Team information: Roles and responsibilities

The team responsibilities are divided by task. Preprocessing and feature set creation will primarily be handled by Alexander and Jacqueline, whereas model training and tuning will be handled by all 3 members. Evaluation will be the responsibility of Rayhaan, and the final report will be handled by all 3 members.

Context

The problem identified is the difficulty of discerning AI generated content from organic, human generated content. With the increased availability of power LLMs like ChatGPT, outsourcing tasks like schoolwork to AI has become very common. Schools and workplaces need a way to combat this by identifying content not generated by a human.

This is a challenging problem because AI is getting better at writing text that could pass off as human created. Even humans can have trouble differentiating between content generated by AI and content generated by humans. Although dystopian, this problem can be solved with the help of machine learning. We can use a neural network to analyze snippets of text to predict if they are AI or human generated.

The tool we are proposing will allow teachers, professors, and other professionals to easily identify AI-generated text. This will help them make more informed decisions about how to identify this kind of plagiarism, rather than making “educated guesses”, thereby reducing false positive and false negative plagiarism claims.

This is relevant to the world as AI will continue to grow and be used in various environments. As AI becomes more easily accessible, it is likely there will be more AI generated content and works. This project would ensure work is human generated, helping maintain integrity in places such as schools, universities, and workplaces.

Dataset

Link to dataset: https://www.kaggle.com/datasets/jdragonxherrera/augmented-data-for-llm-detect-ai-generated-text?select=final_train.csv

For this task, we will use the “Augmented data for LLM” dataset on Kaggle, which was created for the “LLM - Detect AI Generated Text” competition organized by Vanderbilt University. The original aim of the competition was to create a model that can identify AI generated essays. The dataset we are using contains text selections from various human and AI generated essays, the average length of a snippet being around 400 words. There are 346977 unique samples in the dataset. We will most likely only use a subset from this dataset to train our model, as well as possibly reduce the length of each snippet. We also plan on adding to the feature set by computing features such as unique word proportion, average sentence length, and other metrics.

Proposed Solution

This is a predictive problem, as we are trying to predict Variables involved as features are the words themselves, evaluated as N-grams with TF-IDF, average sentence length, unique word ratio, and stop-word frequency. Additional metrics may be used to expand the feature set. The target variable is a binary label indicating 1 for AI generated, and 0 for human generated.

CNN, as it has been shown to be effective in analyzing the sentiment of a text [1] as well as detecting phishing/scam emails [2], which is a task similar to the one in this problem. We will be selecting a subset from the dataset to train on, as the dataset is very large. Of this subset we will split it 80-20 into a training set and test set. This data will be preprocessed using the Word2vec and/or gloVe [3] technique to represent our text samples as vectors, to be processed by the CNN. We will document these results using a confusion matrix, evaluating them primarily using accuracy.

This strategy is optimal as our problem is one of a binary classification, and a confusion matrix is a concise and readable way of displaying the effectiveness of our model. The inspiration for this problem comes from the growing trend of identifying human texts from AI texts. In an MIT Technology Review article [4], the author described the trend and growth of AI texts and the various ways scientists and developers have been developing to differentiate between the two. It is also apparent that this is an emerging problem, as there is a game site called “human or not” [5], where users will decide if the text on the screen was human or AI generated. The fact that there is a game for this indicates it is not straightforward and simple to differentiate AI text from human text.

Existing solutions include the Giant Language Model Test Room, which improved human detection of fake text from 54% to 72%. This is a visual tool that highlights text passages to determine if text likely human or AI generated. The three tests used to determine if the text is human or AI generated are probability of the word, absolute rank of the word, entropy of the predicted distribution [6].

References

[1] https://www.sciencedirect.com/science/article/pii/S1877050917312103

[2] https://siiet.ac.in/wp-content/uploads/2023/12/46.A-Case-Study-on-Deep-Learning-in-Fraud-Detection_compressed.pdf

[3] https://www.alpha-quantum.com/blog/word-embeddings/introduction-to-word-embeddings-word2vec-glove-fasttext-and-elmo/

[3] https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/

[4] https://www.humanornot.ai/

[5] https://aclanthology.org/P19-3019/