From Deception to Detection: Evaluating and Optimizing Language Models (LMs) to Detect LLM-Edited Fake News

Ali Azarsina **Bardia Khalafi** **Majid Faridfar**
{aliazarsina, bardiakhalafi, majid.faridfar}@ut.ac.ir

1 Abstract

As the prevalence of AI-generated content rises, detecting language model (LLM)-edited fake news has become increasingly challenging. This study evaluates the effectiveness of BERT, RoBERTa, and Llama (2,3) models in identifying fake news edited by LLMs. By rewriting a human-written fake news dataset (PolitiFact and GossipCop of FakeNewsNet) with Llama 2, we created a corpus of LLM-edited texts to examine model performance under these conditions. We observed a notable drop in detection accuracy across models on the LLM-edited dataset. To address this, we implemented three optimization strategies: (1) prompt engineering with zero- and few-shot learning using Llama, (2) fine-tuning BERT and RoBERTa on the original dataset, and (3) fine-tuning RoBERTa on the LLM-edited dataset. Results indicate that fine-tuning on the LLM-edited dataset achieved the highest accuracy, suggesting that adapting models to LLM-altered data significantly enhances detection performance. This work highlights the importance of tailored training strategies in combating LLM-driven disinformation.

2 Introduction

The detection and automation of identifying fake news have long posed significant challenges, with extensive efforts dedicated to this domain over the years. As digital information proliferates, the ability to discern truth from falsehood becomes increasingly critical. Recently, the advent of large language models (LLMs) has introduced a new layer of complexity to this problem. Users can now leverage these models to refine and modify fake news, creating texts that are more convincing and harder to detect. Research indicates that automatically identifying the falsehood in news enhanced or generated by LLMs is significantly more difficult than detecting fake news crafted solely by humans .

This project aims to address this emerging challenge by first generating a secondary dataset of fake news. Using a primary dataset of fake news, we will employ the rephrasing capabilities of LLMs to create new, more deceptive versions of these texts. Subsequently, we will assess the performance of current fake news detection models on both the original and LLM-enhanced datasets. This evaluation will help us understand the discrepancies in detection accuracy between human-created and LLM-enhanced fake news. Leveraging techniques acquired during this course, we will then focus on enhancing the robustness and accuracy of these detection models to better handle the intricacies introduced by LLM-generated misinformation.

3 Related work

Chen and Shu (2023) show that "LLM-generated misinformation can be harder to detect for humans and detectors compared to human-written misinformation with the same semantics" and Wu and Hooi (2023) found that "LLM-camouflaged fake news content leads to substantial performance degradation of state-of-the-art text-based detectors (up to 38% decrease in F1 Score)". Also, they introduce SheepDog, which enhances robustness against LLM-empowered style attacks, and Zhang and Gao (2023) examine in-context learning (ICL) and some prompt engineering approaches to enhance LLM's performance on news claim verification task. However, the improvements are ont considerable. Later, Su et al. (2024) realised that "Existing detectors are more prone to flagging LLM-generated content as fake news while often misclassifying human-written fake news as genuine",

so they provide a more improved model with higher performance on both human and LLM-generated news, by which we are inspired. We also found a fake news data repository called FakeNewsNet provided by Shu et al. (2020) that will be used in our project.

4 Baselines

Based on the papers we reviewed on improving LLMs for LLM-edited fake news detection, some models have been implemented to achieve this goal, but their source codes are not publicly available. We can use the results reported in these papers as a baseline (specifically for GossipCop and PolitiFact), but we prefer to use some classic baseline methods so that we can evaluate their performance on our own secondary dataset as well as the primary dataset.

Wu and Hooi (2023) and Su et al. (2024) have examined different models on the FakeNewsNet dataset, which we are using as our primary dataset, and reported the evaluation results in their papers. The results are shown in the tables below:

Method	PolitiFact		GossipCop	
	Acc.	F1	Acc.	F1
dEFEND\c	82.67	82.59	70.85	70.74
SAFE\v	79.89	79.85	70.71	70.64
SentGCN	81.11	80.77	69.38	69.29
DualEmo	87.78	87.76	75.51	75.36
BERT	85.22	84.99	74.60	74.50
RoBERTa	88.00	87.40	74.14	74.05
DeBERTa	86.33	86.30	73.86	73.80
UDA	87.77	87.74	74.28	74.22
PET	85.56	85.51	74.75	74.63
KPT	87.78	87.70	74.38	74.23
GPT3.5	71.11	69.61	61.49	56.30
InstructGPT	67.78	64.59	58.33	50.38
Llama2-13B	65.56	63.15	55.74	53.54
SheepDog	**88.44**	**88.39**	**75.77**	**75.75**

Based on the provided information, DeBERTa and RoBERTa appear to be appropriate baseline methods, demonstrating higher accuracy and F1 scores compared to other methods. These models are advanced variants of the original BERT model, each enhancing it in distinct ways. RoBERTa, developed by Facebook AI, improves BERT by training on a more extensive dataset (160GB vs. 16GB), using dynamic masking for varied training signals, eliminating the next sentence prediction task, and optimizing hyperparameters for better performance. DeBERTa, created by Microsoft, introduces a disentangled attention mechanism that separates content and position for better contextual understanding, uses relative positional encodings instead of absolute ones, and includes an additional decoding-enhanced layer to refine word representations. More importantly, both models are accessible via the transformers library. You can see some other strong models too, but they are not easily accessible.

We have not yet started testing these models, so we are still determining the specific hyperparameters we will set. We will report these details comprehensively in the next phase of our report.

5 Methodology

To address the challenge of detecting LLM-enhanced fake news, our approach will be structured as follows:

5.1 Dataset Creation

- **Primary Dataset:** We will begin with a primary dataset of fake news articles sourced from credible repositories and prior research datasets. This dataset will include a variety of fake news examples that have been manually verified and labeled. A comprehensive description of this dataset, including detailed sample characteristics, will be presented in the Data section of this paper.

- **Secondary Dataset Generation:** In the process of developing our secondary dataset, we leveraged the advanced linguistic capabilities of large language models (LLMs), specifically the Llama 2-7B architecture. This methodology involved the transformation of our primary dataset through a systematic application of the model's rephrasing functions. The procedure entailed inputting the original fake news articles into the LLM and generating semantically consistent but linguistically diverse iterations. Our approach incorporated both zero-shot and one-shot prompting techniques to facilitate this transformation. Through this methodological framework, we successfully generated alternative versions of 500 fake news instances and 500 real news instances from the primary dataset, thereby establishing our secondary dataset. This approach enabled us to investigate the model's capacity to produce variants of misinformation while preserving the fundamental deceptive content. The resulting LLM-

enhanced articles were designed to exhibit increased sophistication and present greater challenges for detection by conventional fake news identification systems. This process not only expanded our dataset but also provided valuable insights into the potential for LLMs to generate more nuanced and complex forms of misinformation, thus contributing to the broader understanding of machine-generated fake news and the challenges it poses to current detection methodologies.

5.2 Model Evaluation

- **Baseline Evaluation:** In our research methodology, we conducted a comprehensive evaluation of three prominent language models: BERT, RoBERTa, and Llama-2. These models were assessed on both our primary and secondary datasets to gauge their efficacy in fake news detection. Initial evaluations of the models without fine-tuning revealed suboptimal performance in the fake news detection task. Specifically, we observed accuracy rates of 0.46 for BERT and 0.48 for RoBERTa, indicating performance only marginally above chance level. This finding underscored the necessity for model refinement to enhance detection capabilities. Consequently, we implemented a fine-tuning process using a subset of our dataset. This step was crucial in developing more robust models capable of achieving the requisite accuracy for our subsequent analyses. The objective was to establish a baseline of high-performance detection against which we could measure potential degradation in accuracy when confronted with LLM-generated fake news. Post-fine-tuning, we observed a substantial improvement in model performance. The BERT model achieved an accuracy of 0.78, while the RoBERTa model attained 0.77. These enhanced accuracy rates represent a significant improvement over the initial, non-fine-tuned performances. This phase of our study provided valuable insights into the capabilities and limitations of these models in fake news detection tasks. It also established a robust foundation for our subsequent investigation into the potential impact of LLM-generated fake news on detection accuracy. The marked improvement in model performance post-fine-tuning highlights the importance of domain-specific training in addressing the complexities of fake news detection.

- **Accuracy Comparison:** Following the creation of our secondary dataset, we conducted an evaluation to assess the impact of LLM-generated content on model performance. We tested both base and fine-tuned models on this dataset to determine if variations in writing style could influence the models' decision-making processes. Our findings revealed a notable decrease in accuracy when models were presented with LLM-generated news articles. Specifically, we observed an average reduction in accuracy of approximately 0.1 across all models. It is worth noting that this decline in performance was more pronounced in base models compared to their fine-tuned counterparts. These results support our hypothesis that LLM-generated fake news can indeed diminish the effectiveness of models in fake news detection tasks. The observed decrease in both accuracy and F1-score underscores the need for developing robust solutions to address this emerging challenge in the field of misinformation detection.

5.3 Comprehensive Analysis

- **Error Analysis:** Our analysis revealed a significant correlation between model performance and the linguistic style of the text, rather than solely its content. This was evidenced by the performance decline observed when models were tested on the LLM-generated and LLM-rephrased news dataset (secondary dataset). Specifically, we observed decreases in F1-scores of 0.12 and 0.08 for the BERT and RoBERTa base models, respectively. The fine-tuned RoBERTa model also experienced a 0.08 reduction in F1-score. Interestingly, the fine-tuned BERT model demonstrated an improvement in F1-score on the secondary dataset, which we attribute to its potentially superior generalization capabilities due to its model size. Detailed results are presented in Table 1. These findings underscore that in most cases, model performance declined when faced with LLM-generated content. This trend highlights a critical issue in fake news detection that requires addressing to ensure robust perfor-

mance across varied writing styles and content generation methods.

- **Model Enhancement:** Based on our previous analysis, we focused on improving the model that exhibited the most significant performance decline, namely the RoBERTa model. Our approach involved further fine-tuning our previously fine-tuned RoBERTa model on a subset of the secondary dataset.

  The rationale behind this strategy was to expose the model to diverse writing styles of the same news content, thereby enhancing its robustness against stylistic variations and encouraging a stronger focus on the news content itself. We hypothesized that this approach would lead to improved generalization and more consistent performance across different writing styles.

  The results of this additional fine-tuning were notably positive. The RoBERTa model, which initially achieved an F1-score of 0.74, demonstrated a substantial improvement post-fine-tuning, attaining an F1-score of 0.96 on the fake news detection task. This marked enhancement in performance underscores the effectiveness of our approach in addressing the challenges posed by LLM-generated content in fake news detection.

5.4 Final Evaluation

- **Performance Benchmarking:** Once the improvements are made, we will conduct a final evaluation of our enhanced models on both datasets. This will allow us to benchmark the performance gains and assess the overall effectiveness of our approach.

- **Reporting:** We will document our findings, methodologies, and results in a comprehensive report, highlighting the key contributions and implications of our work.

By following this structured approach, we aim to develop robust methods for detecting LLM-enhanced fake news and contribute valuable insights to the field of automated misinformation detection.

Table 1: Performance comparison of different methods on Primary and Secondary datasets on different models

Method	Primary Dataset		Secondary Dataset	
	Acc.	F1	Acc.	F1
BERT	46	**63**	37	51
RoBERTa	48	**65**	45	57
Llama 2-7B*	46	59	51	**65**
Llama 3-8B*	74	**69**	0	0
FT-BERT	78	78	80	**81**
FT-RoBERTa	77	**78**	69	74
Llama 2-7B**	66	**79**	55	71
FT-RoBERTa	0	0	96	**96**

* Zero-Shot, ** Few-Shot (2)

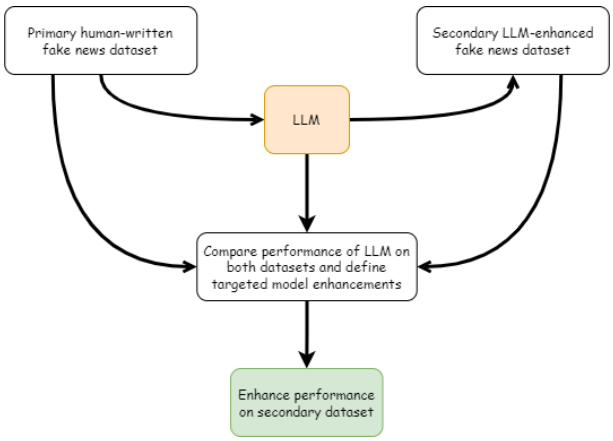

Figure 1: overview of the whole porject diagram

6 Data

In this study, we utilize the FakeNewsNet dataset, which comprises two primary sources: Politifact and GossipCop. These platforms specialize in fact-checking political and entertainment news, respectively. For the purposes of this research, we focus exclusively on the GossipCop subset due to its shorter news text lengths and larger sample size, which are more conducive to our analytical approach.

According to the instruction listed on the Github page, this data is to get the code provided to 'News Content' and 'Tweet' and 'Retweet'. We were able to successfully receive News Content, which for each news, a JSON file containing the Title and Content column, along with a series of other Metadata. But we had to use the API Twitter to get tweets and Retweets. Given that the use of this API has been paid and this website was also restricted to scrap data, we failed to get them. Of course, with a brief review we put into the links, they didn't have a particular effect. Because they were mostly empty or out of reach.

Finally we gained two data. The first is the Politifact, which has 381 Fake Symbols and 520 Real Symbols, and the second, Gossipcop, which has 2988 Fake and 6879 Real Symbols. We also extracted the news for each sample: title + enter character + Content.

We then did a series of Preprocess on these data. In the first step, we deleted the sympathies with both the Title and Content columns in the News Content file. It should be noted that we kept the news that only Title or only content. Because, however, they contain material that can continue our work.

In the next step, we deleted the news that was the same as Title and their content. Because our data is low overall, so the Duplicate data will have a devastating effect on the performance of the model and takes the number of Generalization and the model will be more preserved. By doing so, PolitFact and Gossipcop databases became smaller and their samples were equal to 763 and 5375, respectively.

Also, we had to reduce the number of samples. Because each model is limited to the size of the Context Window, and we cannot give them the samples where the news is longer. This is more evident when using Zero Shot and Few Shot. Because part of the Prompt is occupied by Instruction and there is less room for the news. So finally for each model, we produced and used our own data. For example, the data used to finish bert are all less than 512. While the news used in Llama 2 can have up to 4,000 token.

Here's about 500 of the Gossipcop fake samples using the LLAMA 2 Paraphris. This data, along with 500 Real data from the same data, formed our second data, which is used to deceive LLM.

7 Tools

Our project will employ standard deep learning libraries such as Pytorch, Transformers, and others for task execution. Google Colab's infrastructure will serve as our computational backbone, particularly for enhancing the accuracy of our model. Additionally, we aim to utilize the GPT-3.5 model from the Language Models for the creation of our secondary dataset.

8 Contributions of group members

Our group members cooperated and contributed to this project equally. The placement of authors' names in title section is chosen ahphabetically. The contributions of our group members are as follows:

- **Ali Azarsina:** Performed 100% on gathering data, 20% in evaluation, 25% in creating second dataset, 100% in improving model on secondary dataset, 40% in writing of report
- **Bardia Khalafi:** Performed 60% in evaluation, 20% in writing of report, 50% in creating presentation slides
- **Majid Faridfar:** Performed 20% in evaluation, 75% in creating second dataset, 40% in writing of report, 50% in creating presentation slides

9 Conclusion

In this study, we demonstrated the sensitivity of language models to stylistic elements in fake news content. Our findings reveal that when fake news articles are rephrased using advanced language models (LLMs), they become more challenging to detect by both human readers and automated systems. To address this vulnerability, we proposed a strategy of fine-tuning models on diverse stylistic variations of the same content, aiming to enhance their robustness against such variations and enable a focus on substantive content rather than superficial stylistic cues.

Our results show that this approach significantly improves model performance in detecting fake news across varying writing styles. By exposing models to a wide range of linguistic variations during training, we effectively strengthened their adaptability and detection accuracy. This research highlights the critical need for sophisticated and adaptable detection systems capable of responding to the rapidly evolving landscape of misinformation, particularly as advanced language generation technologies continue to emerge.

References

Chen, C. and Shu, K. (2023). Can llm-generated misinformation be detected? *ICLR*.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*.

Su, J., Zhuo, T. Y., Mansurov, J., Wang, D., and Nakov, P. (2024). Fake news detectors are biased against texts generated by large language models. *NAACL*.

Wu, J. and Hooi, B. (2023). Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. *arXiv preprint arXiv:2310.10830*.

Zhang, X. and Gao, W. (2023). Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *IJCNLP-AACL*.