Comparative Methods for Detecting Bias in News Articles

Text Mining - Final assignment

Sadaf Esmaeili Rad
s.esmaeili.rad@umail.leidenuniv.nl
LIACS, Leiden University

ABSTRACT

Detecting bias in news articles is a challenging task due to the nuanced and subtle ways bias can be expressed at the sentence level. In this paper we evaluate the performance of several transformer-based models, including BERT, RoBERTa, DistilBERT, and DeBERTa, in addition to a state-of-the-art machine learning model, Support Vector Machine (SVM), for the task of sentence-level bias detection. The performance of the models is tested on a dataset specifically designed for this purpose, using preprocessing techniques such as text cleaning, and oversampling. We provide a comparative analysis of different models for sentence-level bias detection.

Our results reveal that some transformer models outperform others, with DistilBERT achieving the best overall performance by balancing computational efficiency with high precision and recall. Interestingly, a traditional machine learning approach like SVM can perform better than certain types of transformer models. These findings highlight the effectiveness of transformer models for bias detection while also showing the potential of traditional machine learning methods when combined with proper pre-processing.

KEYWORDS

Bias detection, Bias in news articles, Bert, DistilBERT, RoBERTa, DeBERTa, SVM , Transformers.

1 INTRODUCTION

News articles are not always written in a completely neutral way. They often reflect bias through specific word choices, writing styles, or even the author's personal opinions. This type of bias, known as news bias, can influence how readers understand social issues, form political opinions, and shape their overall attitudes toward certain topics. Because of this, identifying and addressing news bias is important to maintain fairness, accuracy, and trust in media reporting. However, detecting bias in news articles is not an easy task. Bias often appears in very subtle ways, like using slightly different words or phrases that carry different meanings. For example, referring to "illegal immigrants" versus "undocumented immigrants" or using "climate change" instead of "global warming" can subtly influence how readers interpret an article. On top of this, there aren't many datasets available that focus on detecting bias at the sentence level. Most datasets label bias at the document level or for entire news outlets, which misses the subtle sentence-level details that can shape a reader's opinion. Bhatia et al. [1]

To address these challenges, our research focuses on evaluating different methods for detecting bias in news articles at the sentence level. Specifically, we wanted to investigate the following research question:

"How do different preprocessing strategies and transformer-based models compare in detecting biased sentences in news articles, and how do these methods perform compared to traditional machine learning techniques like SVM when dealing with imbalanced datasets?"

This research question is important because detecting sentence-level bias requires advanced techniques that can understand the context of individual sentences. To answer it, we used a dataset specifically designed for bias detection at the sentence level. Our approach involved applying various preprocessing techniques, such as tokenization and oversampling, and comparing the performance of several transformer-based models, including BERT, DistilBERT, RoBERTa, and DeBERTa. We also trained a Support Vector Machine (SVM) as a baseline for comparison. By analyzing these methods, we aim to understand which approaches work best for detecting bias in news articles, particularly in the context of imbalanced datasets.

2 RELATED WORK

Transformer-based models are now the go-to choice for detecting complex language patterns, including subtle forms of bias in text. These models often outperform traditional methods like Support Vector Machines (SVM), mainly because they capture context at a deeper level. For example, Hugging Face Blog [2] showed that BERT can identify signs of bias that SVMs might miss, underlining the benefits of large-scale pre-training and contextual embeddings. Expanding on BERT's success, Mdpi Electronics [3] investigated bias detection alongside cyberbullying detection. They used DeBERTa, RoBERTa, and DistilBERT and showed that balancing the dataset by creating artificial samples leads to better performance on less common classes. This highlights the importance of data preprocessing and careful handling of class imbalances, which are common in bias detection tasks.

Meanwhile, Springer Authors [5] demonstrated that DistilBERT, a smaller and faster transformer model, can still achieve strong performance while using fewer resources. This makes it an appealing option when efficiency is a priority. In contrast, Vallejo and Others [6] compared BERT, RoBERTa, and DeBERTa for political text analysis, finding that RoBERTa and DeBERTa often outperform BERT due to improved designs and training methods. Their results show that even closely related transformer architectures can vary significantly in how they capture nuanced language patterns related to bias.

Most previous research has focused on bias at the document level, which can overlook how small word choices shape reader opinions. Sora [4] addressed this gap by creating a dataset for detecting bias at the sentence level, making it easier to pinpoint which phrases or terms cause readers to perceive bias. Building on these insights, our work uses the dataset from Vallejo and Others [6] to compare several transformer models—BERT, RoBERTa, DeBERTa, and DistilBERT—along with an SVM model. By applying oversampling and other preprocessing techniques, we aim to find the best approach for detecting sentence-level bias, even when the dataset is imbalanced.

3 METHODS

In this section, we describe the steps taken to detect bias in news articles, including the preprocessing techniques applied to the dataset, the selection of models, and the methodology for training and evaluation.

3.1 Preprocessing

We started with preprocessing to ensure the dataset was ready for training the models. Here are the steps we followed and the reasons behind them:

Data Cleaning : The dataset was already well-organized, but we checked for missing values or inconsistencies. While some columns had missing entries, the key ones—sentence text and bias labels—were complete, so we could use the data directly without modifications. Ensuring clean data was important to avoid errors during model training.

Label Mapping : The dataset contained four labels representing different levels of bias: slightly biased, biased, neutral or unbiased, and very biased. We decided to keep all four labels because our goal was to perform fine-grained classification. Fine-grained classification allows us to identify not just whether a sentence is biased or unbiased, but also how strong the bias is. This level of detail can provide deeper insights into how bias appears in news articles.

Handling Class Imbalance with Oversampling : One of the main challenges with this dataset was the imbalance in label distribution. Most sentences were labeled as neutral/unbiased, while very few were labeled as slightly biased or very biased. To address this, we used oversampling. By duplicating sentences from the minority classes during training, we gave the models more examples of these labels. This helped the models learn better and improved their ability to classify the underrepresented categories.

Stopword Removal: To enhance the performance of the Support Vector Machine (SVM) model, we removed the stopwords. Stopwords, such as "the," "is," and "of," often do not contribute significantly to the semantic meaning of the text but can dominate the vector space in traditional machine learning models like SVM. By removing these common words, we reduced noise in the data and improved the model's ability to focus on meaningful terms that are more indicative of bias. This preprocessing step improved the SVM performance. For instance, the test accuracy of SVM was improved by 1.78% , the test recall improved by 2.00% and the test precision improved by 5.00% after stopword removal.

Splitting the Dataset : We divided the dataset into three parts: 70% for training, 15% for validation, and 15% for testing. Stratified splitting was used to maintain the same label proportions in each set as in the original dataset. This ensured that the validation and test sets were representative of the data the models would encounter in real scenarios.

Preparing the Data for Models : For transformer models, we used their respective tokenizers, which break sentences into pieces that the models can understand. For the SVM, we represented the sentences numerically using TF-IDF. This allowed us to use both traditional and advanced methods on the same dataset for comparison.

3.2 Challenges with Imbalanced Data

One of the main challenges in this task was the imbalance in the dataset. Most sentences were labeled as neutral or unbiased, while fewer were slightly biased or very biased. This imbalance sometimes led to situations where models performed better on the test set than the validation set. Oversampling helped address this issue, but the models occasionally overfitted to the minority classes during training, leading to unexpected variations in performance across the splits. The label distribution for the dataset is shown in Figure 1.

3.3 Model Selection

To evaluate different approaches for detecting bias in news articles, we selected four transformer-based models and one traditional machine learning model. Our goal was to test a variety of methods to see which worked best for this task. Each model was chosen for specific reasons, and we were particularly interested in how they handled the imbalanced dataset and performed on the minority classes.

BERT: BERT was chosen as it is one of the most widely used transformer models for text classification. It is known for its ability to capture contextual relationships in text, which we expected to be useful for identifying subtle biases in news sentences.

DistilBERT: We included DistilBERT as a lighter version of BERT. Since it requires less computational power, we wanted to see if it could still achieve good results while being more efficient.

RoBERTa: RoBERTa was selected because it builds on BERT but is optimized for better performance by removing some of BERT's limitations. We expected RoBERTa to perform well due to these improvements.

DeBERTa: DeBERTa represents the latest advancements in transformer architectures. It has been shown to perform exceptionally well on various NLP tasks. We chose it to test whether its advanced design could better handle the complexity of our dataset.

SVM: We included Support Vector Machine (SVM) as a traditional machine learning baseline. SVMs are known to perform well on small datasets, and comparing its results with transformer models allowed us to understand how much newer models improve over older methods.

4 TRAINING PARAMETERS AND THEIR IMPACT

Training parameters significantly influence the performance of transformer-based models. Among these, the learning rate, number of epochs, and batch size are critical for achieving optimal results. Below, we explain the rationale behind the parameter choices for our models and their observed effects on performance.

4.1 Learning Rate

The learning rate is one of the most important hyperparameters, determining how much the model adjusts its weights during training. A lower learning rate (2e-5) ensures small, steady updates, which can be advantageous for complex models like DeBERTa and RoBERTa. These models are pre-trained with advanced optimization techniques, requiring finer adjustments to perform well on specific tasks like bias detection.

On the other hand, models like BERT and DistilBERT used a higher learning rate (5e-5), allowing for faster convergence. Interestingly, despite having a higher learning rate, DistilBERT performed better overall, likely due to its lighter architecture and reduced parameters, which make it easier to optimize compared to larger models like DeBERTa. The higher learning rate worked well for DistilBERT by quickly finding a good solution without the need for finer adjustments.

4.2 Number of Epochs

All models were trained for 5 epochs, which balanced underfitting and overfitting. This choice ensured that the models had enough time to learn patterns in the data while avoiding overfitting, especially given the small dataset size and the use of early stopping based on validation metrics.

4.3 Batch Size

Batch size influences both training stability and computational efficiency. For DeBERTa, we used a smaller batch size of 8 because of its memory-intensive nature. For BERT, DistilBERT, and RoBERTa, a batch size of 16 was feasible, enabling faster training without memory constraints.

4.4 Observations and Insights

Despite using different learning rates, DistilBERT outperformed the other models, suggesting that its simpler architecture and faster convergence made it better suited for this specific task. Models with lower learning rates (DeBERTa and RoBERTa) required more precise updates but struggled to generalize as effectively, particularly in this dataset. These results highlight that the effectiveness of parameters like learning rate depends not only on the dataset but also on the architecture of the model.

Table 1: Training Parameters for Transformer Models

Parameter	BERT	DeBERTa	DistilBERT	RoBERTa
Learning Rate	5e-5	2e-5	5e-5	2e-5
Epochs	5	5	5	5
Batch Size	16	8	16	16

4.5 Evaluation Metrics

To compare the models fairly, we used several evaluation metrics:

- **Accuracy:** This gave us an overall measure of how many sentences were classified correctly.
- **Precision:** This measured how many sentences predicted as biased were actually biased. Precision was important because we wanted the model to avoid labeling neutral sentences as biased.
- **Recall:** This showed how many biased sentences were correctly identified. It was crucial for ensuring that the minority classes were not overlooked.
- **F1-Score:** We focused on F1-score as it balances precision and recall, making it particularly useful for evaluating imbalanced datasets.

The combination of these metrics helped us understand the strengths and weaknesses of each model. While accuracy provided a general idea of performance, F1-score gave us deeper insights into how well the models performed on the minority classes, which was a key challenge in this task.

5 DATA

The dataset used in this study was introduced by Bhatia et al. [1] in 2020 and was specifically created to detect bias in news articles at the sentence level. It contains 215 sentences, each labeled with one of four levels of bias: slightly biased, biased, neutral or unbiased, and very biased. These labels provide a detailed framework for analyzing bias at the sentence level, which is essential for understanding the subtle ways bias can appear in news reporting.

5.1 Sample Size and Label Distribution

The total number of sentences in the dataset is **215**. The distribution of labels is shown in Table 2.

Table 2: Label Distribution in the Dataset

Label	Description	Count	Percentage
3	Neutral/Unbiased	92	42.79%
2	Biased	72	33.49%
1	Slightly Biased	32	14.88%
4	Very Biased	19	8.84%

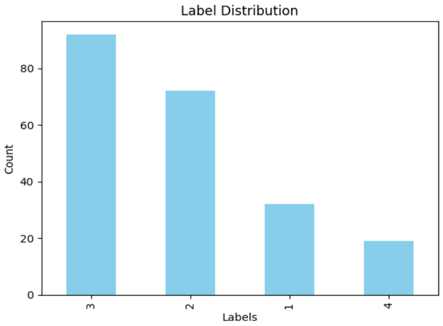

Figure 1: Label Distribution of the Dataset

The label distribution shows that the majority of sentences are labeled as neutral/unbiased, while "very biased" sentences are relatively rare. This imbalance in the dataset presents a challenge for machine learning models, as they may struggle to accurately classify the minority classes.

5.2 Dataset Overview

The dataset includes 46 English-language news articles covering diverse topics such as political scandals, international relations, and social movements. Metadata provided in the dataset includes:

- **Article metadata:** Information such as source, URL, publication date, and article-level bias labels.
- **Sentence content:** Individual sentences with their associated bias labels for fine-grained analysis.

5.3 Missing Data

Although the dataset is well-structured, some columns contain missing values. For example:

- Columns related to specific sentence metadata, like: `s6` to `s19` have varying degrees of missing values.
- However, the main columns relevant to this study, such as sentence text (docbody) and bias labels (`article_bias`), do not have any missing values.

The dataset's structure and annotation provide a solid foundation for analyzing bias at the sentence level, enabling us to evaluate different preprocessing strategies and machine learning models effectively.

6 EXPERIMENTS

6.1 Bert

The results of the BERT model show that it struggled with the task of detecting bias in news articles. The performance was generally low on both the validation and test sets, with accuracy only reaching **30.30%** on the test set and F1-score at **0.2681**. This suggests that the model had difficulty distinguishing between the different levels of bias, which could be due to the subtle nature of the bias in the dataset.

Interestingly, the test set results slightly outperformed the validation set in terms of accuracy and F1-score. This could be because the test set had slightly more examples that matched patterns the model learned during training. However, the difference is not large enough to be considered significant, and it may also indicate some level of randomness or overfitting to the training data.

The most notable issue is the low recall, which means the model missed many biased sentences, especially in the minority classes. This highlights a key challenge with this task: the imbalanced dataset makes it hard for the model to learn patterns for the less frequent classes. Even with oversampling, BERT struggled to generalize well to these underrepresented classes.

Table 3: Results of BERT Model on Validation and Test Sets

Metric	Validation Set	Test Set
Loss	1.4604	1.4139
Accuracy	0.21875	0.3030
F1-Score	0.2210	0.2681
Precision	0.3903	0.3460
Recall	0.2188	0.3030

6.2 DistilBERT

As we already mentioned, we included DistilBERT as a lighter version of BERT. Since it requires less computational power, we wanted to see if it could still achieve good results. Interestingly, the results of DistilBERT show a noticeable improvement compared to BERT, especially in terms of accuracy and precision. On the validation set, the model achieved an accuracy of **50%**, which is significantly better than BERT's performance. The F1-score on the validation set was **0.4063**, indicating that DistilBERT was better at balancing precision and recall during evaluation.

DistilBERT also had higher precision on both the validation and test sets compared to other metrics. For example, the test set precision was **76.74%**, which suggests that the model was good at correctly identifying biased sentences and avoiding false positives. However, this came at the cost of a slightly lower recall, meaning it missed some biased sentences, especially in the minority classes.

The test set accuracy dropped slightly to **48.48%**, which is close to the validation set accuracy. This consistency suggests that the model generalized better than BERT, even though the F1-score decreased slightly to **0.3665** on the test set.

Table 4: Results of DistilBERT Model on Validation and Test Sets

Metric	Validation Set	Test Set
Loss	1.2111	1.1992
Accuracy	0.5000	0.4848
F1-Score	0.4063	0.3665
Precision	0.6356	0.7674
Recall	0.5000	0.4848

6.3 DeBERTa

The DeBERTa model showed slightly better performance than BERT but fell short compared to DistilBERT in some aspects. The accuracy on the validation set was **43.75%**, which is an improvement over BERT but not a significant leap. On the test set, the accuracy dropped to **42.42%**, which suggests that the model struggled to generalize effectively across datasets.

One standout aspect of DeBERTa's performance is its **high precision**, reaching over **75%** on both validation and test sets. This means the model was very confident in the sentences it classified as biased and was rarely wrong in doing so. However, its recall was relatively low, indicating that it failed to identify a large portion of biased sentences, especially from the minority classes.

The F1-scores were lower compared to DistilBERT, with **0.2663** on the validation set and **0.2527** on the test set. This reflects an imbalance between precision and recall, where the model excelled at avoiding false positives but struggled to capture all instances of bias.

6.4 RoBERTa

RoBERTa performed better than BERT in all metrics but was consistently outperformed by DistilBERT. Specifically, RoBERTa achieved a test accuracy of **37.50%**, which is higher than BERT's **30.30%** but lower than DistilBERT's **48.48%**. This shows that RoBERTa's

Table 5: Results of DeBERTa Model on Validation and Test Sets

Metric	Validation Set	Test Set
Loss	1.2528	1.2632
Accuracy	0.4375	0.4242
F1-Score	0.2663	0.2527
Precision	0.7539	0.7557
Recall	0.4375	0.4242

advanced design helps it do better than BERT but not as well as DistilBERT for this task. In terms of F1-score, RoBERTa achieved **35.69%** on the test set, which is a significant improvement over BERT's **26.81%**. RoBERTa's precision and recall are balanced, but its lower F1-scores and accuracy compared to DistilBERT shows room for improvement.

Overall, RoBERTa demonstrated clear improvements over BERT, highlighting the benefits of its enhanced architecture. However, it could not outperform DistilBERT. The performance of the RoBERTa model is summarized in Table 6.

Table 6: Results of RoBERTa Model on Validation and Test Sets

Metric	Validation Set	Test Set
Loss	1.2699	1.3172
Accuracy	0.4000	0.3750
F1-Score	0.3683	0.3569
Precision	0.4707	0.3990
Recall	0.4000	0.3750

6.5 SVM

The performance of the SVM model demonstrates that traditional machine learning techniques can still be competitive in certain scenarios. SVM outperformed RoBERTa on both validation and test sets and slightly outperformed DistilBERT in test accuracy. Specifically, SVM achieved a test accuracy of **53.57%**, compared to RoBERTa's **37.50%** and DistilBERT's **48.48%**. This indicates that SVM, despite relying on simpler feature representations like TF-IDF, can effectively classify biased text when the data is well-preprocessed.

However, SVM underperformed in precision and recall compared to DistilBERT. For example, while SVM achieved a test precision of **58.00%**, DistilBERT's precision was significantly higher at **76.74%**. Similarly, DistilBERT's recall of **48.48%** slightly exceeded SVM's recall of **54.00%**, proving that transformer models handle complex language details effectively.

Overall, SVM displayed a better balance between precision and recall than RoBERTa but was less effective in achieving the higher precision seen with DistilBERT. The performance of the SVM model is summarized in Table 7.

Table 7: Results of SVM Model on Validation and Test Sets

Metric	Validation Set	Test Set
Accuracy	0.4909	0.5357
F1-Score	0.4500	0.5200
Precision	0.4600	0.5800
Recall	0.4900	0.5400

7 DISCUSSION

The results of our experiments show the varying strengths and weaknesses of transformer-based models and state-of-the-art machine learning models for sentence-level bias detection. In this section, we interpret the results of the models that we implemented for this task.

7.1 Transformer Models

Transformer-based models, particularly DistilBERT and DeBERTa, demonstrated strong performance due to their ability to capture nuanced contextual relationships in text. DistilBERT's lightweight architecture allows a balance between computational efficiency and high precision and recall. On the other hand, DeBERTa's high precision makes it appropriate for applications where the goal is to minimize false positives.

BERT, which was our baseline for transformer models in this study, did not perform as well as its successors. This underperformance was expected considering that newer models like RoBERTa and DeBERTa are designed to address specific limitations in BERT, such as better pre-training strategies.

7.2 SVM

The SVM model, despite its simplicity, demonstrated competitive performance, particularly in test accuracy. This suggests that traditional machine learning models, when combined with appropriate preprocessing techniques such as oversampling and TF-IDF , can still provide viable results in certain contexts. On the other hand, SVM's lower precision and recall compared to some of the transformer models, shows limitations in its ability to capture subtle patterns in language.

7.3 Impact of Preprocessing

Preprocessing played a crucial role in improving the performance of all models in this study. Techniques such as text cleaning and TF-IDF feature extraction for SVM ensured that the data was well-prepared for training, improving the results of the models.

Oversampling had a significant impact on the performance of the models. By creating additional examples for underrepresented classes, oversampling helped increase performance metrics for all models. This was particularly noticeable for traditional machine learning models like SVM, which rely on balanced data for effective learning.

Among the transformer-based models, oversampling also led to improved performance, particularly in metrics like F1-score and recall. For example, DistilBERT and RoBERTa showed more balance between precision and recall on the test set after oversampling. These findings show the importance of robust preprocessing and data augmentation strategies when working with imbalanced datasets.

7.4 Insights

The results of the experiments show a trade-off between model complexity, computational efficiency, and task performance. Transformer-based models can handle nuanced contextual relationships but require more resources for training. On the other hand, traditional models like SVM offer a simpler and computationally cheaper alternative that can still perform well with robust preprocessing.

The comparative analysis in this study also highlights the importance of dataset characteristics and preprocessing strategies.

8 CONCLUSION

Detecting bias in news articles, especially at the sentence level, is a difficult but important task. Bias often appears subtly, making it hard for models to identify without advanced techniques. In this study, we compared transformer-based models with a traditional machine learning method to see which performed best at identifying bias. We also applied preprocessing techniques, like oversampling, to handle the imbalance in the dataset.

Our results showed that DistilBERT performed the best overall, balancing accuracy, precision, and recall, while also being more efficient than other transformers. BERT, on the other hand, struggled to perform well on this task, likely due to its older architecture. DeBERTa and RoBERTa showed strong precision but lower recall, meaning they were good at avoiding false positives but missed many biased sentences. Surprisingly, SVM, despite being a simpler model, performed competitively, especially in terms of test accuracy, showing that traditional methods can still be effective with proper preprocessing.

One major challenge across all models was the low recall for minority classes, like "slightly biased" and "very biased." This shows how difficult it is to detect subtle biases, even with powerful models. Oversampling helped improve performance for these underrepresented classes, but it didn't fully solve the problem.

For future work, larger and more diverse datasets could help models perform better and generalize across different topics and languages. Improved preprocessing techniques, such as creating synthetic data or using adversarial training, could address class imbalance more effectively. Adding explainable AI methods would also make these models easier to understand and more reliable for real-world use. Exploring tasks related to bias detection, like stance or fake news detection, could also strengthen the models' abilities.

Our study highlights the strengths and weaknesses of current approaches to bias detection and provides a foundation for further improvements. By building on these findings, future research can contribute to fairer and more transparent news reporting.

REFERENCES

[1] N. Bhatia, S. Flekova, S. Rajendran, and I. Gurevych. Learning to recognize biased text. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1562–1571, May 2020. Available at: https://aclanthology.org/2020.lrec-1.184/.

[2] Hugging Face Blog. Bert for bias detection in text. *Hugging Face*, 2023. Available at https://huggingface.co/blog/maximuspowers/bias-detection-in-text.

[3] Mdpi Electronics. Bias and cyberbullying detection and data generation using transformer models. *MDPI Electronics*, 13:3431, 2023. doi: 10.3390/electronics13173431. Available at https://www.mdpi.com/2079-9292/13/17/3431.

[4] Sora. Sora lrec 2020 biased sentences dataset. https://github.com/skymoonlight/biased-sents-annotation/blob/master/Sora_LREC2020_biasedsentences.csv, 2020. Accessed: January 2025.

[5] Springer Authors. Detecting bias in university news articles: A comparative study using distilbert. In *Advances in Artificial Intelligence and Data Engineering*, pages 42–56. Springer, 2023. doi: 10.1007/978-3-031-47994-6_42. Available at https://link.springer.com/chapter/10.1007/978-3-031-47994-6_42.

[6] Samuel Vallejo and Others. Bert, roberta, or deberta? comparing performance across transformers models in political text analysis. *Journal of Political Science*, 2022. Available at https://svallejovera.github.io/files/bert_roberta_jop.pdf.