Dissertation Proposal: *Media Bias Detection System*
Name: *100498877*
Word count: *2342*

1 Introduction

In this new age of technology, internet is often chosen as the primary source of information compared to traditional media. However, the quest for engagement and profit has brought about editorial bias, emotionally charged journalism, and the proliferation of opinionated content. Cognitive biases like naıve realism and confirmation bias (Nickerson, 1998) worsens the consequences of such media propaganda, also makes it harder for the public to comprehend inducing polarizing opinions (Rodrigo-Ginés et al., 2023).

Despite the abundance of data, the audience remains vulnerable to biased content. While comparative news aggregators and bias indices attempt to offer balanced perspectives, their impact is limited due to fragmented definitions of media bias. In controlled environments like China, state ownership coexists with market dynamics, producing variation in bias and highlighting how media systems adapt to political and economic pressures (Qin et al., 2018).

In order to circumvent these challenges, Natural Language Processing (NLP)-powered expert systems have become vital resources for detecting, categorizing, and reducing media bias. The current solutions lack maturity, facing issues in robustness, accuracy, and multilingual scalability.

This dissertation presents a systematic review of media bias literature and proposes a unified taxonomy of bias types. The paper reviews pre-existing bias detection frameworks ranging from transformers to rule based methods while exploring datasets along with measuring system's efficiency. Using hybrid machine learning pipelines, the main objective of this work is to create a complete media bias detection system that can be used in real-time, multilingual, and scalable applications. So, in order to preserve media integrity and accountability towards their audience, to fortify the foundation of journalism.

2 Related Work

2.1 Media Bias Gaza 2023 - 24

The Centre for Media Monitoring's Media Bias Gaza 2023–24 research offers a thorough analysis of UK and some foreign media coverage of the October 7, 2023, Hamas attack and Israel's military response. The main accusation of this research is that there is a significant bias in the mainstream media about pro - Israel, evidently in areas mainly language, context, framing, allegations, marginalisation of Palestinian sources and Islamophobic undertones.

Crucially, the media frequently decontextualised the asymmetric character of the violence by refusing to impart historical context, especially the occupation of Palestinian territory, and by portraying the conflict as a "Israel-Hamas war." Broadcast coverage hardly used phrases like "occupied Gaza," with Al Jazeera English being the only outlet to mention the more general occupation.

The references to the war substantially favoured Israeli narratives, further outnumbering the mentions to Palestinian rights to resistance. Israeli victims were often referred to "massacred" or "atrocities" with an empathetic overtone while the Palestinian deaths were reported with less accountability. A key issue was the unjustified spread of Israeli military claims, many of which were later debunked, like the false "beheaded babies" accusation.

Despite being groundless, these claims were by major media outlets without proper retraction or correction. This selective reporting fosters a skewed media landscape, one that marginalizes Palestinian voices and distorts public understanding, reinforcing a biased, one-sided narrative.

Thus, the research makes the enlights us with the concern that this level of biased journalism can skew public perception. This showcases an ethical and media failing (Canli and Toklu (2021)).

2.2 Media Bias Taxonomy

The research undertaken between January 2019 and May 2022 is prioritised in this article. The review proposes a Media Bias Taxonomy comprising five categories: linguistic bias, text-level context bias, reporting-level context bias, cognitive bias, and related concepts like framing effects and group biases. The taxonomy, with the use of sentiment analysis and word embeddings, addresses the stereotypes in various frameworks like language, structure, topic selection and audience interpretation.

The article categorizes media bias detection methods into traditional natural language processing (tNLP), machine learning (ML), and graph-based approaches. Traditional NLP methodologies like Word2Vec, were referred to measure Word Frequencies and embedding, to detect implicit biases. This review uses a ML models based on Transformers, fine tuned for identifying framing and spin bias in language, political stances, group bias etc. BERT, RoBERTa and their derivatives were some of the models used. Complementarily, non-transformer ML methodologies, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and bidirectional long short-term memory networks (BiLSTMs), are also employed, particularly in the detection of hyperpartisan news and political stances. Furthermore, non-neural network (nNN) tends to be surpassed by deep learning models, yet they try to maintain relevance due to their interpretability and computational efficiency using some models like support vector machines (SVM), logistic regression (LR), and XGBoost.

Graph based methodologies were implemented by taking advantage of the structural properties of online networks to discern media bias. Graph neural networks and autoencoders, analyse the user interactions and content transmission patterns to clarify ideological statements and the formation of echo chambers.

The article concludes that while various techniques show promise, more research is needed to refine interventions aimed at reducing cognitive media bias. There is insufficient integration of advanced computer science techniques with psychological and social psychological insights related to media bias, which can limit the effectiveness of detection methods. The study found that existing datasets largely neglect to include background information on annotators, which is critical for understanding and improving the accuracy of bias annotations. Different methods, such as encouraging diverse media exposure, mindset changes, and forewarning messages, have yielded small effects. Additionally, interventions may not be equally effective for different political groups, suggesting a need for further testing (Spinde et al. (2023)).

2.3 A Domain-adaptive Pre-training Approach for Language Bias Detection in News

The stereotypes depicted in media has progressively evolved from linguistic cues to transformer-based models, that has the ability to capture nuanced context. Traditional methodologies struggled with subtle forms of biases, due to their reliance on shallow syntactic patterns (Recasens and Jurafsky (2013)). RoBERTa, comparetively performs better than the other models suggesting that encoder-only architectures were effective in detecting bias. Models like ExT5 might offer generalizability but demands significant number of resources. Additionally, the widespread use of F1 score as a primary evaluation is insufficient for capturing the nuanced nature of bias.

Improvment in bias detection tasks(Spinde et al. (2021a)) was demonstrated since using models like BERT, RoBERTa, BART, T5 etc (Vaswani et al. (2017)). However, challenges persists. The main concern is the limited availability of annotated datasets, the narrow focus on sentence level bias, neglecting word/article level nuances. (Spinde et al. (2023)). An exploration using distant supervision (Spinde et al. (2021a)) and multi task (Spinde et al. (2022)) learning were employed to address data limitations, yet this introduces label noise and MTL diffuse bias specific learning and they completely rely on such datasets (Spinde et al. (2021a)). Hybrid models that combine domain adaptation with MTL strategies hold particular promise for improving both performance and interpretability in media bias detection (Spinde et al. (2023)).

2.4 News Bias Detection with Pre-Trained Neural Transformers

The study explores the capabilities of GPT 3.5, GPT 4, and Llama2 using MBIC dataset (Spinde et al. (2021b))through three different phases. From the experiments, they found that a fine - tuned GPT 3.5 performs better in detecting bias in an input compared to GPT 4. However, comparing the two models, GPT 4 outperforms GPT 3.5 when input was evaluated independently. Each model uses uniform prompts as well as bias types, outputs and strength scores (Wessel et al. (2023b)).

The third phase tested prompt engineering strategies: removing or simplifying bias definitions, adjusting prompt clarity, filtering weakly biased outputs (bias score ¡ 0.6), altering prompt structure, and changing temperature settings(Menzner and Leidner (2022)).

The elimination of criteria enhanced F1 but resulted in inconsistent bias classification. The fine-tuned GPT-3.5 exhibited the highest consistency with specified bias categories. The noted issues include the models' incapacity to identify fake information and reported speech, their failure to differentiate between language and meta-language, their lack of contextual understanding, and their propensity to misclassify authentic statements (Menzner and Leidner, 2022).

The MBIC dataset was exacerbated by its U.S.-centricity and annotation litigations. The study builds attention to the trade-offs faced by media bias detection encompasses between prompt design, model architecture, and context. It suggests transparent, responsible deployment and suggests working together to produce a multilingual dataset, a browser plug-in for user awareness, and an open-source, on-premise bias detection program (Menzner and Leidner (2022)).

2.5 Ara-BERT: Media Bias Detection in Israel-Gaza Conflict

This review renders forth current technique, tools, frameworks, and issues with media bias identification, with an objective on traditional machine learning, deep learning, and transformer-based methods.

Detecting media bias in polarised geopolitical issues, such as the Israel-Gaza conflict, is challenging due to the cultural, linguistic, and ideological layers that pervade reporting (SinaLab, 2024).

Traditional machine learning methodologies like SVM, Logistic Regression, Random Forest etc, such supervised learning models along with structured characteristics using TF-IDF and n-grams, remains comprehensible (Nadeem and Raza, 2021) yet they struggle with contextual depth. Unsupervised learning models like K-Means, provide alternative solutions to high resource scenarios, although they garner less attention.

Deep learning models—particularly LSTM, Bi-LSTM, and Bi-GRU, attention mechanisms refine their focus on bias-indicative text. However, languages with complex morphology, like Arabic, hinder their performance without substantial data.

Transformer-based models such as BERT, T5 (Raffel et al., 2019), and AraBERT, E5 (Wang

et al., 2024) offer cross-lingual potential. As per information, the biased content focuses on minority class, caused by class imbalance which is aided using techniques like Borderline-SMOTE (Han et al., 2005).

Some of the prevalent issues noticed are when different annotators are asked to label subjectively, multilingual preprocessing (especially Arabic dialects), and class imbalance (Evans et al., 2024). Even strong models like AraBERT (Antoun et al., 2020) and T5-small struggle on minority classes or fail to generalize without task alignment as cultural sensitivity has to be considered when making accurate detections.

3 Methodologies

In order to detect media bias intuitively, this experiment proposes a transformer-centric approach, using the Media Bias Identification Benchmark (MBIB) corpus (Wessel et al., 2023b) as the basis for evaluation. The five main steps of the procedure are dataset collecting, preprocessing and, model fine-tuning.

3.1 Data Preprocessing

The preprocessing framework is applied on a CSV file containing the cognitive bias dataset. This dataset comprises both the raw text data and the labels that categorise the text as biassed or non-biased. The information is structured in a tabular format with each row denoting a text sample and the label indicating whether cognitive bias is present or not. Following the preprocessing stage, tokenization is performed. Tokenization serves as the first step in transforming human-readable text into a machine-understandable format. Tokenization is performed using the tokenizer that corresponds to the pre-trained language model chosen for the task, such as RoBERTa, GPT-2, which is preferred for this project. The raw text from the CSV file is then encoded into tokens for quicker processing. Transformer models do not require the data to be converted to numbers. Following tokenization, the next step is to generate Attention masks are binary arrays that indicate which tokens in a sequence are actual words and which are padding tokens. Padding, to ensure that all input sequences are of the same length, shorter sequences are padded with special tokens. An attention mask gets generated for every tokenised sequence, with padding tokens having a value of 0 and genuine tokens having a value of 1. The model only pays attention to the significant portions of the input sequence. Well the array differs from a normal supervised learning like Random Forest compared to Transformer-based models so the chances of using either of them is indecisive.

3.2 Transformer - based models

In this project, we use Transformers module from Hugging Face in developing a transformer-based sequence classification pipeline. Utilising stratified k-fold cross-validation for performance assessment, one of the main goals is to assess the pre-trained transformer models on a bias classification problem.

The dataset, stored as a CSV file named after the task, is loaded and then tokenized using the appropriate tokenizer from the model specifications used for this project. The model employs tokenizer, and learning rate of the respective transformer architecture, such as BERT, RoBERTa, BART, ConvBERT, GPT-2, or ELECTRA which will be taken into consideration for evaluating model efficiency.

Training is managed in the fitting process, which supports early stopping based on validation loss. For each epoch, gradients are reset, and a forward pass is performed. Loss is computed and used for backpropagation. A cosine learning rate scheduler helps optimize convergence.

This is one of the most commonly used methodologies when using transformers. No gradient computation is performed during evaluation to save memory.

Overall, I prefer the said models from Hugging Face as they are preferred more often over other models that require substantial hardware and sufficient memory for detection. Where models like Roberta use Bi-directional understanding for better detection.

3.3 Evaluation

Our model's evaluation is done using stratified k-fold cross-validation that guarantees a balanced label distribution between the training and validation sets. The model is tested on test / validation samples to evaluate its performance. Performance is assessed using using a weighted F1-score that balances recall and precision while addressing class imbalance. The process is repeated multiple times, and the final result is the average F1-score, which is often a reliable measure of the model's ability to detect biased text.

4 Analysis and Risk

There is some ethical risk associated with this study because it employs publicly available text data rather than sensitive or private information. Another potential issue is the model's misclassification, which might be biased as a neutral content. The study ensures that AI is used ethically for bias detection by adhering to ethical norms for data processing, reproducibility and transparency. From **Fig: 1** we comprehend that the model used in this dataset performs decently on BigNews approximately showing 0.67 on F1 score yet having a score of 0.3 on Liar, indicating that detecting intentional deception can be challenging. A potential risk in vague labelling or lack of context can lead to mis classification of bias types.

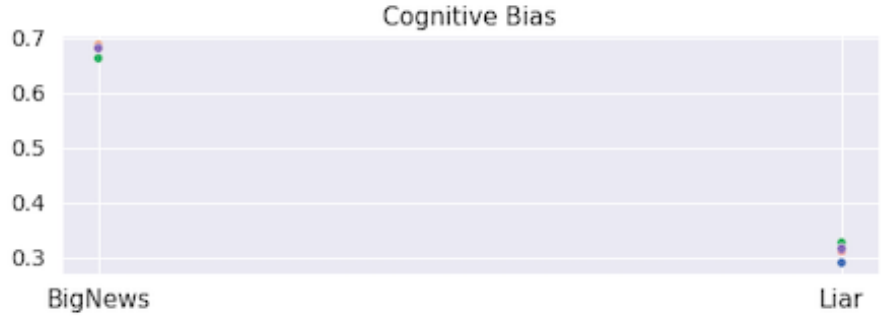

Figure 1: Cognitive Bias for Fake News Detection (Wessel et al., 2023a)

5 Project Plan and Gantt Chart

The project enters into the implementation phase, beginning with Data Collection (MBIC dataset), which has to be gathered, cleaned and scaled properly for our use. After setting up the necessary Environmental packages, we have to configure and test the library tools, frameworks along with other models. Next, involves loading the transformer-based model and a tokenizer to handle textual input. After the setup completion, Initialize the training and debugging which could be performed using a small subset of the data to fine tune the training loop and performance. Following that, the full training and validation stage will take place, utilizing the entire dataset and k-fold validation to ensure reliability. Model evaluation is conducted using F1 - score, accuracy score, confusion matrix. This leads to error analysis. This improved model is compared against other traditional models.

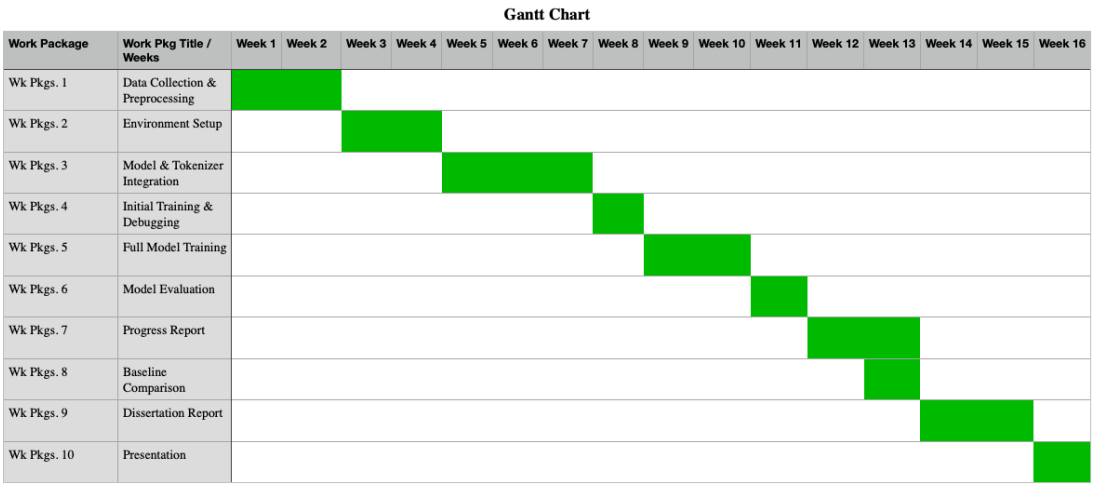

Figure 2: Gantt Chart

6 Conclusion

In an era of increasingly biased content, the ability to critically assess media content is more vital than ever. This project aims to address the urgent need to detect and reduce media bias by cognitive biases and fragmented journalism. By proposing a media bias detection system, the application enhances transparency, encourage accountability, and support the public in navigating complex news landscapes. An ethical obligation to the audience exists, necessitating that the tool honours diverse perspectives and sensitive contexts, thereby strengthening the foundations of journalism, fostering balanced narratives, and enabling the audience to engage with information meticulously. Media integrity is a societal imperative not just a technological challenge.

References

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Canli, H. and Toklu, S. (2021). Deep learning-based mobile application design for smart parking. *IEEE Access*, 9:61171–61183.

Evans, A. S., Moniz, H., and Coheur, L. (2024). A study on bias detection and classification in natural language processing. Preprint.

Han, H., Wang, W., and Mao, B. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, volume 3644 of *Lecture Notes in Computer Science*, pages 878–887. Springer.

Menzner, T. and Leidner, J. L. (2022). Experiments in news bias detection with pre-trained neural transformers. In *Proceedings of the Conference on Bias and Fairness in AI (BFAI)*, Coburg University of Applied Sciences and University of Sheffield. Contact: leidner@acm.org.

Nadeem, M. and Raza, S. (2021). Detecting bias in news articles using nlp models. In *Proceedings of the International Conference on Natural Language Processing (ICNLP)*. Conference paper.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Qin, B., Strömberg, D., and Wu, Y. (2018). Media bias in china. *American Economic Review*, 108(9):2442–2476.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint.

Recasens, M. and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659. Association for Computational Linguistics.

Rodrigo-Ginés, F.-J., Carrillo-de Albornoz, J., and Plaza, L. (2023). A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 233:121641. Open access under CC BY license.

SinaLab (2024). Biasfignews: A multilingual corpus of facebook posts annotated for bias and propaganda. https://github.com/SinaLab/BiasFigNews. GitHub repository.

Spinde, T., Hinterreiter, S., Haak, F., Ruas, T., Giese, H., Meuschke, N., and Gipp, B. (2023). The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. Technical report, University of Göttingen. arXiv preprint, DOI: 2312.16148.

Spinde, T., Krieger, J.-D., Ruas, T., Mitrović, J., Götz-Hahn, F., Aizawa, A., and Gipp, B. (2022). Exploiting transformer-based multitask learning for the detection of media bias in news articles. In *Proceedings of the iConference 2022*, Virtual event.

Spinde, T., Plank, M., Krieger, J.-D., Ruas, T., Gipp, B., and Aizawa, A. (2021a). 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, Dominican Republic. Association for Computational Linguistics.

Spinde, T., Rudnitckaia, L., Sinha, K., Hamborg, F., Gipp, B., and Donnay, K. (2021b). MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. In *Proceedings of the iConference 2021*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. arXiv preprint.

Wessel, M., Horych, T., Ruas, T., Aizawa, A., Gipp, B., and Spinde, T. (2023a). Introducing mbib - the first media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, New York, NY, USA. ACM.

Wessel, M., Horych, T., Ruas, T., Aizawa, A., Gipp, B., and Spinde, T. (2023b). Introducing MBIB – The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*.