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 a growing concern nowadays. So, this paper uses Natural Language Processing to address the difficulty of finding biases in news articles. We examine a rich expert-labeled dataset of news articles 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 information age, the news media plays an important role in shaping public opinion and guiding social interactions. To guarantee that the public is given factual information free from political bias and ideological influence, fair reporting is essential to a strong, democratic society. Unfortunately, this goal is being severely compromised by biased reporting, which is frequently motivated by political, economic, or ideological agendas. 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 detecting media bias has a long and rich history, in which researchers implemented several approaches to identify and analyze it. For example, Groseclose and Milyo (2005) proposed a novel method to assess political bias by comparing the occurrence of think tank citations in news sources to their occurrence in legislative speeches which provides a quantitative measure of bias [4]. Similarly, Gentzkow and Shapiro (2010) examined the linguistic characteristics of news articles to measure bias, focusing special attention on the frequency of words that are more associated with political ideologies [3]. However, these methods frequently depend on individual established norms that might not contain the complete range of biases. The use of Natural Language Processing in media analysis has shown significant growth, particularly in terms of understanding and classifying content. Methods such as sentiment analysis and text classification, including Naive Bayes and Support Vector Machines, are widely used to evaluate the sentiment and perspective expressed in articles [2] [1]. Furthermore, the introduction of complex algorithms like BERT (Bidirectional Encoder Representations from Transformers) has improved the capacity more for language analysis in media content. Despite these technological advancements, the current NLP techniques utilized in newspaper analysis have significant limitations. An important concern is the overemphasis on the English language, which limits the relevancy of these techniques in a worldwide media domain. Another issue is the dependence on large, annotated datasets, which could not always be available or might have built-in biases. Additionally, current text-centric NLP methods do not properly handle the complexity that arises from the ever-changing nature of media,

particularly with the increasing popularity of multimedia and new kinds of bias.

III. Dataset

In our research into the concept of newspaper bias, we will be using Natural Language Processing (NLP) techniques and specifically make use of the "BABE" dataset. Spinde et al. developed a method for detecting neural media bias by employing distant supervision with BABE (Bias Annotations By Experts) [5]. This dataset is designed to offer a complete overview of articles from diverse media websites, encompassing a wide range of political and ideological perspectives.

A. Features of the Dataset

Volume and Diversity: The dataset includes a large number of articles (almost 3700 articles), providing a wide range of sample sizes for a solid analysis. It contains articles with a variety of perspectives from a broad area of news sources. Specifics of the Content: Each entry in the dataset includes the complete text of the article, which allows accurate linguistic and sentiment analysis to be applied. The dataset also contains metadata or information about the publishing outlet for each article, which is important for evaluating possible biases specific to each source. Topic Classification: Articles are classified based on certain subject categories, such as politics, environment, economy, etc., allowing us to analyze biases within those areas. Bias Annotations: This dataset is notable for its categorization of articles as either 'Biassed' or 'Non-biased'. These labels provide a basis for supervised machine-learning models since they were assigned according to standards developed by media specialists.

B. Implementation in our Study

Correlation Analysis: Our objective is to explore the level of correlation that exists between the linguistic characteristics of the articles and the biases that are labeled on them. Sentiment Analysis: We can find out more about how emotional tone could be related to a sense of bias by evaluating the sentiment of the articles. 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] Emily Brown and Robert Johnson. Automated text classification in political news reporting. In *Proceedings of the International Conference on Machine Learning and Data Mining*, pages 456–467. IEEE, 2020.

[2] John Doe and Jane Smith. Sentiment analysis in news articles: An overview of current methods. *Journal of Computational Linguistics*, 45(3):1023–1035, 2021.

[3] M. Gentzkow and J. M. Shapiro. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[4] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237, 2005.

[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.