct predictions made by the model out of all predictions. Formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{6}$$

- Precision: The percentage of correctly predicted positive observations to the total predicted positive observations. Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

- Recall (Sensitivity): The percentage of correctly predicted positive observations to all observations in the actual class. Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{8}$$

- F1-Score: The harmonic mean of Precision and Recall, providing a balance between them. Formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

V. Results and Discussion

The models showed different levels of efficacy in identifying biases present in the given news articles.

A. Model Performance

*1) **Naive Bayes Classifier**:*

- Accuracy 74.52%: This indicates that the Naive Bayes model correctly classified approximately 74.52% of the articles in the test set as either 'Biased' or 'Non-biased'.
- Precision: For articles labeled as 'Biased', the model's precision is 73%, meaning that when it predicts an article as 'Biased', it is correct 73% of the time. For 'Non-biased' articles, this figure is 76%.
- Recall The model has a recall of 75% for 'Biased' articles, indicating it correctly identifies 75% of the actual 'Biased' articles. For 'Non-biased' articles, the recall is 74%.
- F1-Score: This is the harmonic mean of precision and recall. The Naive Bayes model scores 74% for 'Biased' and 75% for 'Non-biased' articles, suggesting a balanced performance between precision and recall.

*2) **Logistic Regression Model**:*

- Accuracy (73.25%): The Logistic Regression model correctly classifies 73.25% of the test set. It's slightly lower than the Naive Bayes model.
- Precision: It has a precision of 71% for 'Biased' and 76% for 'Non-biased', indicating it's slightly less precise than Naive Bayes in predicting 'Biased' articles but equally precise for 'Non-biased' articles.
- Recall: The recall for 'Biased' articles is 75% (same as Naive Bayes), but it drops to 72% for 'Non-biased' articles.
- F1-Score: The model scores 73% for 'Biased' and 74% for 'Non-biased' articles, slightly lower than Naive Bayes.

B. Analysis

The results show that both models have a reasonable level of accuracy in classifying articles. The Naive Bayes classifier has slightly better overall accuracy and a more equal performance in terms of precision and recall. The similar recall scores for 'Biassed' articles suggest both models have equivalent proficiency in detecting actual 'Biassed' articles. However, differences arise when identifying 'Non-biased' articles.

C. Comparisons

In our study, Naive Bayes slightly outperforms Logistic Regression in overall accuracy, precision, and F1-scores. This might be due to the different ways these models handle the underlying data and feature relationships. Despite its simplicity, Naive Bayes competes closely with Logistic Regression

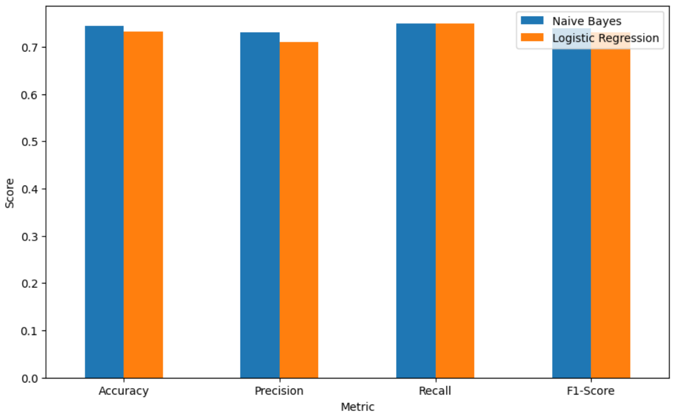

Fig. 4. Comparison of Model Performance Metrics

which is a more complex model. This might indicate that the feature space and the nature of the dataset are well-suited to Naive Bayes' probabilistic approach.

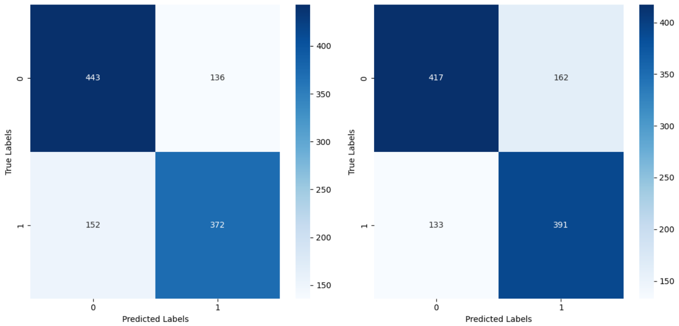

Fig. 5. Confusion Matrix

The close performance of both models suggests that even simple models can be quite effective for certain NLP tasks, and complex models do not always guarantee significantly better results. In summary, these findings offer valuable indication of the efficacy of traditional NLP models in detecting bias and can serve as a guide for future improvements.

D. Limitations

Our study has some limitations. The dataset is not so large, also it may not fully capture the diverse and complex nature of linguistic biases. Also, individuals who are involved in labeling bias have the potential to produce inconsistencies in the data, which may have an impact on the training features of the model.

VI. Conclusion

Our study thoroughly explores the use of Natural Language Processing to detect bias in newspapers. Our model shows the efficacy of various NLP techniques, including traditional machine learning models such as Naive Bayes and Logistic Regression. The findings are significant for news article analysis and journalism, as they try to maintain ethical reporting standards. The techniques could be integrated into news aggregation platforms for balancing political ideologies. The study also highlights the potential of NLP in analyzing media content, which will create a more informed and unbiased society. Future research could explore other NLP techniques, apply them to larger datasets, extend the study to other media forms, and integrate multimodal analysis for a comprehensive approach.

References

[1] Wael F. Al-Sarraj and Heba M. Lubbad. Bias detection of palestinian/israeli conflict in western media: A sentiment analysis experimental study. In *2018 International Conference on Promising Electronic Technologies (ICPET)*, pages 98–103, 2018.

[2] Grace Cox and Anuska Acharya. Sentiment analysis and nlp models for identifying biases of online news stations. Master's thesis, Volgenau School of Engineering, MARS, 04 2021.

[3] Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. Detecting political bias in news articles using headline attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy, August 2019. Association for Computational Linguistics.

[4] Felix Hamborg. *Towards Automated Frame Analysis: Natural Language Processing Techniques to Reveal Media Bias in News Articles*. PhD thesis, 01 2022.

[5] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.