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 [3]. 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 [2]. 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 [1]. 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 [4]. 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

Many algorithms can be used to detect bias in newspapers such as sentiment analysis and ML models such as Text Classification Models. We are researching some of the algorithms and have made a model that can be useful to reach our research goals.

V. Result

Our model is almost ready. As soon as we are done with our testing we will attach our results here.

VI. Conclusion

We are still working on our research and have yet to analyze the results that are to be used to meet our needs. We haven't reached any conclusion yet.

References

[1] 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.

[2] 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.

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

[4] 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.