Turkish AI Generated Content Detection

Halil Erkan
Computer Engineering
TOBB ETU
Ankara, Türkiye
herkan@etu.edu.tr

Zeynep Meriç Aşık
Artificial Intelligence Engineering
TOBB ETU
Ankara, Türkiye
zeynepmericasik@etu.edu.tr

I. Idea

Today, AI-generated texts are present in social media, news websites, and essentially everywhere the internet exists. However, determining whether these texts are written by AI is not straightforward. This can lead to issues such as misinformation, loss of originality, and plagiarism, which can have profound impacts on societal structures.

According to recent studies, the number of publications has increased exponentially in recent years, raising the question of whether all of these were written by humans individually. We are now familiar with the similar patterns frequently used by AI, and these patterns are often found in some articles.

The goal of this project is to detect whether content in various writings is generated by AI. Among the many studies conducted on this topic, we found that there has not been much work focused on Turkish. We aim to investigate whether we can apply models and methods commonly used for English to a Turkish dataset.

II. Planned Work

The objective of our project is to effectively detect and classify AI-generated texts. To achieve this, we plan to use adversarial training (AT) and TuringBench test methodologies. Initially, we will prepare a comprehensive dataset, which will include texts written by both humans and AI. We will then train various deep learning and machine learning models on this dataset. These models include BERTurk, RoBERTa, CNN, and T5.

Additionally, to compare model performance, we will use baseline models such as BoW, Bayes, SVM, and Linear Regression. These baseline models require feature extraction, and we will use the following methods for this purpose: Frequency-based features (e.g., TF-IDF), N-grams, LIWC, Sentiment analysis, Syntactic features, and Readability scores For general data preprocessing, we plan to use the following methods: Tokenization, Lowercasing, Stop-word removal, Stemming, and Lemmatization.

Adversarial training will be used to enhance the robustness of our model and improve its ability to detect AI-generated texts more effectively. This method will test the model's resilience against potential challenges and aim to improve overall performance.

In addition, we will use the TuringBench test framework to evaluate our model's capability to distinguish between AI-generated and human-written texts. TuringBench provides a benchmark environment with examples of texts generated by various AI text generators and compares them to human-written texts. In our project, we will test the Turkish version of this methodology with data collected and created from different sources.

III. Datasets to be Used

For this project, we need datasets that include diverse texts produced by both humans and AI, grouped into two categories.

For human-written texts, we will use the following datasets containing Turkish articles from Wikipedia:

- Wikipedia Dataset [5]
- Turkish Wikipedia Dataset [6]

We were unable to find a dataset of AI-generated texts in Turkish. Therefore, we will generate these texts ourselves using AI. Different language models will be utilized in this process, including ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, Claude AI, and Gemini. For each of these LLM models, we will generate between 50-100 sentences, which will then be augmented to increase the dataset size.

References

[1] Tharindu Kumarage, Garima Agrawal, Paras Sheth, Raha Moraffah, Aman Chadha, Joshua Garland, Huan Liu 'A Survey of AI-generated Text Forensic Systems: Detection, Attribution, and Characterization'

[2] Morgan Sandler, Hyesun Choung, Arun Ross, Prabu David 'A Linguistic Comparison between Human and ChatGPT-Generated Conversations'

[3] D. Hee Lee and B. Jang, "Enhancing Machine-Generated Text Detection: Adversarial Fine-Tuning of Pre-Trained Language Models," in IEEE Access, vol. 12, pp. 65333-65340, 2024, doi: 10.1109/ACCESS.2024.3396820.

[4] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, Dongwon Lee 'TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation'

[5] https://www.kaggle.com/datasets/mustfkeskin/turkish-wikipedia-dump

[6] https://huggingface.co/datasets/umarigan/turkish$_{w}ikipedia_{d}ataset_{N}ER$