True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning

Chahat Raj*
George Mason University
Fairfax, Virginia, USA
craj@gmu.edu

Anjishnu Mukherjee*
George Mason University
Fairfax, Virginia, USA
amukher6@gmu.edu

Ziwei Zhu
George Mason University
Fairfax, Virginia, USA
zzhu20@gmu.edu

ABSTRACT

The dissemination of information, and consequently, misinformation, occurs at an unprecedented speed, making it increasingly difficult to discern the credibility of rapidly circulating news. Advanced large-scale language models have facilitated the development of classifiers capable of effectively identifying misinformation. Nevertheless, these models are intrinsically susceptible to biases that may be introduced through numerous ways, including contaminated data sources or unfair training methodologies. When trained on biased data, machine learning models may inadvertently learn and reinforce these biases, leading to reduced generalization performance. This situation consequently results in an inherent "unfairness" within the system. Interpretability, referring to the ability to understand and explain the decision-making process of a model, can be used as a tool to explain these biases. Our research aims to identify the root causes of these biases in fake news detection and mitigate their presence using interpretability. We also perform inference time attacks to fairness to validate robustness.

KEYWORDS

misinformation, bias, fairness, interpretability, security

ACM Reference Format:
Chahat Raj, Anjishnu Mukherjee, and Ziwei Zhu. 2023. True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3600211.3604760

1 INTRODUCTION

Machine learning classifiers consistently exhibit discriminatory tendencies, favoring one demographic group over another across various domains based on specific characteristics. In the context of news, political leaning represents one notable characteristic wherein biases have been observed and documented. Such bias may deteriorate public trust and exacerbate political polarization [3]. Given the potential for bias in the news related to political leaning and the severe implications this can have, it becomes crucial to understand the decision-making process of these black-box models. We also want to see if fake news detection techniques carry any biases. However, it is yet unclear what ideal measures should be used to evaluate fairness realistically. Interpretability is valuable in determining whether a model has genuinely acquired knowledge or is merely producing predictions through random guessing. We aim to identify the most crucial information that language models utilize for classifying fake news. Hence, we propose the following research questions:

RQ1: What dimensions of fairness should be considered to evaluate the performance of language models in fake news detection?
RQ2: Do existing language models demonstrate bias in detecting fake news across different political ideologies?
RQ3: Can integrating interpretability techniques in misinformation detection aid in identifying and mitigating the sources of bias?

2 PROPOSED RESEARCH

Experiment Settings: We utilize the NELA-GT-2018 dataset [2], comprising news articles from various fact-checking sources. The original dataset includes 713k news articles labeled with source-level credibility and political leaning indicators. We rely on credibility labels provided by NewsGuard and political leaning labels provided by BuzzFeed. We exclude articles lacking labels from both NewsGuard and BuzzFeed, resulting in 163k articles.

The experiments [1] are conducted using a fine-tuned DistilBERT, which, according to our preliminary investigations, outperforms the original BERT in terms of key performance indicators such as accuracy and F1 score. Existing work [3] employs traditional machine learning classifiers, with Random Forest demonstrating the highest overall accuracy of 87.87%. Our approach results in a new state-of-the-art accuracy of 91.36%. Additional relevant metrics are presented in Table 1. To our knowledge, this represents the first reported results on this dataset using a transformer-based language model.

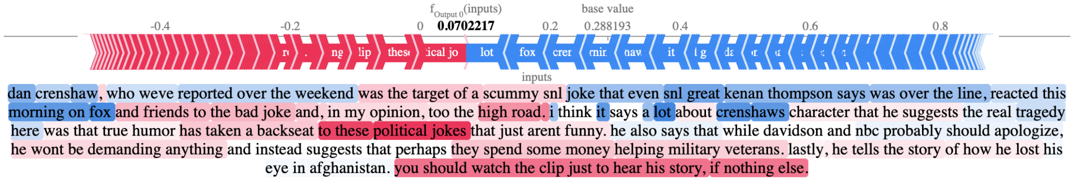
Fairness Formalization: We extend the scope of fairness assessment beyond the conventionally used metrics, Statistical Parity Difference (SPD) and Disparate Impact Ratio (DIR). We incorporate two additional metrics, Equal Opportunity Difference (EOD) and Average Odds Difference (AOD), to comprehensively evaluate algorithmic fairness. Moreover, we highlight the underlying bias by contrasting the precision and recall scores between the privileged and unprivileged groups. Additionally, we highlight the discrepancies in precision and recall scores, broken down by categories of real and fake news (Table 1). We discover significant biases manifested through these category-specific differences in precision and

[*] Both authors contributed equally to this research.

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604760

[1] Code and data are available at https://github.com/chahatraj/true-and-fair



dan crenshaw, who weve reported over the weekend was the target of a scummy snl joke that even snl great kenan thompson says was over the line, reacted this morning on fox and friends to the bad joke and, in my opinion, too the high road. i think it says a lot about crenshaws character that he suggests the real tragedy here was that true humor has taken a backseat to these political jokes that just arent funny. he also says that while davidson and nbc probably should apologize, he wont be demanding anything and instead suggests that perhaps they spend some money helping military veterans. lastly, he tells the story of how he lost his eye in afghanistan. you should watch the clip just to hear his story, if nothing else.

Figure 1: An interpretability example using SHAP depicting salience of terms across a misclassified news item

recall, biases which are often overlooked when using conventional fairness measures such as SPD and DIR.

Model Interpretation: In the next phase, our primary objective is to investigate the most influential tokens in BERT's decision-making process. The underlying hypothesis is that the traditional salience approaches, such as SHAP [1] and LIME [4], in addition with newer techniques like Integrated Gradients [5], can effectively identify vital linguistic identifiers employed by the model to favor one category over another.

We utilize Named Entity Recognition (NER) to identify entities like person names or city names and average their salience scores. We calculate salience scores relative to the predicted class for each category of political leaning, yielding positive and negative scores for tokens. A positive score indicates that removing that token would likely reduce the model's confidence in its prediction, while a negative score suggests the opposite. Figure 1 presents a use case of SHAP for identifying important tokens.

We use two experimental designs to analyze global salience scores. The first experiment considers all articles' top 100 salient words, categorized by political leaning. The second one focuses on the top 100 most frequent words across all documents. These words are then used to execute data injection attacks during inference, leading to an attack success rate of 3.8%. This result establishes a baseline for our naive attack approach.

Results: Table 1 highlights the discrepancy in the model's performance across the left-leaning and right-leaning groups, which indicates a potential bias. It shows higher precision for left-leaning news (0.96) than right-leaning news (0.85), suggesting it's less likely to misclassify left-leaning news as fake. The model also has a higher recall for left-leaning news (0.92) than for right-leaning news (0.90), implying it's more adept at correctly identifying fake news if it's left-leaning. These discrepancies indicate that the model may not treat news items from different political leanings equally.

According to Table 2, the negative SPD (-0.39) and AOD (-0.03) indicate potential disparities in the overall prediction rates between the two groups. The EOD of -0.013892 suggests a slight difference in true positive rates, while the DIR of 0.47, being less than 1, indicates a potential bias towards the unprivileged group i.e., right-leaning.

3 CONCLUSION AND FUTURE WORK

In this work, we 1) introduce fairness aspects to be considered in the context of fake news detection using transformer models. 2) Through experiments, we demonstrate the extent of bias in current language models' performance in fake news detection across different political ideologies. 3) We employ interpretability techniques to gain insights into the behavior of these models, which not only aids in the development of more robust and effective models but also informs the design of debiasing strategies.

Table 1: Classification scores using DistilBERT classifier on all data, left and right-leaning (0: fake class, 1: real class)

	A	P	R	F1	P (0)	P (1)	R (0)	R (1)
Data	0.91	0.91	0.91	0.91	0.92	0.91	0.92	0.91
Left	0.91	0.96	0.92	0.94	0.75	0.96	0.87	0.92
Right	0.92	0.85	0.90	0.88	0.95	0.85	0.92	0.80

Table 2: Fairness metrics evaluated (SPD, EOD, DIR, and AOD)

Fairness Metrics	Value
Statistical Parity Difference (SPD)	-0.394171
Equal Opportunity Difference (EOD)	-0.013892
Disparate Impact Ratio (DIR)	0.472067
Average Odds Difference (AOD)	-0.031709

We propose two future directions: 1) exploring alternative methods to more effectively aggregate salience scores for named entities and other tokens, aiming to facilitate a global analysis rather than scrutinizing at a granular article level. 2) Enhancing the attack success rate by employing alternative strategies, such as word removal, word swapping, and context-preserving modifications instead of arbitrarily inserting or deleting words.

The rationale behind executing attacks on the model using interpretability lies in identifying vulnerable data points that exhibit bias towards the privileged group. Consequently, this enables the development of model-agnostic debiasing methods that surpass the capabilities of existing model-based debiasing approaches, thus increasing fairness in fake news detection tasks.

REFERENCES

[1] S. M Lundberg and S. Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30.
[2] J. Nørregaard, B. D. Horne, and S. Adalı. 2019. NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (2019), 630–638.
[3] J. Park, R. Ellezhuthil, R. Arunachalam, L. Feldman, and V. Singh. 2022. Toward Fairness in Misinformation Detection Algorithms. *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media* 16 (2022).
[4] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA). 1135–1144.
[5] M. Sundararajan, A. Taly, and Q. Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. 3319–3328.