1 Introduction

Artificial Intelligence language models are increasingly used to generate content across a wide range of fields, including informative texts, user reviews on multiple platforms, and news articles. While their considerable benefits in terms of efficiency are evident, their widespread accesibility has generated a growing need for distinguishing between human-written and AI generated texts. This problem becomes considerably relevant in academic assesments, online platforms that rely on real user input, and the malicious widespread of fabricated news.

In this project we address the problem of detecting AI generated texts, by focusing on genre-based comparisons between human and machine generated texts across multiple writing styles.

1.1 Contributions

The main contributions of the team members on the project are:

Ungureanu Eduard Mihai: Ungureanu Robert Gabriel:

1.2 Summary of the approach

Our approach consists of three main stages. First, we construct a human written text corpus based on three main wirting styles: informative, news and reviews. Second, we construct an AI generated text corpus that covers the same writing styles. Finally, we combine all texts, human written and AI generated, into a single dataset, that explicitly includes both the source label and the style used. This will be used in further stages for training and evaluating AI-generated text detectors.

1.3 Why we chose to approach this subject

We chose to approach this subject due to the increasing use of AI-generated text across numerous platforms and academic contexts, accompanied by the concerns about missuse and lack of transparency in written communication. By approaching this subject we are aiming to conduct deep research into the performance of detection AI in relationship with the control for genre effects, a problem which makes it unclear wether models detect AI-specific patterns or exploit domain differences.

1.4 What other research has been done in this area

Among the notable works on this problem we highlight the research into the Stylometry "Artificial Writing and Automated Detection" by Brian Jabarian and Alex Imas, where the authors investigate the viability of current AI-generated text detectors, while tested in realistic conditions. They base their work on cross-domain data, proving the variation of performance between different domains. They also highlight the critical implications of False Negatives (AI classified as human), which render the detector useless, and the False Positives (human classified as AI) which can affect real users. Another notable paper is "Benchmarking AI Text Detection" by Shushanta Pudasaini et al. which focused on the lackluster generalisation of detectors between different domains, and the way that multiple simple evasion strategies destroy performance.

2 Approach

2.1 Repository and Project Organization

All code, data processing scripts and intermediate datasets of this project are organized in repository, structured to clearly separate raw data, human and AI generated text and analysis scripts. The repository link can be found at https://github.com/Eduard815/Styllometry-on-LLM. At the current state of the project, the focus is set on providing reliable datasets for human and machine generated text, which forms the foundation for future modelling and training.

2.2 Software Tools and Libraries

This project is implemented entirely in Python. We have so far used the following tools and libraries:

- pandas for data filtering and dataset manipulation
- matplotlib and seaborn for visual analysis through graphics
- WordCloud for lexical analysis
- OpenAI API for generating AI-written texts with different stylistic settings

- LaTeX for writing the documentation and reports itemize)

2.3 Training and Proccessing

The structure of the dataset has been implemented to ensure the comparability between human and AI-generated texts First, we created three separate human-written sections, to cover the stylistsic genres of informative texts, news and reviews. These texts have been selected from publicly available datasets, and have been processed to include exactly 100 texts from each category of various sizes. Second, AI-generated texts on the three specified genres have been created using a large language model. Constraints by genre-specific prompts and text sizes have been applied to ensure compatibility with the human written texts. Third, all text subsets have been combined and merged into a single dataset designed to label the source of the text (human or AI), and the genre they belong to. This dataset serves as the foundation for later development and exploratory data analysis. (The introduction of at least one more genre has been planned)

2.4 Exploratory Data Analysis

For this study we used a balanced corpus of 600 paragraphs - 300 human written and 300 AI-generated. Paragraphs have beeen uniformly distributed into three categories: informative, news and reviews. Our final(in progress) corpus is described by three fields: text, label and style. Our EDA visualizes graphs for class distributions (human vs AI), style distribution (informative, news, reviews), paragraph lenght distribution (number of words) and two WordClouds, one for human and one for AI-generated paragraphs to highlight the most used words in each category.

2.5 Used Models and Hardware Requirements

At the current stage of the project, no machine learning models have been trained. However, the models we will be using are the classic ones (Logistic Regression, Support Vector Machine, Naive Bayes, XGBoost) as well as opensource trained models like RoBERTa (probably). Hardware usage:

- no GPU required
- RAM < 2 GB (at this stage of the project we don't need more than 100 MB of RAM but we plan on extending our datasets)

- CPU : I5 13420h - I713620H which are more than sufficient for the task itemize)

2.6 Evaluation Methods

To evaluate the performance of classic and trained models we will use the following metrics:

- FPR (False Positive Rate) - indicates the probability that a passage written by a human is mistakenly identified as AI
- FNR (False Negative Rate) - indicates the probability that a passage generated by AI is mistakenly identified as human
- AUROC (Area Under ROC) - indicates that a random AI generated paragraph is ranked above a human written paragraph

- delta-Mean - indicates the average score of the AI-generated text minus the average score of human-written text, assigned by the detector. The higher the value, the more likely a detector is to distinguish between human and AI-generated text Each model will be evaluated under the same conditions, using the same corpus and metrics. The results will be summarized in tables showing each model's performance. itemize)