Leveraging Large Language Model for Bias Detection in News Articles

Hirbod Gholamniaetakhsami
Linköping University / Linköping City
hirgh815@student.liu.se

Abstract

In this study, a machine learning approach is developed to detect and measure bias in news articles. The methodology involves fine-tuning a language model specifically for the task of bias classification. To ensure robustness, the model was also trained with and without oversampling method and it was tested in various random states. The model achieved a good precision of 84% and 83% for SG1 and SG2 datasets, respectively, ultimately aiming to minimize False Positive results in classifying biases.

Additionally, the model scored high F1 scores on both datasets compared to best baselines (63% on SG1 and 71% on SG2), ensuring the balance between precision and recall measures. The project categorizes bias into three types: distortion, content, and decision-making biases, and discusses their impact on media and democracy. The successful implementation of this project could lead to the development of tools that enable readers and journalists to better understand and check for biases, thereby contributing to a more informed and balanced public discourse.

The paper provides a comprehensive discussion of the models used, the data collected, the evaluation methods employed, and the results obtained, offering an alternative to automated bias detection in journalism.

1 Introduction

In the era of information overload, the ability to discern bias in news articles is more critical than ever. This research project aims to develop an innovative, machine learning-based approach to identify and quantify bias in news articles. The proposed methodology will involve using natural language processing techniques to analyze the linguistic patterns in the text, which often indicate a certain bias. The project contains two stages which are comprised of 1- fine-tuning a language model for the classification task of labeling the news article text and 2- evaluating the model through two methodologies.

This project, if successful, has significant implications for the field of journalism. It can provide a tool for readers to understand better the biases in the news they consume, and for journalists and editors to more effectively check their work for unintentional bias. Ultimately, this could contribute to a more informed and discerning public discourse. The next sections of this project are as follows: Section two provides the necessary context and theoretical foundation for the study. It includes a literature review highlighting relevant previous work in the field of news bias, particularly the studies involving automated tools for this purpose. It also outlines the motivation for this study and states the research problem and objectives. Section three details the data, preprocessing steps, and the process of fine-tuning the RoBERTa model. It also explains the implementation of strategies and the application of oversampling techniques. Section four presents the outcomes of the fine-tuning process and the performance of the model on the validation set. It includes detailed analysis and interpretation of the results, supported by appropriate statistical measures and visualizations. Any patterns, relationships, or trends observed in the results are highlighted and discussed. The final section wraps up the study by summarizing the key findings and their implications. It discusses the strengths and limitations of the current study, and how these findings contribute to the existing body of knowledge.

2 Background

Bias, whether intentional or unintentional, can significantly influence public opinion and discourse. Understanding the background and implications of this issue is crucial to this research. Essentially

Bias can be categorized in many forms (Mastrine, 2019), including:

- Unsubstantiated Claims: These are statements that are not supported by evidence.
- Opinion Statements Presented as Facts: This occurs when journalists present their personal opinions as if they are objective truths.
- Sensationalism/Emotionalism: This involves using exciting or shocking stories, languages, or visuals at the expense of accuracy to overhype an issue and generate public curiosity.
- Flawed Logic: This involves using faulty reasoning to make an argument.
- Commission: This involves permitting errors or false assumptions that support a specific point of view.

Each of these types of bias can influence how news is reported and how audiences perceive events and issues. Therefore consumers of news need to be aware of these biases to critically evaluate the information they receive. Entman (Entman, 2007) signifies the importance of bias in media. The author also emphasizes the importance of understanding these biases to comprehend how media influences the distribution of power and affects democracy. In this regard based on the content the type of bias that is delivered to the audience can be described as follows:

- Distortion Bias: This refers to news that allegedly distorts or falsifies reality.
- Content Bias: News that favors one side in a political conflict over another, rather than providing equivalent treatment to both sides.
- Decision-Making Bias: Biases related to the motivations and mindsets of journalists who produce the content.

The significant influence of media on society and the responsibility of avoiding biased content is undeniable. (Chen et al., 2020) presents a methodology for automatic political bias detection. The dataset utilized in this research contains a corpus of 6964 articles. The target of this analysis is the study of the distribution of bias and how it manifests at different levels of text granularity, from word to entire articles. The authors further reveal some common patterns of bias at various text levels, noting that the last part of an article tends to be the most biased.

Deductive Content Analysis is a common method used in qualitative studies to interpret and analyze data (Graneheim et al., 2017). Such social media models have been used for decades in the field of media bias analysis. (Hamborg et al., 2019) combines manual inspection methods from social sciences with natural language processing techniques. The authors introduced an automated identification of Media Bias by word choice and labeling in news articles. This approach involves extracting potential instances of bias, merging similar semantic concepts, and analyzing the framing of these instances to reveal bias. It is shown to achieve an F1 score of 45.7% which can be described as one of the best-performing models outside of machine learning approaches.

There are various research on how to detect media bias effectively. (Benson and Cruickshank, 2024) developed a method to cluster cable news programs based on their biases. By analyzing the topics discussed (using Named Entity Recognition) and how they are discussed (through Stance Analysis), programs with similar biases were grouped.

(Wu et al., 2022) proposed a novel to mitigate biases in evidence-based fake news detection. The causal intervention is used as a means to mitigate biases that are introduced during the data collection phase. As a model-agnostic method, it can be applied to various models for fake news detection. Additionally, the framework has shown promise in improving the robustness of augmented models.

(Hu et al., 2022) shows another application of causal inference on fake news detection. In this article, a novel framework(CLIMB) for multi-modal fake news detection was introduced. This approach addresses the problem of image-text matching bias in fake news detection. The task of fake news detection was formulated as a causal graph to reflect cause-effect factors. It has been shown that this model effectively improves fake news detection on real-world datasets.

(Arruda et al., 2020) proposed a tripartite model to analyze three types of bias: Selection bias, coverage bias, and statement bias. Assuming the bias is a deviation from the mainstream behavior, the authors introduce an outlier detection framework to gain insight into the existence and their nature in media outlets. Despite the inherent lim-

itation of their proposed methodology, it shows promise to identify specific biases. (Mansouri et al., 2020) discusses a semi-supervised learning method for detecting fake news in social media using a novel deep learning technique called SLD-CNN. This method combines convolutional neural networks(CNNs) with Linear Discriminant Analysis (LDA) for complex feature extraction and class separation respectively. As a method usable for both labeled and unlabeled data, this approach is especially useful in case of real-world application and data scarcity.

Subjective bias can significantly impact the validity and reliability of research findings, leading to distorted results and wrong conclusions. (Pryzant et al., 2020) aims to address the issue of subjective bias in texts that are expected to be objective such as news articles. The authors propose two encoder-decoder baseline algorithms for this task. In this paper, a human evaluation of four domains of encyclopedias, news headlines, books, and political speeches shows that both proposed algorithms are capable of reducing bias in texts.

To summarize the literature revolving around news bias detection and identification of fake news involves one or more than the following steps:

- Bias detection: To detect whether a news article is biased or not.
- Bias recognition: To recognize the biased words or phrases from the news articles.
- De-biasing: To de-bias the data by replacing the biased words or phrases from the news article with unbiased or at least less biased word(s).

(Hamborg et al., 2018) discusses the impact of media bias on public perception and the importance of unbiased news in shaping opinions. The authors review interdisciplinary approaches to analyzing media bias, combining social sciences and computer science methods with a focus on automated methods for identifying media bias in news articles, particularly using natural language processing (NLP). The next two review articles (Rakhecha et al., 2023; Rohera et al., 2022) provide two comprehensive surveys on news bias detection based on deep learning methods and fake news classification. The first paper reviewed a variety of deep learning models such as BERT and Long Short-Term Memory(LSTM) and Machine Learning algorithms such as logistic regression, While in the latter paper in addition to the introduction of models, authors also implemented models on a self aggregated dataset containing 6335 rows and four columns. According to the second paper, the highest accuracy belongs to the LSTM model which also shows the highest Recall. The primary reason for mentioning the task of fake news classification alongside the main focus of this project is that these tasks are highly similar and often show correlation:

1. Shared Objective: Both fields aim to assess the credibility and objectivity of news content. While fake news classification focuses on distinguishing between true and false information, news bias identification seeks to determine the presence of any partiality or prejudice in the news.
2. Similar Techniques: Both fields often employ similar computational and linguistic techniques for analysis. These include Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) algorithms to analyze text and identify patterns.
3. Interconnected Nature: The presence of bias can sometimes be a strong indicator of fake news. Biased news articles may distort facts or present misleading information, which is a characteristic of fake news. Therefore, identifying bias can be a crucial step in the process of fake news detection.

By providing a more objective measure of bias, this research could significantly enhance the transparency of news reporting and empower readers to make more informed judgments about the news they consume.

3 Methods

3.1 Data

The primary dataset utilized in this study is derived from the paper 'Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts' (Spinde et al., 2021). Despite the methodology used in the literature toward bias identification, it remains a challenging task due to its nature and lack of universal bias indicators. The BABE dataset is offered in two versions specified by the subgroups they were annotated:

1. SG1: 1700 sentences annotated by eight experts.

2. SG2: 3700 sentences annotated by five experts.

The features provided in the original dataset are as follows:

- **text**: The actual text of the sentence.
- **news link**: The link to the original news article from which the sentence was extracted.
- **outlet**: The media outlet that published the news article.
- **topic**: The topic of the news article.
- **type**: The type of bias (if any) present in the sentence.
- **label bias**: whether the sentence is biased or not.
- **label opinion**: The type of opinion expressed in the sentence, if any.
- **biased words**: The words in the sentence that show bias in the Python style list.

The distribution of bias in both SG1 and SG2 datasets can be seen in Figure 1. Both datasets appear to have a similar amount of labels belonging to both classes. The code for drawing the plots is available in the 'Datasetanalysis.ipynb' in the repository. Additionally, the count of label bias votes for each expert is given in Figure 2. In addition to these datasets, this project also contains scripts for extracting full-textual data from some news outlets. The data collection process for the evaluation phase of this project involved the extraction of textual data from news articles, specifically the title, highlight, and main body of each article. This was accomplished through the use of web scraping techniques, using the BeautifulSoup (Richardson, 2007) to parse the HTML content of the web pages.

During the data collection process, certain challenges were encountered. Some of the links to the news articles were invalid or required the use of web proxies, rendering them inaccessible for the purposes of web scraping. To address this issue, these links were either replaced by querying the text of the article through the DuckDuckGo search engine, or they were discarded entirely.

Despite these challenges, the web scraping script performed effectively, extracting data from

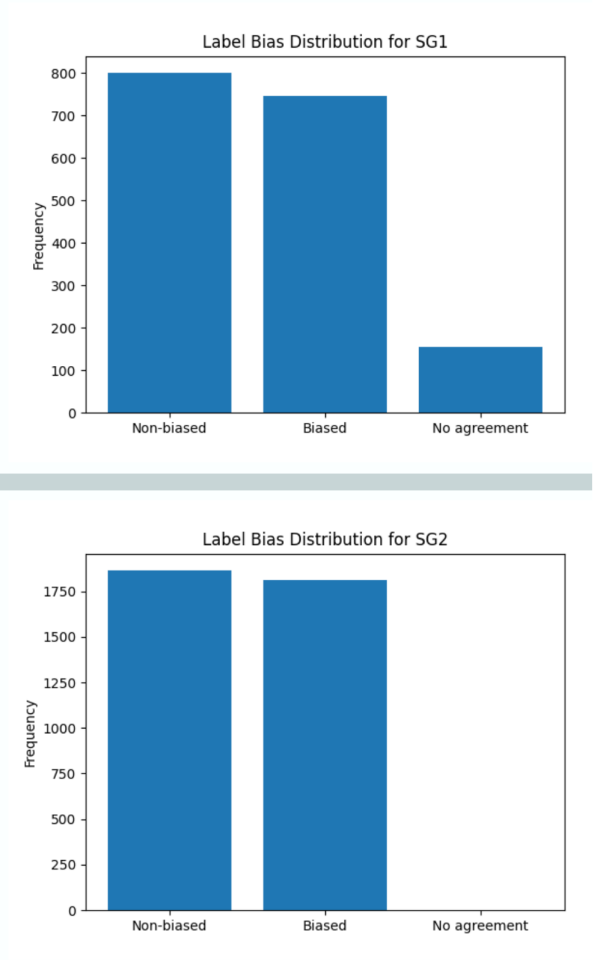

Figure 1: Distribution of Bias across datasets

the majority of the available articles across various news outlets. The performance ensured enough data to analyze the model's behavior and also for the future needs of possible extensions of this study.

Notice: The data extracted for this project is strictly used for academic purposes. It is used solely for the purpose of developing and evaluating machine learning models for media bias classification. No personal data is collected or used, and all data is handled in accordance with ethical guidelines and privacy standards.

3.2 Models

In this project, I employ a variety of models as a means of investigation for the main model. comprising three baseline models and a Large Language Model to perform text sequence classification. In the result section the models are compared according to the Accuracy, Precision, Recall, and F1 scores. The Baseline models used here include:

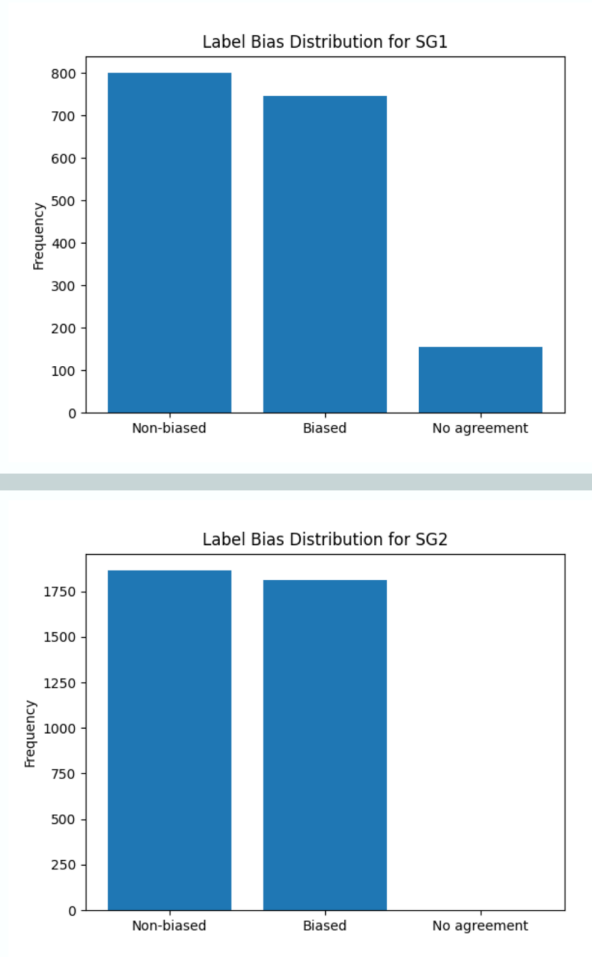

Figure 2: Individuals label bias tagging for SG1 and SG2 subgroups

- **Majority Class classifier**: This is a simple classifier that always predicts the most frequent label in the training set. It does not learn any information from the input features and in literature it is often used as a baseline to compare with other (real) classifiers. It's useful to provide a sanity check and to compare performance against the complex model.

- **Random Guesser**: This classifier generates predictions uniformly at random. The input features are not learned in this model as well. According to the dataset used in this project, using random predictions improves the results since the distribution of 'Biased' and 'Non-Biasd' are similar in this case.

- **LogisticRegression**: Logistic Regression is a statistical model that uses logistic function to model binary variables. published the news article. This method calculates the probability of belonging to a class by learning from input. As a linear model, it is an effective solution to provide a baseline prediction of classes in the model.

The main methodology in this project follows the implementation as in (Spinde et al., 2021). As the classifier RoBERTa (Liu et al., 2019) was implemented. RoBERTa is a variant of BERT which was trained on more data and for a longer amount of time. This model outperforms BERT and other state-of-the-art models on a variety of natural language processing tasks (a20, 2020). The model was trained and evaluated on a Google Colab environment in several epochs through checkpointing. The specific hardware used in this project includes a V100 VRAM 16GB GPU and The default CPU, an Intel Xeon CPU with 2 vCPUs (virtual CPUs).

3.3 Adding extra layers

As a transformer-based model excelling at various NLP tasks, RoBERTa can capture the complex structure of the task. During the implementation of the model, a few additional layers were added to improve the base model's performance. The aim of this step is to improve the capability of the model by improving the identification of local patterns(Conv1D) and capturing long-term dependencies(GRU). The customized model with additional layers can be observed in 9.

3.4 Evaluation

The evaluation of the model is a crucial aspect of this project, essentially providing a measure to show the goodness of the model's performance in bias identification tasks. The evaluation is divided into two subsections: 1- Traditional Evaluation Metrics and 2- Interpretive Model Evaluation.

3.4.1 Traditional Evaluation Metrics

In the first steps, the performance of models was tested using the most commonly used metrics. The metrics provide a comprehensive overview of performance while also considering both classes:

- **Accuracy**: It is the ratio of correctly predicted observations to the total observations. It provides an intuitive measure of the overall correctness of the model:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: This is the ratio of correctly predicted positive observations to the total predicted positive observations:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: Recall, or Sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual class. It provides a measure of the model's ability to find all the positive samples:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score**: It is the weighted average of Precision and Recall. It tries to find the balance between precision and recall and is particularly useful when dealing with imbalanced classes(The Classes in my dataset are mostly balanced):

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

to enhance the interpretation of the model, the text embedding is extracted from a BERT-style model $bert - tiny$ (Turc et al., 2019; Bhargava et al., 2021). For this purpose, [CLS] token was chosen, which is a special token used for classification tasks in BERT (Devlin et al., 2018). After the extraction of text embeddings representing sentences, the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique for visualizing high-dimensional data in the two-dimensional plot is employed (Com and Hinton, 2008; Arora et al., 2018). Using contextual embedding, the underlying structure is more insightful.

3.4.2 Interpretative Model Evaluation

In addition to the traditional metrics, a creative evaluation alternative is employed to provide a deeper understanding of the model's performance. This alternative involves the use of the Anchors method for model interpretation. Essentially, The Anchors method is a technique that can explain individual predictions of black-box classification by discovering a rule that sufficiently "anchors" the prediction locally – such that changes to the rest of the feature values of the instance do not matter (Ribeiro et al., 2018).

Interpretation of Model Predictions on Extracted News Articles The second strategy involves extracting some news articles from the news articles in the original data, some using web scraping and some manually. The Anchors method is then run with the model's predictions on these texts.

This strategy allows for a more detailed analysis of the model's decision-making process. By applying the Anchors method to individual sentences or sections within the articles, you can see which parts of the text the model is focusing on to make its predictions. If the anchors identified in these new articles are similar to those found in the original dataset, it suggests that the model is applying what it has learned to new data. Conversely, if the anchors are very different, it could indicate that the model is struggling to generalize.

4 Results

This section presents the results of the fine-tuning process applied to the RoBERTa language model for the task of media bias classification.

4.1 Model Fine-tuning

The RoBERTa model was fine-tuned using a train-validation-test approach. This method ensures that each fold of the dataset contains an equal proportion of each targeted class, providing a robust estimate of the model's performance.

4.2 investigating Class Imbalance

The two datasets used in this project did not exhibit class imbalance, This is evident by looking at the differences between the number of classes for dataset SG1 and SG2(54 and 53 respectively).

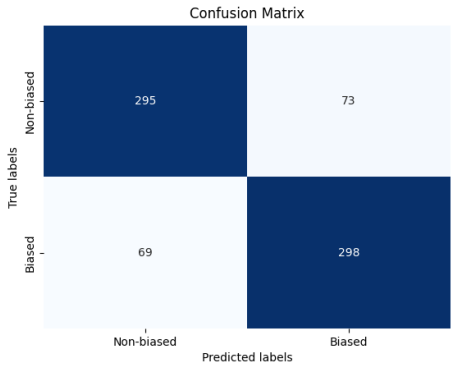

Figure 3: confusion matrix for SG1 dataset with RoBERTa after oversampling [value 1 indicate bias]

Oversampling involves randomly duplicating ex-

amples from the minority class in the training data to balance the class distribution. Theoretically oversampling ensures the model is exposed to enough examples of each class during training, improving its ability to generalize to under-represented classes.
However, due to the small difference in classes, oversampling did not manage to improve the classification and, at best managed to achieve a result similar to the original model. For the purpose of this comparison, some results are shown in 1.

4.3 Traditional Evaluation Metrics

Considering the difference in the number of sentences within each dataset, they are analyzed separately, here is a brief summary of results from the SG1 dataset:

- **Majority Class classifier**: It always predicts class 0. That's why the recall for class 0 is 1.00 (it correctly identifies all instances of class 0), but the recall for class 1 is 0.00 (it fails to identify any instances of class 1). The overall accuracy of 0.5387 indicates that approximately 54% of SG1 data belong to class 0.

- **Random Guesser**: The precision, recall, and F1-score are roughly equal for both classes, indicating that it's equally likely to guess either class. The overall accuracy of 0.5097(rounded at 4 digits) is close to 0.5, as you'd expect from random guessing.

- **LogisticRegression**: This model has higher precision, recall, and F1-score for both classes compared to the other two classifiers, indicating that it is doing a better job of identifying both classes. The overall accuracy of 0.6323 means it is correct about 63% of the time.

Table 1: Comparison of weighted macro F1-Score for the two datasets and the corresponding models

Dataset	Oversampling	Custom Model	F1 Score
SG1	False	True	0.8187
SG1	False	False	0.8049
SG1	True	True	0.8030
SG1	True	False	0.7935
SG2	False	True	0.8081
SG2	False	False	0.8148
SG2	True	True	0.7952
SG2	True	False	0.7971

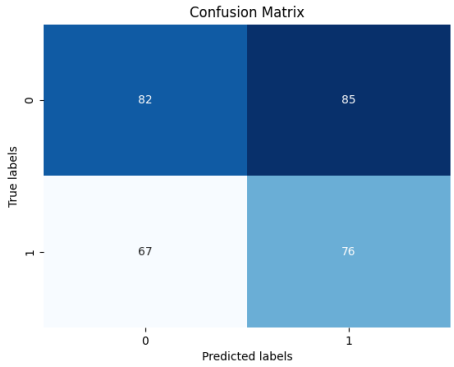

Figure 4: confusion matrix belonging to the random guesses SG1[value 1 indicate bias]

Table 2: Baselines weighted average results across datasets

Model name	Accuracy	Precision	Recall	F1	Dataset
Majority Class	0.5387	0.2902	0.5387	0.3772	SG1
Random Guesser	0.5097	0.5742	0.5097	0.5102	SG1
Logistic Regression	0.6323	0.6346	0.6323	0.6328	SG1
Majority Class	0.5116	0.2617	0.5116	0.3463	SG2
Random Guesser	0.5143	0.5143	0.5143	0.5143	SG2
Logistic Regression	0.7075	0.7082	0.7075	0.7075	SG2

In these numbers, weighted macro avg was selected as the default averaging scheme for the metrics(the full result is present in 'finetuning.ipynb'). Please note that the results of these algorithms do not necessarily indicate the distribution of labels in the corresponding dataset, this is mainly due to randomized train/test split before applying the models.

The metrics from the two datasets show a consistent pattern. Additionally, according to logistic regression results it can be observed that the increase in the number of data in each group has a substantial effect on learning from data. On the other hand, the language model shows a very promising result for both SG1 and SG2 datasets. Regarding the SG1 dataset and the best rates, the model exhibits an accuracy rate of 0.82 which is significantly better than all the baselines, similarly, a recall of 0.78 in the best SG2 model suggests that the model identifies 78% of all True positive instances.
To better understand the dataset, using the $bert - tiny$ model, the embeddings were extracted and visualized through t-SNE. Figure 5 and Figure 6 show the assigned labels to sentences by Logistic regression and RoBERTa model respectively.

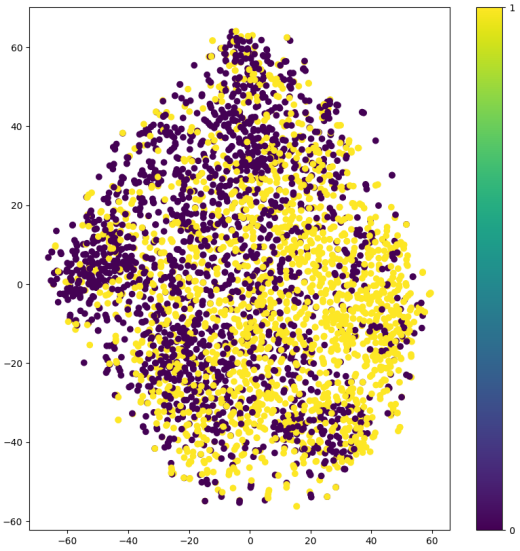

Figure 5: low dimensional representation of data with Logistic Regression labels[value 1 indicate bias]

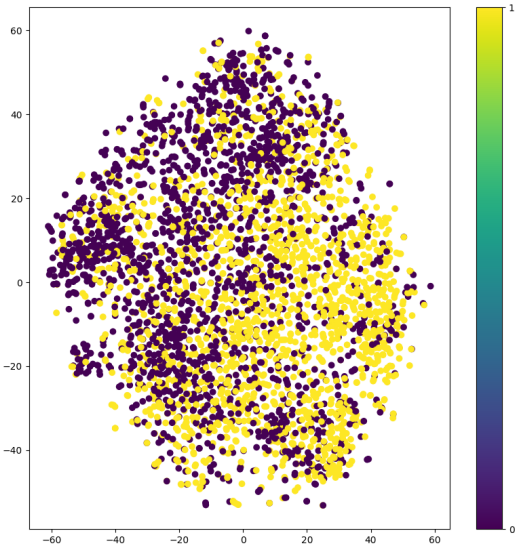

Figure 6: low dimensional representation of data with RoBERTa labels[value 1 indicate bias]

And if we divide the data into five regions we can see the distribution of misclassified biases are not exactly equal(Figure 7 and Figure 8)

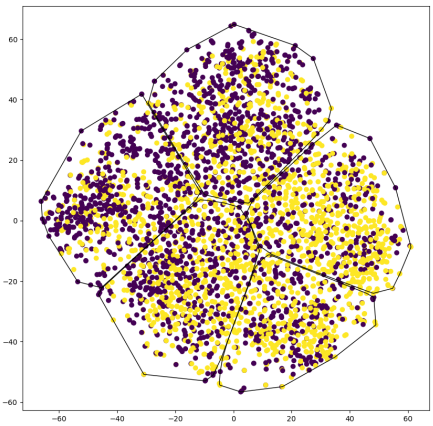

Figure 7: low dimensional representation of data with Logistic Regression labels with regions[value 1 indicate bias]

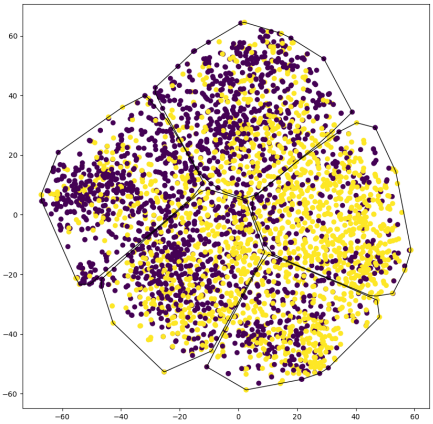

Figure 8: low dimensional representation of data with RoBERTa labels with regions[value 1 indicate bias]

According to table 3 we can see that region 5 has the least misclassified items, ten sentences from region are:

1. Sentence 1(TrueL 1): A young, slender woman with a sign reading “End All Violence” stood at the doors and pleaded with the looters to stop.

2. Sentence 2(TrueL 1): As Congress begins debating the Equality Act, the Heritage Foundation warns that if the bill becomes ...

3. Sentence 3(TrueL 1): Bloomberg also supports passage of the Equality Act, which

would bar governments and sports organizations from recognizing the biological/physiological difference between ...

4. Sentence 4(TrueL 0): Democrats and Republicans have long conflated “policing” with “public safety,” ...
5. Sentence 5(TrueL 0): Fake Political Media Eager Partners in Joe Biden’s Fake Press Conference
6. Sentence 6(TrueL 0): Fox News has gone so deep into white nationalism that Donald Trump now believes it’s how he’ll win in 2020
7. Sentence 7(TrueL 0): George Soros does not “own” ANTIFA or Black Lives Matter.
8. Sentence 8(TrueL 0): Graffiti alleging that the ‘MET POLICE SERVE TRUMP’ was also seen ...
9. Sentence 9(TrueL 0): In Haiti, President Bill Clinton admitted—after he left office—to a “devil’s bargain” on rice tariffs ...
10. Sentence 10(TrueL 1): In the dominant political fight in D.C., Donald Trump wants a giant wall along the U.S./Mexico border.

Table 3: Distribution of Misclassified objected for RoBERTa across regions(coordinates rounded to 2 digits)

Region	X axis range	Y axis range	Misclassification count
1	[3.01,59.51]	[-25.33,33.38]	48
2	[-49.55,7.97]	[-56.19,7.68]	54
3	[-63.08,-11.57]	[-24.42,43.03]	38
4	[-5.64,51.42]	[-57.66,-11.89]	31
5	[-29.60,37.11]	[4.86,63.54]	74

In this context we need to improve the rate of correctly predicted bias labels, therefore we have to maximize accuracy and precision to enhance the overall correctness and lower the false positive rate.

4.4 Interpretative Model Evaluation

In this section, the model is evaluated using a variety of text among the extracted data from news articles. Here are six chosen examples from the SG2 dataset and others extracted from news articles that demonstrate the model’s capabilities:

- ’Civil rights activists across the United States for years have called for scaling back traditional policing, and the efforts have gained momentum...’: This is an instance of Non-biased text, the model classifies it as Non-Biased, Anchors method chose ’called’, ’protests’, ’activists’, ’death’, ’scaling’, and ’momentum’ as the indicator for this decision
- ’An attorney for President Donald Trump on Friday blasted U.S. House Democrats’ request for six years of Trump’s tax returns as “a misguided attempt” to politicize...’: This is an instance of Biased text. The experts identified the keyword ’blasted’ as the indicator for bias, RoBERTa classified this sentence as Biased, and based on Anchors the decision is also based on the ’blasted’ keyword.
- ’Japan has reported about 1300 deaths from COVID-19 but new cases in Tokyo have fallen in the last few weeks’: This instance was extracted from a news article. Based on the model prediction this is Non-Biased and the indicator are the keywords ’fallen’, ’few’, and ’last’. This is likely true since it is a factual text and presents information about the trend of new cases in Japan.
- ’The idea of a ”rigged” economy that both benefits and is shaped by the rich and powerful was a popular theme in 2016 and is sure to play a big role again this cycle...’: This is a second instance extracted from news articles. The model predicted Non-Biased with Anchors indicators of big, idea, rich, theme, powerful, rigged, play, popular, campaign, ... After looking at the corresponding text, it is most likely a False Negative. This sentence offers a subjective viewpoint and uses the words ’Rigged’, ’Rich’, and ’powerful’ to imply a negative view of the corresponding groups.
- ’As you know, this is a very high priority for the president, ending surprise medical bills’: The model predicts this instance as Non-Biased and the anchor indicators are very, surprise, medical, ... This is factual information, so it is True Positive.
- ’The findings suggest childhood vaccination efforts nearly ground to a halt between March 13 — when the national emergency was declared — and April 19’: The model predicted this text as Non-biased with the Anchors keywords of ’halt’, ’efforts’, ’declared’, ’child-

Table 4: language model performance metrics according to the best models per datasets(rounded to two digits)

Loss	Accuracy	Precision	Recall	Micro F1	Weighted F1	Dataset
1.29	0.82	0.84	0.77	0.82	0.82	SG1
1.17	0.81	0.83	0.78	0.81	0.81	SG2

hood', 'between'. This proved to be a True sentence after a web search.

5 Conclusion and Discussion

This project aimed to classify media bias in news articles using a fine-tuned RoBERTa model and compare its performance with three baseline models. The results demonstrated the effectiveness of the RoBERTa model in this task greatly. As expected, the model performed better on the SG1 dataset due to a smaller overall amount of annotated sentences.

In addition to traditional evaluation metrics, an interpretative evaluation strategy was employed to provide a deeper understanding of the model's decision-making process. The Anchors method was used to identify the words that the model relied on most heavily to make its predictions, providing insight into what the model had learned and how it was making its decisions.

Regarding the limitations of this work, firstly, despite the effectiveness of the RoBERTa model in bias identification, the model the training process was computationally expensive due to the complexity of the model, potentially limiting its accessibility.

Secondly, for the sampling step of anchor, $en\_core\_web\_md$ model from spaCy was used. Larger models(such as $en\_core\_web\_lg$) have more parameters and potentially richer word vectors, which might lead to different perturbations being generated during the sampling process, this could be a potential direction for future improvements.

Lastly, the original labels annotated by experts were categorized as 'Biased', 'Non-Biased', and 'No agreement'. The third class indicates uncertainty or lack of consensus among experts, which could introduce additional noise; Therefore to improve consistency the labels from the third class were removed before the finetuning process. A more sophisticated approach would be to employ consensus-building techniques among experts to resolve disagreements before labeling.

Additionally, another research direction for this project would be to look more into the interpretation of the model's behavior including the use of other model-agnostic methods.

Overall, this project demonstrates the potential of transformer-based models like RoBERTa in media bias classification. It also highlights the importance of interpretability in machine learning, showing how methods like Anchors can prove valuable. The provided mode can help readers and journalists understand and check for biases, contributing to more informed public discourse.

Figure 9: Custom model with additional layers

References

2020. Overview of roberta model.

Sanjeev Arora, Wei Hu, and Pravesh K. Kothari. 2018. An analysis of the t-sne algorithm for data visualization.

Gabriel De Arruda, Norton Roman, and Ana Monteiro. 2020. Analysing bias in political news. *JUCS - Journal of Universal Computer Science*, 26:173–199.

Seth P Benson and Iain J Cruickshank. 2024. Developing a natural language understanding model to characterize cable news bias. *IEEE Access*, pages 1–1.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics. *arXiv (Cornell University)*, pages 125–135.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. *arXiv (Cornell University)*.

Lvdmaaten@gmail Com and Geoffrey Hinton. 2008. Visualizing data using t-sne laurens van der maaten. *Journal of Machine Learning Research*, 9:2579–2605.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Robert M. Entman. 2007. Framing bias: Media in the distribution of power. *Journal of Communication*, 57:163–173.

Ulla H. Graneheim, Britt-Marie Lindgren, and Berit Lundman. 2017. Methodological challenges in qualitative content analysis: A discussion paper. *Nurse Education Today*, 56:29–34.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2018. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20:391–415.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated identification of media bias by word choice and labeling in news articles. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Linmei Hu, Ziwei Chen, Ziwang Zhao Jianhua Yin, and Liqiang Nie. 2022. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Reza Mansouri, Mahmood Naderan-Tahan, and Mohammad Javad Rashti. 2020. A semi-supervised learning method for fake news detection in social media.

Julie Mastrine. 2019. How to spot 16 types of media bias.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:480–489.

Khushi Rakhecha, Simran Rauniyar, Muskan Agrawal, and Aruna Bhatt. 2023. A survey on bias detection in online news using deep learning.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Dhiren Rohera, Harshal Shethna, Keyur Patel, Urvish Thakker, Sudeep Tanwar, Rajesh Gupta, Wei-Chiang Hong, and Ravi Sharma. 2022. A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access*, 10:30367–30394.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with babe - bias annotations by experts.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv (Cornell University)*.

Junfei Wu, Qiang Liu, Wei Xu, and Shu Wu. 2022. Bias mitigation for evidence-aware fake news detection by causal intervention. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.