COSE474-2021F: Final Project
Machine Generated Journalism Detection Using Deep Learning

Minseo Kim

1. Introduction

In the recent years, Large Language Models (LLMs) have shown a remarkable ability to generate human-like text, making them a potentially valuable tool for automated journalism. However, this same capability also poses several significant risks.

Firstly, the widespread use of LLMs in journalism could lead to an increase in misinformation. Given that these models generate content based on the data they are trained on, they are susceptible to replicating and amplifying any biases present in that data. This could potentially lead to the creation and dissemination of biased or misleading news articles.

Secondly, the use of LLMs in journalism raises ethical concerns around plagiarism. Since these models are trained on large quantities of data, it is possible that they could inadvertently generate text that closely resembles existing articles, violating intellectual property rights.

Lastly, the advent of LLMs could potentially undermine public trust in journalism. If audiences are unable to distinguish between human and machine generated content, it could lead to a general distrust in the information they consume, further exacerbating the current crisis of misinformation.

On the other hand, while several models exist to classify English articles as human-written or machine generated, the Korean language presents unique linguistic and structural facets which these models may not effectively capture. Therefore, there is a pressing need to specifically train a model capable of classifying Korean articles. This would contribute to ensuring the veracity of journalism in Korean language and help to maintain the integrity of information consumed by the public.

To tackle this problem, this project aims to train a model capable of distinguishing human written and machine generated text. In particular, the project aims to train a model capable of distinguishing human written and machine generated text in Korean, where the topic is relatively less explored.

2. Source Code

The source code to this project can be found in the following github repository:

https://github.com/ms-2k/COSE474_Final

However, while the dataset was generated from publicly available data, to adhear to distribution rights they are not included in the github repository.

3. Problem Definition

The primary objective of this research endeavor is to develop a computational model that can accurately differentiate between news articles authored by humans and those generated by machines, utilizing a transformer-based architecture. The intended model should possess the capability to analyze and interpret any given news article text written in the Korean language and subsequently produce a binary output. This binary output, in the form of a label, will indicate whether the given article is a product of human intellect or a result of machine generation.

Despite the ambitious nature of this project, it is important to acknowledge the constraint of limited available computational resources. Therefore, to overcome this limitation and ensure efficient use of resources, the final model will not be built from scratch. Instead, we intend to fine-tune an existing model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture. This approach leverages the established capabilities of BERT while allowing us to tailor the model to our specific task of distinguishing between human-written and machine-generated news articles in Korean.

4. Contribution

The sole main contributor to this project is Minseo Kim.

5. Related Works and Baseline

While there are few studies on detecting machine generated Korean journalism, similar attempts have been made by numerous researchers to detect AI generated content in general. Recent works such as RADAR (Hu et al., 2023) and DetectGPT (Mitchell et al., 2023) have attempted to distinguish human and machine generated text, where the former in particular utilized an adversarial model with human generated text along with machine generated text generated and paraphrased from the human written text.

Regrettably, these preceding works, while innovative in their own rights, have not been trained on Korean corpora. This presents a significant challenge when attempting to utilize them as a baseline for comparison with our model. The linguistic nuances and specificities inherent to the Korean language render these models less applicable in the context of our research. Consequently, in order to establish a more suitable benchmark for evaluation, we have opted to use foundation models trained on a more diverse corpus as the baseline. These foundation models, having been trained on diverse and extensive corpora, provide a more relevant and robust point of comparison for assessing our model's performance.

6. Method

There were a two primary challenges that had to be addressed prior to training the model. Firstly, it was difficult to obtain a large enough dataset consisting of both human written journalism and machine generated journalism. The problem was especially apparent when limited to Korean corpora.

Secondly, there are far more publicly available human written journalism than machine generated journalism. This imbalance in data could lead to a class imbalance problem, where the model could be biased towards predicting articles as human written.

To address the above problem we have decided to gather our own datasets by obtaining publicly available online Korean news, and then paraphrasing the articles with the help of an ensemble of large language models capable of text to text generation. This will allow us to obtain a dataset that is equally comprised of human written and machine generated news, and potentially be large enough to train a robust model with.

7. Overall Structure

insert image here later

8. Significance & Novelty

9. Experiments

10. Dataset

As described above the dataset is comprised of a corpus consisting of publicly available news articles from numerous Korean news outlets, and corresponding machine generated news articles paraphrased from human written articles by an ensemble of large language models capable of text to text generation.

11. Computing Resources

12. Experiment Setup

13. Comparison with Baseline

14. Conclusion

References

Jiang, G. (2023, May 30). Is AI-generated content actually detectable?: College of Computer, Mathematical, and Natural Sciences: University of Maryland. College of Computer, Mathematical, and Natural Sciences — University of Maryland. https://cmns.umd.edu/news-events/news/ai-generated-content-actually-detectable

Hu, X., Chen, P. Y., & Ho, T. Y. (2023). RADAR: Robust AI-Text Detection via Adversarial Learning. arXiv preprint arXiv:2307.03838.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305.

kr-FinBert-SC, Kim, Eunhee and Hyopil Shin (2022). KR-FinBert: Fine-tuning KR-FinBert for Sentiment Analysis. huggingface https://huggingface.co/snunlp/KR-FinBert-SC