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. These techniques need to be trimmed and used with combinations and separately to find the optimum data augmentation method in a way that contributes to the model training the most. For now, the baseline models were trained without any augmentation, with using a few of these methods, and with using all of them together. Even though the methods are not proven to be always beneficial, they helped with this case and the last decision was to utilize these augmentation methods all together through the whole dataset for both human and AI-generated.

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). For pre-processing, the methods explained above were also applied to the human data.

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.

TuringBench is a comprehensive benchmark environment designed to evaluate the capability of various models to distinguish between human-written and AI-generated texts. This framework is particularly relevant given the advancements in generative language models that produce text almost indistinguishable from that written by humans.

TuringBench includes a dataset of 200K samples comprising both human and AI-generated texts across 20 different labels. These labels represent various generative models and a human label, encompassing models like GPT versions, GROVER, CTRL, XLM, XLNET, and others.

There are two main benchmark tasks within TuringBench: Turing Test (TT) and Authorship Attribution (AA). Turing Test involves a binary classification problem where the goal is to classify texts as either human or AI-generated. It is modelled after the classical Turing Test, which assesses a machine's ability to exhibit human-like intelligence. The test contains subtasks for each pair of human and machine model. Authorship Attribution task extends beyond binary classification to identify which specific neural model generated a given text if it is determined to be AI-generated.

The framework also features a website with leaderboards that track the performance of various models on the benchmark tasks, providing a competitive and open platform for researchers. Thus, it can also be used for evaluation and comparison with the other models. However, preliminary results from TuringBench experiments suggest that newer models like GPT-3 and FAIR\_wmt20 generate text that is highly indistinguishable from human writing, posing challenges for current detection methods. So, it might be hard to surpass the Machine Learning models that have been on use forever as in accuracy or recall metrics.

For this project on detecting AI-generated complaint reviews in Turkish, TuringBench offers a robust framework for developing and evaluating specific models. Given that TuringBench primarily deals with English texts, it will be needed to train or fine-tune the models on the Turkish dataset. This involves either adapting existing models within TuringBench to understand Turkish through transfer learning. TuringBench’s tasks are also designed for general text, but they can be customized to focus specifically on complaint reviews. This might involve adjusting the types of prompts used for generating machine text which is already available in the newly generated Turkish dataset or the features which used in model training to better capture the nuances of complaint language.

The evaluation metrics are going to stay as usual traditional ones such as precision, recall, F1 scores, and accuracy to evaluate the models. Given the class imbalance in real-world scenarios (possibly more human than machine-generated texts), and also in this project which the human dataset is many times greater than AI-generated, these metrics can help assess the effectiveness of the detection framework.

To leverage the benchmarking aspect of TuringBench, setting up similar leaderboard systems to compare different models’ performances on Turkish complaint review dataset would be appropriate. This could foster a collaborative environment and push for further improvements in the models.

By integrating TuringBench’s methodologies and adapting its tasks and datasets for Turkish, the project can establish a pioneering framework for detecting AI-generated texts in underrepresented languages and specific domains like complaint reviews.

The Adversarial Fine-Tuning framework focuses on enhancing the detection of AI-generated text through adversarial fine-tuning of pre-trained language models (PLMs), such as BERT (Bidirectional Encoder Representations from Transformers).

The framework generates adversarial examples that mimic human modifications to texts. These examples are crafted using the T5 model, which modifies the input text by introducing subtle perturbations that are typically indistinguishable to humans but can mislead machine learning models. These adversarial examples are then used in training the PLMs. The process involves re-training the PLMs with a mix of original and adversarial texts, which helps the models learn to differentiate between human-like AI-generated text and genuine human text.

The PLMs are fine-tuned using a binary classification approach where the model learns to classify texts as either AI-generated or human-written. The fine-tuning process leverages adversarial examples to improve the robustness and accuracy of the models in detecting nuanced differences in text.

The adversarially fine-tuned models show a significant improvement in detecting AI-generated texts, reducing misclassification rates and enhancing metrics like accuracy and F1 score compared to traditional fine-tuning methods.

To apply this framework into this project, the first step is to utilize a model like T5 to create adversarial examples from the Turkish dataset. Since the focus is on complaint reviews, the adversarial modifications should mimic common expressions and nuances specific to complaint narratives in Turkish. Since the original framework is demonstrated primarily on English data, adapting the language model to understand Turkish is crucial. This might involve pre-training the model on a large corpus of Turkish texts or using a multilingual model that includes Turkish in its training data. So, there might be a need for a big Turkish dataset extra from the human-AI dataset that is generated for other models.

The adversarially fine-tuned model will be trained using both the original and adversarially modified Turkish complaint review texts. This training will help the model learn the specific characteristics of AI-generated versus human-generated complaint texts.

To evaluate the model’s performance, standard metrics like precision, recall, and F1 score will be used. Additionally, considering the creation of a validation set that reflects the real-world distribution of human and AI-generated texts will be essential for assessing the model’s practical effectiveness.

Based on initial results, to better capture the subtleties of AI-generated texts in the context of Turkish complaint reviews, the adversarial examples and fine-tuning processes will be iteratively refined.

By utilizing the adversarial fine-tuning approach, this project can significantly advance the detection of AI-generated text in a less commonly studied language and application area, providing valuable insights into the capabilities and limitations of current NLP technologies in new domains.

Initial Results