471 Progress Report

Turkish AI-Generated Review Detection

Abstract

The increase of AI-generated content, poses new challenges in distinguishing between human and machine-generated texts. This project focuses on the detection of AI-generated reviews in Turkish, leveraging classical machine learning algorithms as a baseline while also implementing two novel frameworks. It is aimed to compare the effectiveness of traditional models with these different approaches, addressing the unique linguistic features of the Turkish language. Initial results indicate that integrating language-specific adaptations significantly enhances the detection accuracy, offering promising directions for further research in AI-generated content identification in Turkish language.

Introduction

In the evolving landscape of text generation, the distinction between human and machine-generated content is becoming increasingly blurred. This has significant implications, particularly in areas like consumer reviews where authenticity impacts consumer trust and business reputation. To address this challenge, our research focuses on the detection of AI-generated reviews in the Turkish language, a linguistic area that is underrepresented in current literature.

For our baseline, we utilize classical machine learning algorithms—Support Vector Machines (SVM) and Naive Bayes—employing TF-IDF vectorization, and Linear Regression with n-gram vectorization. These methods have proven effective in various text classification tasks so with the two frameworks that is going to be applied to the Turkish data, the aim is to check whether they will work better than these well-established methods in the field of Natural Language Processing.

In addition to these traditional approaches, the two cutting-edge frameworks that is decided to be tested are yet to be explained. The first, as detailed in "TuringBench: A Benchmark Environment for Turing Test in the Age of Neural Text Generation" [atıf], offers a comprehensive suite of tests designed to challenge the capabilities of text generation models under diverse conditions. The second framework, titled "Enhancing Machine-Generated Text Detection: Adversarial Fine-Tuning of Pre-Trained Language Models," [atıf] describes an innovative approach involving adversarial fine-tuning of language models to improve detection accuracy.

The goal of this project is not merely to apply these frameworks but to adapt and optimize them for the Turkish context. By doing so, it is aimed to contribute to the broader discourse on machine-generated text detection, offering insights and methodologies that could be adapted for other languages and settings.

Literature Review

Related Work

-Datasets

Before testing on existing human-written reviews, the generation of AI-produced Turkish reviews was necessary, as no dataset of such content existed. To create this dataset, three of OpenAI’s GPT Language models—GPT-3.5, GPT-4, and GPT-4o—were utilized, with each model producing 100 reviews. The prompts used to generate these texts were included within the dataset. Due to the small initial size of the AI-generated dataset, augmentation techniques such as tokenization, lowercasing, stop-word removal, stemming, and lemmatization were employed, expanding the dataset to approximately 6000 entries.

For the human-written component, an existing dataset [atıf] was utilized and cleaned for training purposes. It was down-sampled to achieve a balanced dataset when combined with the AI-generated texts (AI/Human).

Initially, the project aimed to predict whether any text was written by a human or an AI. However, it was later realized that to accurately perform such broad detection, a much larger dataset would be required than could feasibly be produced manually. At that point, a general form of data had already been produced and trained using Turkish Wikipedia data. Although the results exhibited higher than expected accuracy and recall, they were not reliably indicative of real-world performance. Consequently, the focus was shifted to producing and utilizing a dataset specifically comprising user reviews, which could be more practically relevant for both companies and consumers.

-Baseline Models

The baseline models employed are SVM and Naïve Bayes using TF-IDF vectorization, along with Logistic Regression which utilizes N-gram vectorization. Optimal parameters for these models have not yet been determined. Preliminary testing indicates that the best performance is achieved when unigram, bigram, and trigram vectorizations are used concurrently. This may be due to the similarity in wording of complaints, regardless of the author. However, large language models like GPT often generate unique phrases spanning 2-3 words, which may not typically be used in manual complaint submissions. This phenomenon occurs irrespective of the model being instructed to produce outputs in daily, aggressive, or formal language tones, potentially explaining the superior efficacy of the combined uni-, bi-, and tri-gram approach.

Each model—SVM, Naïve Bayes, and Logistic Regression—also has its unique parameters that require fine-tuning. A dedicated run for hyperparameter optimization is necessary, though it has not yet been conducted. The outcomes from the preliminary tests and their comparative analysis are discussed in the section on Initial Results.

- Frameworks to be Utilized (TuringBench/Adverserial)

Upon accurate evaluation of the machine learning models, a separate experiment will be conducted using specialized frameworks designed to detect AI-generated text. These frameworks, however, are typically trained to identify AI-written content primarily in English and are not specifically tailored for user complaint reviews. The objective is to adapt these models to effectively operate on the uniquely assembled dataset for this project, which focuses on a specific use case.

The first, as detailed in "TuringBench: A Benchmark Environment for Turing Test in the Age of Neural Text Generation" [atıf], offers a comprehensive suite of tests designed to challenge the capabilities of text generation models under diverse conditions. The second framework, titled "Enhancing Machine-Generated Text Detection: Adversarial Fine-Tuning of Pre-Trained Language Models," [atıf] describes an innovative approach involving adversarial fine-tuning of language models to improve detection accuracy. An examination of the models and their operational mechanisms is presented below.

Turing Bench

Adversarial Fine-Tuning

Initial Results