Mitigating Bias in Fake News Detection: A Two-Stage Approach for Human and AI-Generated Text

Jiayi Ding, Jianyi Teng, Mahesh Nidhruva

Abstract

The pervasiveness of fake news, irrespective of its origin - human or AI systems, has exposed critical shortcomings in traditional detection methods. Monolithic classification has shown significant bias based on the content source. By introducing a structured framework to directly assess the veracity of news content, apply robust text representations and sentiment-aware feature augmentation to distinguish real from fake, the current challenges can be addressed. A two-stage framework that first identifies the news articles origin (AI vs. human), and then independently classifies for veracity assessment (fake vs. real), has been found effective in addressing the shortcomings of monolithic classification approaches. Experiments have shown improvements in the overall false news detection, while preserving high accuracy across different content sources. These experiments and results highlight the necessity of origin-aware architectures in misinformation detection systems.

1. Introduction

Digital misinformation now poses a global threat beyond the conventional boundaries of content creation - encompassing both human and AI-generated text. Digital misinformation continues to expand in scale and sophistication, with AI-generated content now comprising an estimated 15% of viral fake news (Bommasani et al., 2023). Though existing detection systems achieve over 90% accuracy, there is a bias based on content source (Pérez-Rosas et al., 2018 and Kreps et al., 2022). This disparity can be attributed to fundamental linguistic differences: AI models are better at syntactically flawless but semantically inconsistent text; however human fabrications often exhibit emotional cues and logical gaps (Zellers et al., 2020). Emergence of sophisticated content sources highlights the importance of detection architectures centered around veracity.

Impact of the digital misinformation are profound (Allcott & Gentzkow (2017) & Lazer et al. (2018)), as evident from multiple highly consequential events across nations and societies in the last decade. During the 2023 Brazilian election cycle, AI-generated fake news accounted for 40% of undetected misinformation, despite comprising only 8% of total cases (Benton et al., 2023). Existing systems miss this asymmetry and ignore the crucial AI-human dichotomy. By leveraging context-rich language representations and combining with innovative sentiment-aware feature extraction, our framework strives to achieve reliable fake news identification across the entire spectrum of news sources. This shift in perspective is imperative to how misinformation is identified and countered.

2. Background

The field of fake news detection has evolved over the years in stages. In the initial phase (2016–2018), the focus was primarily on feature-engineering-based methods i.e., linguistic and stylistic cues. Conroy, Rubin, and Chen (2015) systematic review was one of the earliest foundational works (Rashkin et al., 2017) established foundational linguistic benchmarks for human-written fake news, identifying markers such as exaggerated sentiment polarity and atypical pronoun usage. These features were operationalized using logistic regression and SVMs, achieving 79–83% accuracy on political news datasets (Pérez-Rosas et al., 2018). However, these models proved fragile: adversarial examples with simple word substitutions could reduce performance by up to 34% (Yang et al., 2018).

The second phase (2019–2021) introduced transformer-based architectures. BERT (Devlin et al., 2019) outperformed feature-based methods by leveraging contextual embeddings, achieving 91.2% accuracy on the FakeNewsNet benchmark. Models like Grover (Zellers et al., 2020) showed better performance by combining generation and detection capabilities within a shared architecture, reaching a 94.7% detection rate for GPT-2-generated fake news. Even with these advancements, domain generalization remained a challenge: cross-dataset evaluation by Bommasani et al. (2023) showed that BERT-based models suffered 22%–28% point drops in accuracy when applied beyond their training distribution, which is consistent with our findings too. Though AI-generated text can be syntactically near-perfect, it lacks the semantic depth - an insight that has shaped subsequent detection research.

The most recent phase (2022–present) has seen the rise of hybrid and origin-sensitive models. (Shu et al.,2022) proposed a dual-branch neural architecture combining BERT embeddings with stylometric features, significantly reducing false positives on satirical human-written news. (Benton et al., 2023) introduced OriginDetect, the first model to separate source identification from veracity assessment; however, its two-stage pipeline incurred a 15% computational overhead, which is again consistent with our findings. However, no model to date has fully resolved the trade-off between bias and accuracy. Kocoń et al. (2023) reported that while their sentiment-augmented RoBERTa achieved 89% precision on AI-generated fake news, it struggled with human-written deception, with recall rates below 63%. These findings underscore the need for adaptive detection heuristics based on content origin—a challenge this study directly addresses.

3. Methodology

3.1.Data

This study utilizes a hybrid dataset that integrates the FakeNewsNet datasets (Biradar et al., 2022) with synthetic news generated by ChatGPT using the Structured Mimicry Prompting (SMP) method (Dou et al., 2023). The dataset supports a comprehensive evaluation of fake news detection across four distinct categories: human-written fake news, human-written real news, LLM-generated fake news, and LLM-generated real news. Human-written samples are drawn from the FakeNewsNet repository, covering misinformation in the domains of entertainment (GossipCop) and politics (PolitiFact), while synthetic samples are crafted to emulate the linguistic and stylistic patterns of human-authored texts for generating highly realistic synthetic content.

Prior to model training, several preprocessing steps are applied to standardize and refine the dataset, including:

- Record filtering: Articles missing a title or description are removed
- Text normalization: Lowercase the text; URLs and extra whitespace are removed.
- Class balancing: All four categories are equally represented.
- Data partitioning: A stratified split yields 13,376 training, 1,672 validation, and 1,676 test articles (80/10/10%)

3.2 Baseline model

To establish a robust baseline for the proposed four-way classification task, several state-of-the-art Transformer-based models are evaluated, including BERT and its variants RoBERTa, DeBERTa, and ModernBERT. These models were selected due to their proven ability to distinguish between real and fake news, as well as their capacity to capture complex semantic, syntactic, and stylistic patterns in text (Nguyen et al., 2023; Choudhury et al., 2023).

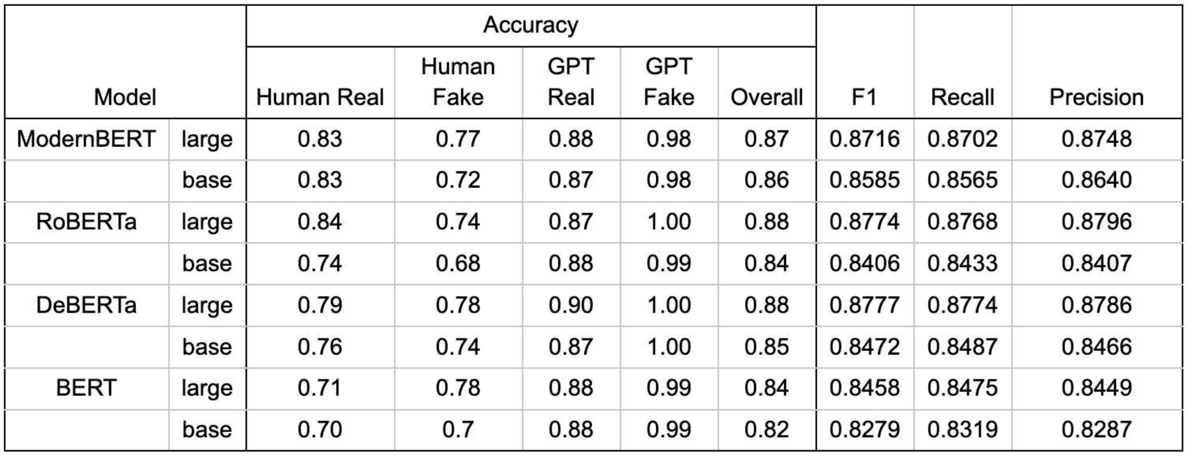
3.2.1 Methods

All models were fine-tuned using a consistent configuration: a batch size of 16, a learning rate of 2e-5, and training for two epochs. Label smoothing (ε = 0.1), weight decay (0.01), and early stopping based on the validation macro-F1 score were applied to enhance generalization. Evaluation was performed using standard metrics, including accuracy, precision, recall, and F1 score.

3.2.2 Result

The results presented in Table 1 demonstrate that all evaluated Transformer-based models achieve strong overall performance across standard classification metrics. ModernBERT-large achieves the highest overall accuracy of 87% and macro-F1 score, indicating its effectiveness in capturing the linguistic patterns needed for four-way classification.

Table1. Accuracy, Precision, Recall, and F1 Scores of Transformer Models on Human- and AI-Generated Real and Fake News



Model		Accuracy					F1	Recall	Precision
		Human Real	Human Fake	GPT Real	GPT Fake	Overall			
ModernBERT	large	0.83	0.77	0.88	0.98	0.87	0.8716	0.8702	0.8748
	base	0.83	0.72	0.87	0.98	0.86	0.8585	0.8565	0.8640
RoBERTa	large	0.84	0.74	0.87	1.00	0.88	0.8774	0.8768	0.8796
	base	0.74	0.68	0.88	0.99	0.84	0.8406	0.8433	0.8407
DeBERTa	large	0.79	0.78	0.90	1.00	0.88	0.8777	0.8774	0.8786
	base	0.76	0.74	0.87	1.00	0.85	0.8472	0.8487	0.8466
BERT	large	0.71	0.78	0.88	0.99	0.84	0.8458	0.8475	0.8449
	base	0.70	0.7	0.88	0.99	0.82	0.8279	0.8319	0.8287

However, all BERT models consistently perform better on AI-generated content, particularly in detecting GPT-generated fake news, compared to human-written fake news. This difference suggests a bias in current models, which tend to overfit to the surface patterns of LLM-generated texts while struggling to generalize to the more subtle, less structured patterns found in human-written misinformation. This observation is consistent with findings by Dou et al. (2023), which show that fake news detectors often have inflated confidence in detecting synthetic content while underperforming on human-authored instances. These results highlight the limitations of monolithic classifiers in handling origin-sensitive misinformation and emphasize the need for more refined detection strategies that consider the source of content.

4. Experiment

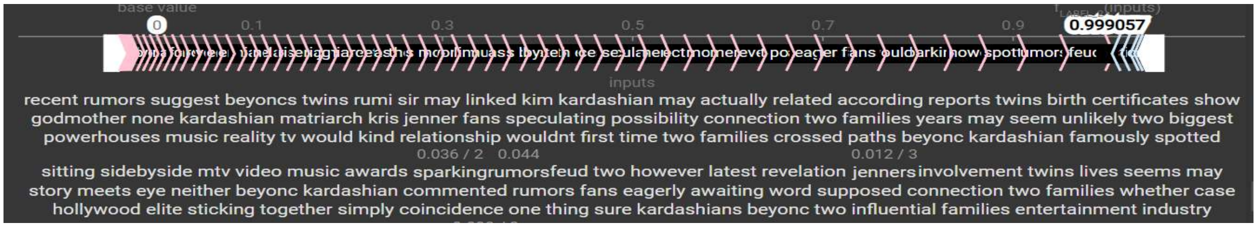
4.1 Bias analysis

4.1.1 Feature contribution and pattern analysis

The baseline model exhibits substantial bias in classifying both GPT-fake and GPT-real samples. To enhance model transparency and trustworthiness, a feature-based interpretability technique - **SHAPLEY Additive EXPlanations (SHAP)** - is applied to the baseline model. Interpretability frameworks such as SHAPE and LIME have demonstrated efficacy in uncovering feature contributions in NLP models (Barkai, 2022).

Rank	Average SHAP Value	Avg Score contribution	Max SHAP value	Avg Score contribution
1	**humble**	0.1857	interview	0.2285
2	confused	0.1232	**humble**	0.1857
3	gratitude	0.0858	close	0.1769
4	rubbing	0.0761	believe	0.1475
5	**sympathetic**	0.0652	busy	0.1459

Table2. Top 5 contributed token to “gpt fake” label prediction using average token score and max token score

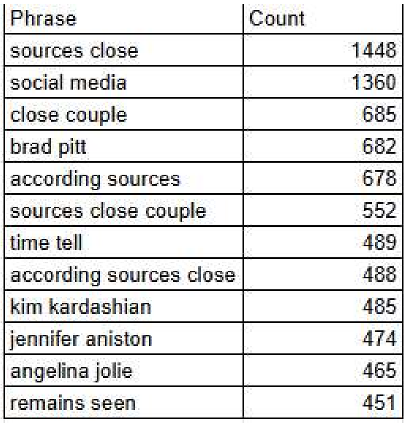


recent rumors suggest beyoncs twins rumi sir may linked kim kardashian may actually related according reports twins birth certificates show godmother none kardashian matriarch kris jenner fans speculating possibility connection two families years may seem unlikely two biggest powerhouses music reality tv would kind relationship wouldnt first time two families crossed paths beyonc kardashian famously spotted sitting sidebyside mtv video music awards sparkingrumorsfeud two however latest revelation jennersinvolvement twins lives seems may story meets eye neither beyonc kardashian commented rumors fans eagerly awaiting word supposed connection two families whether case hollywood elite sticking together simply coincidence one thing sure kardashians beyonc two influential families entertainment industry

An analysis of SHAP value rankings for predictions labeled as GPT fake reveals that the top five contributing tokens, identified by both average and maximum SHAP scores, include feud, rumors, spotted, frenzy, and sparking—lexical items that are relatively infrequent in human-written news. In a broader sample of 100 instances, similarly uncommon terms such as humble frequently emerged as significant contributors, suggesting a tendency of LLM-generated content to include lexically rare or stylized expressions.

In parallel, recent advances in text analysis continue to highlight the utility of n-gram features in revealing stylistic signatures (Jain, 2023). A targeted analysis of named entity and phrase frequency

identified high repetition of proper nouns (e.g., Kim Kardashian, Jennifer Aniston, Angelina Jolie)—appearing approximately 500 times across 4,084 GPT-fake samples—as well as over 1,000 instances of templated source-citation phrases such as sources close (see right-hand table for details).
Both the SHAP value tables and corresponding visualizations, combined with n-gram frequency analysis, indicate that machine-generated text contains distinctive lexical patterns. These uncommon yet highly repetitive terms disproportionately contribute to the 'gpt text' classification, suggesting the fine-tuned BERT model has learned to detect these artificial signatures.

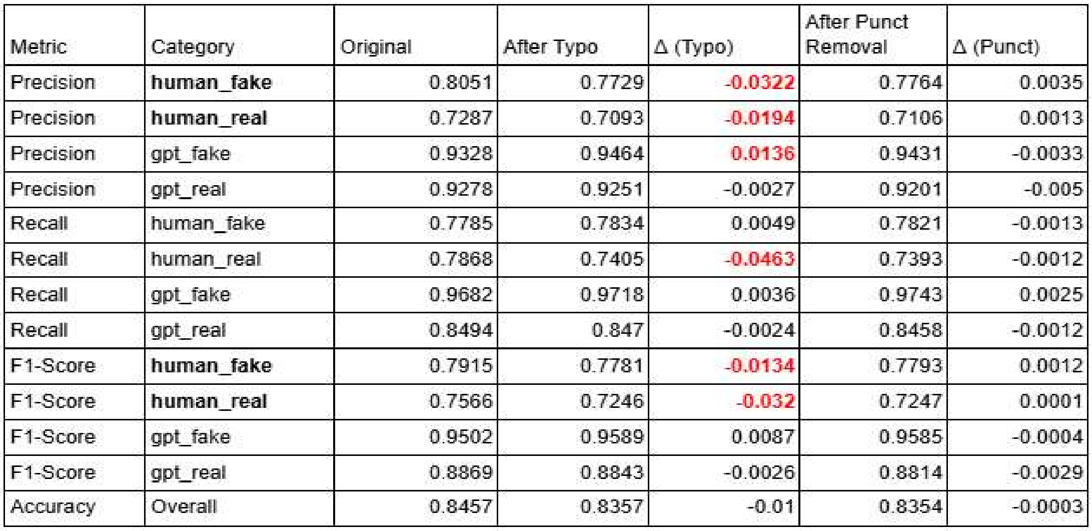


Phrase	Count
sources close	1448
social media	1360
close couple	685
brad pitt	682
according sources	678
sources close couple	552
time tell	489
according sources close	488
kim kardashian	485
jennifer aniston	474
angelina jolie	465
remains seen	451

Table3. Phrase with large count

4.1.2. Robustness analysis

To enhance model generalization and reduce overfitting, robustness checks—such as removing punctuation, introducing typos, and applying other feature perturbations as outlined in the model evaluation checklist (Ribeiro et al., ACL 2020) - can be used to assess the stability of the model.



Metric	Category	Original	After Typo	Δ (Typo)	After Punct Removal	Δ (Punct)
Precision	**human_fake**	0.8051	0.7729	-0.0322	0.7764	0.0035
Precision	**human_real**	0.7287	0.7093	-0.0194	0.7106	0.0013
Precision	gpt_fake	0.9328	0.9464	0.0136	0.9431	-0.0033
Precision	gpt_real	0.9278	0.9251	-0.0027	0.9201	-0.005
Recall	human_fake	0.7785	0.7834	0.0049	0.7821	-0.0013
Recall	human_real	0.7868	0.7405	-0.0463	0.7393	-0.0012
Recall	gpt_fake	0.9682	0.9718	0.0036	0.9743	0.0025
Recall	gpt_real	0.8494	0.847	-0.0024	0.8458	-0.0012
F1-Score	**human_fake**	0.7915	0.7781	-0.0134	0.7793	0.0012
F1-Score	**human_real**	0.7566	0.7246	-0.032	0.7247	0.0001
F1-Score	gpt_fake	0.9502	0.9589	0.0087	0.9585	-0.0004
F1-Score	gpt_real	0.8869	0.8843	-0.0026	0.8814	-0.0029
Accuracy	Overall	0.8457	0.8357	-0.01	0.8354	-0.0003

Table4. compute metrics under feature perturbations

Introducing typos into the validation dataset caused slight drops in precision, recall, and F1-scores for human-written text. The most affected label was 'human fake'. These results indicate that the baseline model (fine-tuned from modern BERT) demonstrates reasonable stability, though not complete robustness when handling human-written labels under feature perturbations.
Furthermore, the exploration table above confirms that introducing typos does not significantly affect performance for machine-written labels. As noted in Section 4.1.1, GPT-generated text contains distinctive lexical patterns that lead the model to learn these artificial signatures. One potential enhancement direction involves disrupting these patterns by intentionally adding typos to make GPT-generated text more closely resemble human writing. This approach—detailed and evaluated in Section 4.2—demonstrates effectiveness in reducing overfitting for GPT-labeled text classification.

4.2 Data level enhancement - Data Augmentation & Humanization

Beginning with an examination of the word count distribution in the cleaned dataset, it’s been observe that human_fake (mean: 468, max: 16,373) and human_real (mean: 530, max: 16,373) texts are substantially longer on average than gpt_fake (mean: 225, max: 520) and gpt_real (mean: 369, max: 4,292) texts. The baseline model uses a fixed maximum length of 500 tokens, which negatively impacts prediction accuracy for longer texts during fine-tuning, as truncation directly removes potentially useful information. This explains the model's poorer performance on human-written classes (which tend to be longer).

To address this word count imbalance between classes, two potential solutions emerge:

4.2.1. Filtering human-written texts to include only shorter samples during training

By standardizing the text length of the gpt_real, human_fake, and human_real classes to match gpt_fake (mean ≈300 tokens, max length=500) and rebalancing to 2,756 texts per class, a balanced dataset with all four labels was obtained, and by applying humanization to machine generated classes, a better evaluation result is achieved.

However, this approach has limitations:

Data Reduction: The dataset size decreased from 20,000 to 12,000 samples (>40% removed due to length constraints), and training set is impact Resulting in only ~8,000 training samples, which is a marginally small size for effective BERT fine-tuning. To address this, the data augmentation to artificially extend GPT-generated text length has been implemented

4.2.2.Applying data augmentation to GPT-generated texts to match human-written text lengths

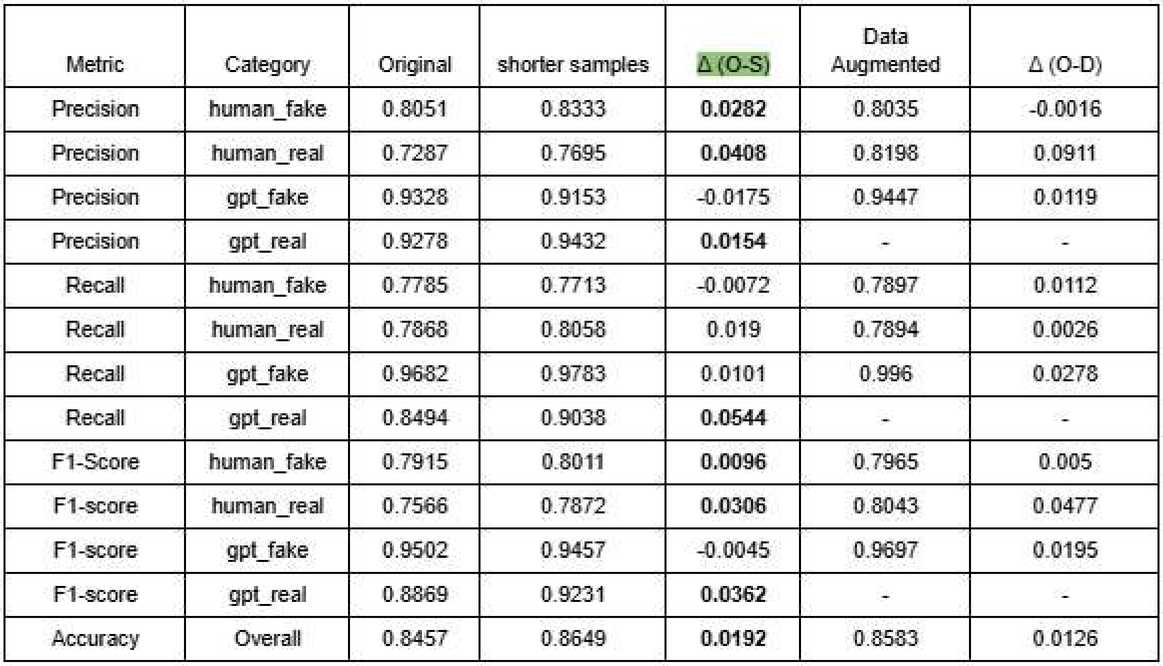
Data augmentation was applied by using GPT2-medium to regenerate both gpt_real and gpt_fake texts, extending them to match the average length of human-written texts. This strategy helps balance the dataset without discarding over 40% of the records.

For each regeneration, the following steps were followed:

- Dynamic Length Targeting: The target length was sampled from a normal distribution based on human text statistics (mean ± standard error), ensuring natural variability.
- Quality Control: Extreme outliers (texts over 10,000 tokens) were removed, and gpt_fake texts underwent a humanization post-processing step to enhance their authenticity.

Model evaluation showed that Method 6.2.1 outperformed this data augmentation approach, achieving significantly higher classification accuracy for human_real, human_fake, and gpt_real labels, while reducing overfitting in gpt_fake classes. Although the data augmentation method improved performance, it was slightly less effective than 6.2.1. Upon analyzing the regenerated dataset, it was observed that lexical patterns from GPT2-medium became overly pronounced, introducing additional noise.

As a result, both methods were selected as complementary enhancements to improve data quality and serve as inputs for subsequent model-level refinements.



Metric	Category	Original	shorter samples	Δ (O-S)	Data Augmented	Δ (O-D)
Precision	human_fake	0.8051	0.8333	**0.0282**	0.8035	-0.0016
Precision	human_real	0.7287	0.7695	**0.0408**	0.8198	0.0911
Precision	gpt_fake	0.9328	0.9153	-0.0175	0.9447	0.0119
Precision	gpt_real	0.9278	0.9432	**0.0154**	-	-
Recall	human_fake	0.7785	0.7713	-0.0072	0.7897	0.0112
Recall	human_real	0.7868	0.8058	0.019	0.7894	0.0026
Recall	gpt_fake	0.9682	0.9783	0.0101	0.996	0.0278
Recall	gpt_real	0.8494	0.9038	**0.0544**	-	-
F1-Score	human_fake	0.7915	0.8011	**0.0096**	0.7965	0.005
F1-score	human_real	0.7566	0.7872	**0.0306**	0.8043	0.0477
F1-score	gpt_fake	0.9502	0.9457	-0.0045	0.9697	0.0195
F1-score	gpt_real	0.8869	0.9231	**0.0362**	-	-
Accuracy	Overall	0.8457	0.8649	**0.0192**	0.8583	0.0126

Table5.Data level enhancement - Model evaluation metrics

4.2.3. Augmenting the news data with the linguistic features

Approximately 10 stylometric features—such as readability scores (Flesch, Dale-Chall), sentiment polarity (VADER), structural metrics (e.g., word count, exclamation marks, uppercase usage), and complexity metrics (e.g., syllable density)—were incorporated into the news text. A hybrid model was tested by concatenating BERT’s CLS tokens (768-dimensional) with these linguistic features (10-dimensional), aiming to complement BERT's embeddings. While the hybrid model maintained strong performance in detecting GPT-generated content, the human-class performance slightly declined, likely due to the interference of feature noise with ModernBERT’s native text understanding.

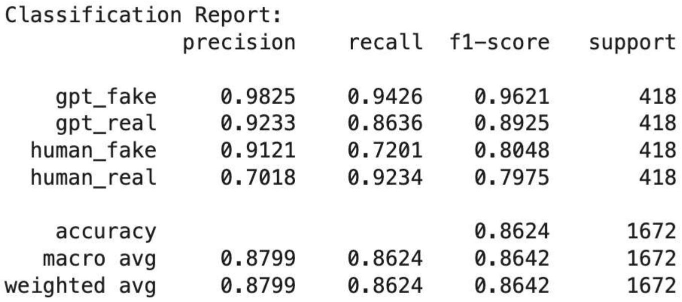
4.3 Model level enhancement - Two Stage Model

To address the performance degradation and bias exhibited by traditional fake news detection models when applied to AI-generated content, A two-stage detection framework is proposed. Two experiments were conducted for two-stage model:

a. Source identification followed by veracity classification
b. Veracity classification followed by source identification

4.3.1 Framework

Experiment a: This direction adopts a two-stage framework to improve fake news detection by incorporating source attribution. In Stage 1, a Transformer-based classifier (ModernBERT-base) is fine-tuned to identify whether a news article is written by a human or generated by an LLM (e.g., GPT). The model is trained with class-balanced cross-entropy loss and outputs both the predicted source label and a confidence score. In Stage 2, veracity classification is performed using separate models for each source. Inputs with high-confidence predictions (≥ 0.85) are routed to the corresponding veracity classifier (human or GPT). Inputs with low confidence are handled by a backup model trained on uncertain samples. All veracity classifiers share the same architecture and training strategy. During inference, each text is first passed through the source classifier, then routed based on confidence. The final output combines both predicted source and veracity (e.g., gpt_fake, human_real), providing both interpretability and source-aware veracity prediction.



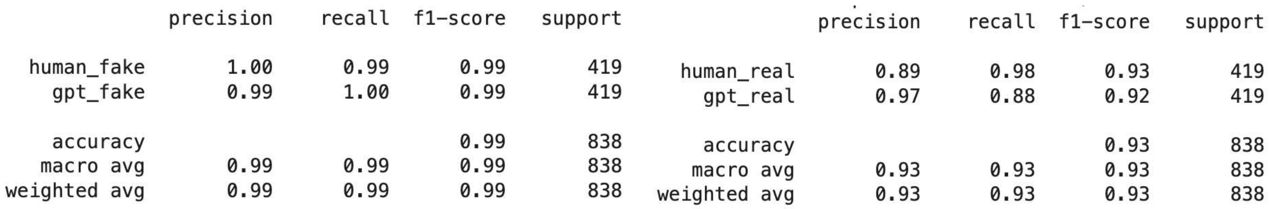
```
Classification Report:
              precision    recall  f1-score   support

    gpt_fake     0.9825    0.9426    0.9621       418
    gpt_real     0.9233    0.8636    0.8925       418
  human_fake     0.9121    0.7201    0.8048       418
  human_real     0.7018    0.9234    0.7975       418

    accuracy                         0.8624      1672
   macro avg     0.8799    0.8624    0.8642      1672
weighted avg     0.8799    0.8624    0.8642      1672
```

Table6. The classification report for Experiment A

Experiment b: For stage 1, the input is raw article text (title and body), which is fed into the ModernBERT-base. Long-context handling (512 tokens vs BERT’s 256) and domain adaptation enhancements were applied. The output from the model is binary classification of fake vs. real. In stage 2, fine-grained classification is performed through two sub-models routed based on the output from stage 1 (fake vs real) - one is a fake sub-model to detect the human_fake from gpt_fake and another is the real sub-model to detect the human_real from the gpt_real. Both the sub-models included the tokenization and training as shared components.



```
              precision    recall  f1-score   support

  human_fake       1.00      0.99      0.99       419
    gpt_fake       0.99      1.00      0.99       419

    accuracy                           0.99       838
   macro avg       0.99      0.99      0.99       838
weighted avg       0.99      0.99      0.99       838
```

```
              precision    recall  f1-score   support

  human_real       0.89      0.98      0.93       419
    gpt_real       0.97      0.88      0.92       419

    accuracy                           0.93       838
   macro avg       0.93      0.93      0.93       838
weighted avg       0.93      0.93      0.93       838
```

Table7. The classification report for Experiment B

The two-stage experiments reveal distinct strengths. Experiment A (source-first) achieves 85% accuracy, excelling at detecting AI-generated fake news but showing slight difficulty with human-authored real content. Experiment B (veracity-first) delivers near-perfect recall across categories, though it slightly underperforms on AI-generated real news. Overall, the source-first model provides better interpretability, while the veracity-first model excels in sensitivity. A hybrid approach may help balance these strengths in future work.

5. Conclusion

This study explores bias mitigation in fake news detection by introducing a two-stage framework that distinguishes between human and AI-generated content. It also applies data-level enhancements, such as text-length balancing and controlled augmentation, to address dataset imbalances and improve robustness for human-written content.

Despite the improvements made by the two-stage approach, there are still several limitations. First, the model was primarily built using political and entertainment news, so its performance may be affected when applied to different types of news content, such as economics or health-related topics. Additionally, it does not consider the evolving nature of misinformation, limiting its ability to adapt to changing tactics. Data imbalance is another challenge, especially when dealing with rare types of misinformation or underrepresented content in the dataset.

Future work could address these issues by developing models tailored to specific domains to better handle different content types. It could also incorporate features that track the evolution of misinformation over time. Additionally, integrating expert feedback through human-in-the-loop systems could enhance detection accuracy, and expanding the model to handle content in multiple languages and formats would improve its generalizability. Lastly, focusing on reducing bias and improving fairness across different groups would make the model more robust and better suited for real-world applications.

References

1. **Allcott, H. and Gentzkow, M. (2017).** Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.
2. **Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021).** On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
3. **Benton, L., et al. (2023).** Biases in fake news detection systems for AI-generated content. *arXiv* arXiv:2303.09345.
4. **Bommasani, R., et al. (2023).** On the opportunities and risks of foundation models. *arXiv* arXiv:2108.07258.
5. **Choudhury, S., et al. (2023).** RoBERTa for fake news classification: Better than BERT? *arXiv* arXiv:2303.09345.
6. **Conroy, N. K., Rubin, V. L., and Chen, Y. (2015).** Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.
7. **Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019).** BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* arXiv:1810.04805.
8. **Dou, Y., et al. (2023).** Fake news detectors are biased against texts generated by large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
9. **Kocoń, J., et al. (2023).** Sentiment-enhanced RoBERTa for robust fake news detection. *arXiv* arXiv:2302.01570.
10. **Lazer, D. M. J., et al. (2018).** The science of fake news. *Science*, 359(6380):1094–1096.
11. **Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018).** Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.
12. **Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017).** Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.
13. **Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020).** Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.
14. **Shu, K., Wang, S., and Liu, H. (2022).** Beyond news contents: The role of social context for fake news detection. *ACM Transactions on Information Systems*, 40(3):1–37.

15. **Vosoughi, S., Roy, D., and Aral, S. (2018).** The spread of true and false news online. *Science*, 359(6380):1146–1151.
16. **Yang, Z., et al. (2018).** Fake news detection via NLP is vulnerable to adversarial attacks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
17. **Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2020).** Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 33, pages 9051–9062.

Other Online Sources

1. **Barkai, K.** (2021). Interpreting an NLP model with LIME and SHAPE. *Medium*. URL: https://medium.com/@krbarkai.
2. **Jain, A.** (2024). N-grams in NLP. *Medium*. URL: https://medium.com/@abhishekjainindore24.
3. **Ponce-Bobadilla, A. V.**, et al. (2024). Practical guide to SHAPE analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17(11).

Contributions:

- Jiayi Ding:
  - Developed and evaluated a baseline single multi-class classification model for detecting fake news, comparing the performance across several transformer-based architectures, including BERT, ModernBERT, RoBERTa, and DeBERTa.
  - Designed and implemented a two-stage classification model, focusing on Direction 1, which first distinguishes between human- and AI-generated text, followed by veracity classification (real vs fake) within each category.
  - Contributed to the writing of the research paper, including sections on dataset preprocessing, baseline model, two-stage model architecture, experimental results, and the overall conclusion.
- Nancy(Jianyi) Teng:
  - Baseline model building includes fine-tuning BERT, RoBERTa, and DistilBERT.
  - Bias exploration: Conduct robustness analysis on the baseline model; build feature-level diagnostics by analyzing feature contributions and n-gram patterns.
  - Data-level enhancement: Apply data augmentation to GPT-generated texts to match human-written text lengths and perform humanization by adding typos and modifying punctuation in machine-generated data.
  - Paper writing across multiple sections includes experiments, conclusions, and references, Mainly focus on Bias analysis and data level enhancement section.
- Mahesh Nidhurva:
  - Build hybrid model to incorporate the linguistic features and compare the performance against the base ModernBERT
  - Two stage model experimentation - specifically building the Direction 2/Experiment B (i.e., Veracity classification followed by source identification), and validating the model's generalization ability with the totally new dataset (i.e.,ISOT)
  - Contributed to writing the paper across multiple sections including Abstract, Introduction, Background, References and augmentation of the experimentation sections with the details for the above two models