Project: ChatGPT Detection
Team: Rachel Buttry, Derek Kielty, Dora Puljić, Jack Arbunich

Overview: Generative LLM such as ChatGPT are becoming increasingly prevalent in society. In this project, we trained a classification model to detect whether text was generated by a human or ChatGPT. This problem is important to defend against malicious uses of ChatGPT (eg. phishing attacks and intentionally misleading news publications) and because it is important to understand the limitations of existing generative LLMs. In this project, we were motivated by the possibility of using ChatGPT for plagiarism.

Stakeholders: Universities and schools

KPIs: Accuracy (proportion of true positives to the total number of positives) and precision (proportion of true negatives and true positives to all instances)

Approach: We began by finding an open-source dataset of 16K essays written by students in grades 6-12. We then used ChatGPT to generate 16K essays from prompts created by doing a topic analysis of the student essays. We trained a deep learning model using the resulting data set of 32K essays. The model consisted of (1) an LSTM block, (2) a block of metrics, and (3) and linear neural network. After the preprocessing which removes punctuation and corrects most spelling errors, the *(1)* block reads the text word by word, allowing for the model to learn word meaning and placement. Meanwhile, *(2)* calculates metrics on the original text. These metrics were developed from natural language processing analyses – burstiness and sentiment. We defined burstiness as variance in sentence length. Texts also received a sentiment score, from the trained VADER sentiment intensity analyzer, which calculates four values between -1 and 1. One of these values is known as the compound score, where values close to -1 indicate a negative sentiment, and values close to 1 indicate a positive sentiment. Finally, *(3)* combines the output of *(1)* and *(2)* to compute the classification probability.

Results: Our model achieved an accuracy of 98.96% and a precision of 96.78%. The average burstiness of human-written text is 324.48 with a standard deviation of 1577.52, and the average burstiness of ChatGPT-generated text is 89.14 with a standard deviation of 21.93. The average sentiment score of human-written text is 0.59 with a standard deviation of 0.70, and the average sentiment score of ChatGPT-generated text is 0.92 with a standard deviation of 0.34.

Future work: Possible directions include expanding the dataset to include shorter text, such as one-sentence or one-paragraph length. We could also include non-academic text in our dataset depending on the intended purpose of the classifier. This could include phishing emails and spam bot texts for a phishing attack detector, or text from journals (both written by humans and AI-generated) for the detection of AI-generated websites containing misinformation. We could also investigate if other metrics could be useful in a classifier.

Just for kicks: When feeding the above into our model, we find that the probability this executive summary text was Chat GPT created is 0.0 % (which is correct!). The text has a burstiness of 447.892