Executive Summary: Detecting LLM Authorship

Andrew Sack and Alireza Tehrani

June 2025

1 Introduction

Large language models (LLM) have been a massively disruptive technology across many areas of fields. In particular, within the academic, educational, and social media landscape. Many users that utilize LLM can quickly synthesize large amounts of text geared to a specific purpose using prompts. These texts can pose challenges in assessing students' ability to write, synthesizing unoriginal academic content, or using human bots to pose as humans on social networks. This greatly warrants the need to decipher whether a given text is written in human form or LLM written. The goal of this project is to solve a classification problem whose input is text files.

1.1 Stakeholders and key performance indicators

Our stakeholders consist broadly of those interested in spam or fraud detection. These include academic institutions, social media companies, survey companies, etc.

For our project, as the dataset is relatively balanced our primary key performance indicator was accuracy. However, as the model has the potential to be used to accuse somebody of fraud, we also consider the rate of false positives of flagging a piece of text as LLM written.

2 Dataset & Features

2.1 Dataset

For this task we used the AuTexTification dataset, a dataset used for a competition from the 5th Workshop on Iberian Languages Evaluation Forum at the SEPLN 2023 Conference. This pre-cleaned dataset was a collection of both human and LLM generated text, with the task of distinguishing between each. Furthermore, a test set was reserved in the competition which we reserved as our test set. The text within the AuTextification dataset is grouped into five domains: legal, wikipedia, tweets, reviews and news. The training set consists only of the tweets, legal and wikipedia domains, whereas the test set consists of news, and reviews.

2.2 Feature Extraction

One of the primary challenges in this project was extracting numerical features from a text dataset. We considered a very large number of features that text could have. Some simple features that we extracted include

- The total number of words.
- The relative frequencies of different punctuation.
- The relative frequencies of different letters.
- The average lengths of words and sentences.
- Lexical diversity, defined as $\frac{\text{UniqueWordCount}}{\text{WordCount}}$ or the proportion of words that appear once, $\frac{\text{Number Of Words Appear Once}}{\text{Total Words}}$

We also employed some existing natural language processing libraries to extract features including

- The number of (complex) verbs, nouns, contractions, adjectives, adposition and many more.
- The number of emotional word used, the text's polarity (positive, negative, or neutral) and sentiment (personal opinion and factual information).
- The grade-level or years of education needed to understand the text (Flesch reading ease, and Gunning Fog index).
- The relative frequency of different vowel sounds.

For each text, the total amount of features that was extracted is 316.

3 Methods & Results

We performed a stratified train/validation split with a 80/20 % split on the training data which consisted of 33,845 data points and only includes the domain

we reserved the test dataset which consisted of 21,832 data points, approximately 40% of the total number of datapoints. The validation set was used to hyperparameter tune our various models. We compare our models using its classification accuracy on the test set, defined as the division of the total correct classification divided by total number of predictions. We utilize the following classification models to select the best candidate:

- Random Forest
- Light Gradient-Boosting Machine (Light-GBM)
- Feedforward Neural network (FFNN)

When evaluating on the validation set, we achieve an accuracy of 80% using a random forest model, 79% using a FFNN and 81% using a Light-GBM. However, when evaluating on the (out-of-domain) test set, we have achieved an accuracy of 62% using a random forest model, 59% using a FFNN, and 65 % using a Light-GBM. This was due to the test-set being part of a different distribution than the training set.

4 Future Goals

4.1 Better Quality Dataset

The number of sentences in the dataset is small making it hard to discern between text written by a LLM and a human. Furthermore (excluding social media tweets), real-life usage of LLM generates substantial more text, e.g. when students are writing essays.

4.2 Realistic Prompts

The prompt on the AuTextification dataset is based on obtaining the initial part of the human written text, and autocompleting the rest of the sentence. Majority of users of LLM do not provide prompts of this kind, but rather provide context, and give directions for completion. Alternatively, majority of users would provide a human-written sentence to the LLM and ask for it to modify it partially or completely, making it a substantially more realistic problem to consider for LLM authorship.

4.3 Detecting Multiple LLM authorship

Another important problem to consider is determining which LLM wrote a piece of text. This helps with detecting the usage of unapproved LLM that can pose legal or safety risk, determining competitors LLM usage, or preventing intellectual property contamination.