Ram Mukund Kripa

Perspectives on Computational Modeling

Winter 2024

Final Project Proposal

Generative Artificial Intelligence, especially those powered by Large Language Models, have captured the global imagination. They have proven useful at a large variety and volume of tasks. However, concerns over its impact on education, specifically in its use in generating essays on behalf of students, have arisen. Several competitions have been hosted to address this need (Kaggle). Hence, several researchers and companies have developed models to address the need of detecting LLM-generated text, specifically essays.

To solve this problem, researchers have tried a variety of approaches. Trained detectors like those used by ZeroGPT, GPTZero, Originality, Quill, etc. have obtained some level of traction in the market today. However, they are shown to be systematically biased against non-native English speakers (Zou et al). In an experiment carried out with real US 8th Grader essays and real TOEFL essays written by non-native English speakers, it was shown that the TOEFL essays were significantly more likely to be flagged as LLM-written by popular trained detectors. Furthermore, it is shown that by prompting Chat-GPT to “elevate the text with literary language”, the detection rate of these TOEFL essays is reduced, which is a damning indictment of modern GPT-detectors.

Recent advances in the field like DetectGPT (Mitchell et al) use probability curvature for zero-shot detection of machine-generated text and achieve a great level of accuracy in specific fields. However, this method is far more computationally expensive than trained detectors, and hence impractical to scale. Simple models in the past have shown good accuracy in predicting whether text is human or AI generated (Frohlig and Zubiaga), but one of their key assumptions is a lack of syntactic and linguistic diversity, a feature of non-native English writing, as demonstrated in Zou et al.

Hence, the question arises of whether lightweight and simple models can detect the use of LLMs in the generation of essays.

This research question will be answered in my final project. My data sources will be both the datasets used in 'GPT detectors are biased against non-native English writers' (Zou et al) and the dataset published in the 'LLM - Detect AI Generated Text' (Kaggle) competition. The first dataset (available here https://huggingface.co/datasets/WxWx/ChatGPT-Detector-Bias) consists of 91 real TOEFL essays written by non-native English speakers, as well as their counterparts after prompt engineering with the 'literary language prompt'. The dataset also contains 88 US 8th Grade essays and 70 college essays written by native English speakers. The second dataset (available here https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data?select=train_essays.csv) consists of 1378 essays. Though their origin is unknown, it is reasonable to assume that this dataset consists mainly of essays written by native English speakers. The two datasets will be appended and split into train, validation, and test datasets which I will use to train my models.

The data are labeled as Human and AI generated, and the prompts attached to them are also available. Model architectures both with and without the prompt will be tried to see which will yield greater accuracy. The only explanatory variable is the text, but it will be split into several variables using tokenization, stemming/lemmatization, and other natural language processing techniques. The outcome variable is a classification between human and AI generated. Some models will also output a probability associated with the likelihood of being human or AI generated. Hence, it is a supervised learning task.

References

Fröhling L, Zubiaga A. 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. PeerJ Computer Science 7:e443 https://doi.org/10.7717/peerj-cs.443

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 'GPT Detectors Are Biased against Non-Native English Writers'. Patterns (N. Y.) 4, no. 7 (July 2023): 100779.

Mitchell, Eric, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 'DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature'. arXiv [Cs.CL], 2023. arXiv. http://arxiv.org/abs/2301.11305.

Kaggle. https://www.kaggle.com/competitions/llm-detect-ai-generated-text