I had simply used a  Random Forest Classifier model and XGB Classifier to test the accuracy of my ml model.

But I had created a lot of features in order to machine understand the difference between llm-generated text and human generated.

Firstly , I had created perplexity and burstiness scores as features :

Perplexity is a measure of how well a probability distribution or probability model predicts a sample. In the context of text detection, a lower perplexity value can indicate that a given piece of text is more likely to be "in line" with the language model's training data.

Burstiness is a property of a sequence of events where certain elements occur in clusters or bursts rather than being uniformly distributed over time or space. If certain words or phrases occur in clusters, it could be an indicator of specific themes or topics that deviate from the norm. In the context of text detection, burstiness score can be used to identify such patterns.

Then I counted number of different types of POS present in the text . Part-of-speech analysis emphasizes the dominance of nouns in ChatGPT texts, implying argumentativeness and objectivity, while the dependency parsing analysis shows that LLM texts use more determiners, conjunctions, and auxiliary relations.

Sentiment analysis, on the other hand, provides a measure of the emotional tone and mood expressed in the text. Unlike humans, large language models tend to be neutral by default and lack emotional expression. So, I have calculated pos, neg and neu features.

As humans are  tend to be more expressive in less words they often used exclamation and question marks.

The vocabulary features offer insight into the text’s word usage patterns by analyzing characteristics such as average word length, vocabulary size, and word density. Data have shown that human-authored texts tend to have a more diverse vocabulary but shorter length. So I had created different features like number of words, avg word length etc.

Then I passed the text through bert model to get the word embedding .

With these features , I had passed through different classifiers  to test my model.

Due to lack of fast computer , I had compressed the dataset set to run the code fastly . I tried to increase the size but the laptop is automatically getting restart.

Thank you for the project , learned a lot of things .