COSE474-2021F: Final Project Proposal
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. Problem definition & challenges

The primary goal of this work will be distinguishing human written news and machine generated news written in Korean. In order to achieve this, a transformer based model will be fine tuned from existing, reliable large language models. Additionally, to ensure the model is robust, the model will be trained on a dataset consisting of human written news and corresponding machine generated news, generated from the human written news dataset.

3. Related Works

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.

4. Datasets

The model will be trained on a corpus consisting of publicly available news articles from numerous Korean news outlets, and corresponding machine generated news articles of the same topic from various large language models capable of text to text generation, such as GPT, and LLaMA. The goal is to have a machine generated news article for each human written article in the training dataset.

5. State-of-the-art methods and baselines

While there are no comparable models for machine generated Korean news article detection, works such as RADAR and DetectGPT were able to achieve AUROC scores of up to 0.95 and 0.98, respectively in AI generated text in general. Other generic text classification models, such as KR-FinBert-SC (Kim, et al., 2022) were able to achieve similar metrics for other datasets of different domain.

6. Schedule

˜ 2023-11-15 : Determine base model, set up environment
˜ 2023-11-30 : Acquire dataset
˜ 2023-12-15 : Train and evaluate model

References

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