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

Minseo Kim

1. Introduction

With the advent of generative large language models it has become easier to generate journalism with little human intervention. However, this could potentially be dangerous for numerous reasons, such as plagiarism, proliferation of misinformation, and misuse. The purpose of this work is to train a model capable of distinguishing human written journalism from machine generated journalism with transformer based architecture. In particular, this project will focus on jounalism in Korean, where the topic is relatively less explored.

2. Motivation

In an era of highly sophisticated LLMs, there is a real danger of misuse of machine generated journalism. Research has shown that it is difficult to distinguish between human generated content and machine generated content, especially if the latter is generated by paraphrasing human generated content. (Jiang, G.) There is real danger of bad actors mass generating and distributing believable machine generated fake news.
Furthermore, while there are plenty of machine generated content detection services in English, few are trained on Korean corpora. Thus, there is a necessity to train a model capable of distinguishing human generated content from machine generated content in Korean.

3. Problem Definition

The goal of this project is to train a model capable of distinguishing human written news from machine generated news with transformer based architecture. The model should be capable of accepting any article written in Korean, and output a binary label indicating whether the article is human generated or not. However, as available computing resources are too limited to train the model from scratch, the final model should be fine tuned from an existing BERT based model.

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.
However, neither of these works are trained on Korean corpora. Thus for this project we will be using various foundation models as the baseline to see if our fine tuned model is better.

6. Method

7. Significance & Novelty

8. Overall Structure

9. Experiments

10. Dataset

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