RSM8521: Final Project Report

I. Introduction and Motivation

The advance of AI driven tools such as ChatGPT has provided students with numerous benefits such as the fast access to information, language learning and most importantly, the assistance on essay writings (Rejeb et al., 2024). On one hand, the introduction of these technologies has greatly improved the student's productivity via the tool's capability to extract the most relevant information within minutes and enhanced the student's learning experience. However, the wide use of these tools such as the GPT has also resulted in the exploitations that lead to violations of academic integrity. Therefore, the primary objective of this project is to develop a sophisticated analytical model that can accurately differentiate between essays written by humans and those generated by AI and thereby assisting institutions in preserving academic honesty.

II. Background: the Dataset and the BERT Model

<u>The Dataset</u>

The dataset used in this project is a dataset comprising essays written by middle and high school students alongside text generated by various Large Language Models. This dataset contains 1,375 documents that are labeled as human-written and 3 as AI-generated.

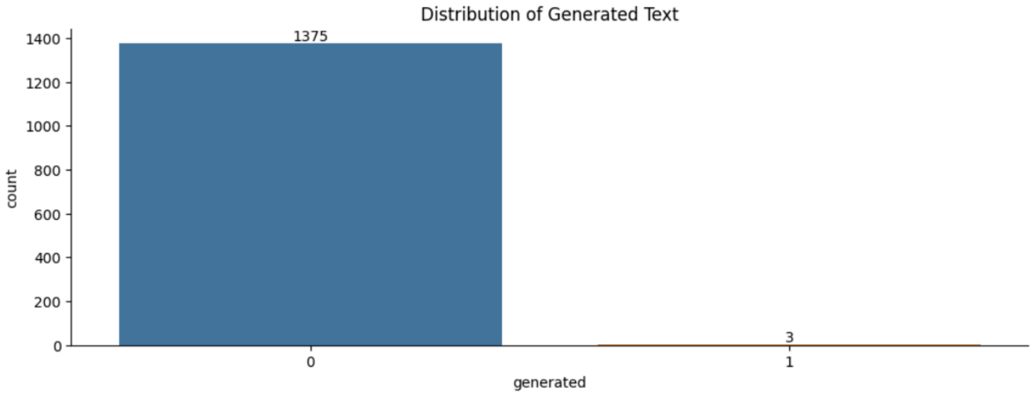

Exhibit 1: Distribution of the Generated Text from the Original Dataset

Since the current dataset is heavily unbalanced with only 3 essays labeled as AI-generated, an additional dataset from Kaggle, which contains another 44206 observations, has been added to the dataset in order to balance the dataset and to enhance the model's learning capabilities. Exhibit 2 shows the distribution of the generated text from the final dataset.

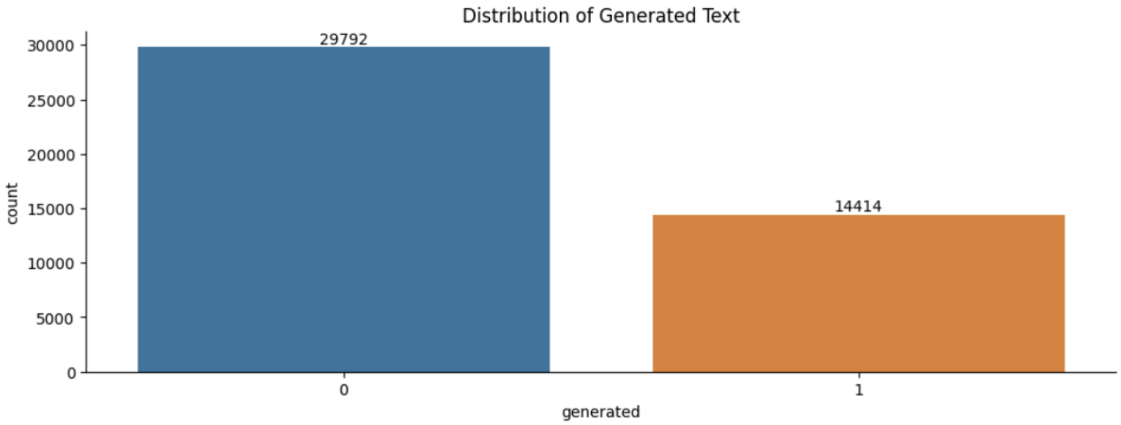

Exhibit 2: Distribution of the Generated Text from the Final Dataset

Additional Exploratory Data Analysis

Additionally, the number of words in each essay in the training set was counted and visualized in the following histogram:

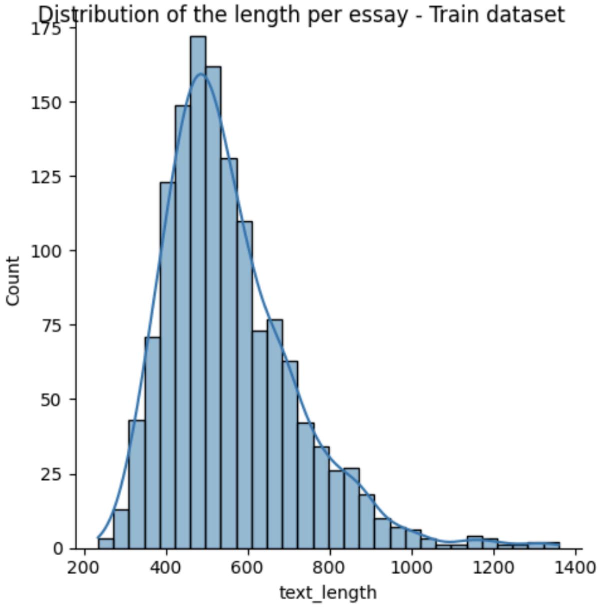

Exhibit 3: Distribution of the Length for each Essay in the Training Set

Furthermore, the cutoff point for outliers was calculated to be about 716. This implies for any essays that were longer than 716 words, they would be identified as a “long” essay.

The BERT Model

The primary model used for this project is the Bidirectional Encoder Representation Transformers (BERT) model. Unlike the traditional language models that process text in a single direction manner (i.e. from left to right or vice versa), BERT reads the entire sequence of the input text all at once, and hence the name: bidirectional. This property allows the model to learn the context of any given words based on its neighboring words and therefore enhances the model's understanding of any subtle and complex features in the given text such as tone and any implicit meanings. Furthermore, the BERT architecture includes a self-attention mechanism, which computes representations of its input and output without using sequence-aligned RNNs or convolution neural networks. The inclusion of the self-attention mechanism allows the BERT model to consider the entire sentence context, as opposed to only the words that came before it (as in traditional RNNs) or local neighborhoods (as in convolutional neural networks).

III. Description of the Method: The DistilBERT Model

The DistilBERT model is a distilled form of the BERT model. Fundamentally, the DistilBERT model leverages the technique of knowledge distillation during the pre-training phase to create a streamlined version of BERT, which reduces the size of the model by 40%. This process involves training a "student" model to emulate the output distributions of the "teacher" model, BERT, thus transferring its knowledge despite the reduction in complexity and size. Moreover, the DistilBERT model uses a 6-layer Transformer architecture in comparison to the 12-layer used in BERT. This reduction not only makes DistilBERT lighter and faster, enhancing its utility on devices with limited processing power, but also preserves about 97% of BERT's performance on language understanding benchmarks.

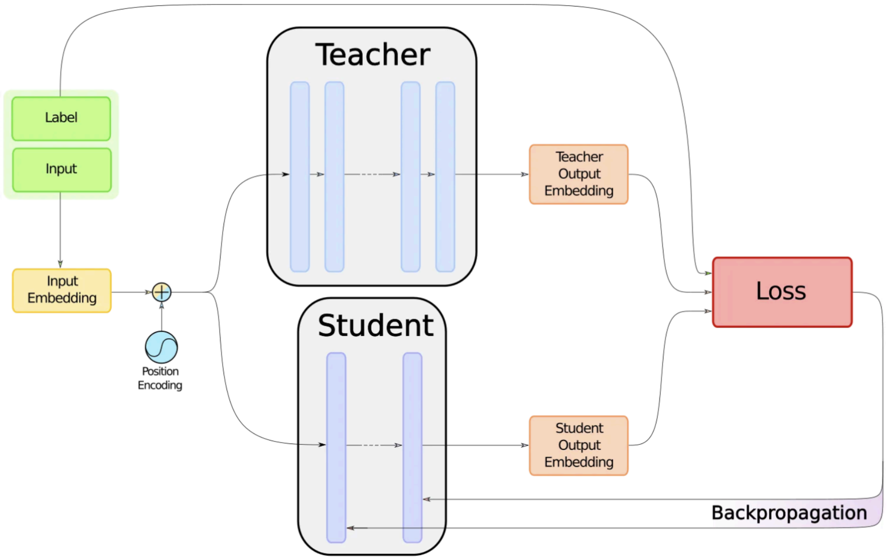

Exhibit 4: The General flow of distillation for BERT-like models[1]

Knowledge Distillation

The fundamental idea behind Knowledge distillation involves training a smaller, student model to reproduce the behavior of a larger, teacher model (Sanh et al., 2019). The training procedure involves training the student model so that the distillation loss is minimized. Specifically, the full distillation loss is composed of three main parts:

[1] Image Source: "Distillation of BERT-Like Models: The Theory", Reboul

$$L_{distillation} = \frac{L_{problem} + L_{cross} + L_{cosine}}{3}$$

where:

1. $L_{problem}$: Classic Loss

   This refers to the loss function which the teacher trained. Since the student model and the teacher model are composed of the same attention layers with the same problem-specific head, the loss can be calculated directly by plugging in the student's embeddings and labels.

2. L_{cross}: Teacher-Student Cross Entropy Loss

   The loss focuses on reducing the gap between the student's and the teacher's probability distribution. The calculation of the loss is given as follows:

   $$L_{cross} = \sum_{i=1}^{n} t_i * log(s_i)$$

   where:

   - $T(x) = (t_1,..., t_n)$: the probability distribution outputted by the teacher model after a forward pass for a given input x
   - $S(x) = (s_1,..., s_n)$: the probability distribution outputted by the student model after a forward pass for a given input x

3. L_{cosine}: Teacher-Student Cosine Loss

   - Unlike the cross-entropy loss that is focused on bringing the probability distribution between the student and teacher closer, the focus of the cosine loss is merely to align hidden vectors in teacher and student models. Specifically, the cosine loss can be calculated using:

     $$L_{cosine} = 1 - cos(T(x), S(x))$$

     where $T(x)$, $S(x)$ are defined as above.

Temperature

One last note on DistilBERT is that the DistilBERT model uses the notion of **temperature (θ)**, which in essence is used to help soften the softmax. The temperature is a variable $\theta \geq 1$ which lowers the 'confidence' of a softmax as it goes up. The normal softmax layer has a temperature of 1. Both the teacher and the student model should have the same temperature during the training process and the temperature should be set to 1 during the inference process.

IV. Experiments

Experimental Setup & Implementation

Below showcases the experiments conducted to distinguish between human-written essays from LLM-generated essays that involve training and evaluating the effectiveness of the BERT-based model. In this experiment, we configured a text classification model using the DistilBert architecture, which is a simplified version of the more complex BERT model, tailored for natural language processing tasks. The keras_nlp.models API helps cover the complete user journey of converting strings to tokens, tokens to dense features, and dense features to task-specific output.

For our task we have set the input sequence length to 512, which is the maximum for DistilBert.The implementation began with the application of the Keras DistilBertPreprocessor layer, which was utilized to tokenize and normalize the text inputs. This preprocessor prepares the data by creating a dictionary with keys "token_ids," "segment_ids," and "padding_mask." These keys correspond to the inputs required by a DistilBert model, facilitating direct integration.

Following preprocessing, we then employed a pre-trained DistilBert model from Keras specifically for English text. This adds a classification layer on top of the DistilBert backbone, allowing us to classify the essays into human or AI-generated. This configuration allows the model to transform backbone outputs into logits, which are then used in classification tasks. The model's parameters in the backbone (feature extraction part) are frozen here to prevent updating during training, optimizing the model's learning on our specific classification task with a modified learning rate. This setup aims to efficiently leverage pre-trained knowledge while adapting to new classification challenges.

After setting up our DistilBert text classification model, we prepared our dataset for training and testing. The dataset consisting of essays and their corresponding labels was divided into training and testing subsets. Specifically, 67% of the data was allocated for training to fine-tune the model, and 33% was reserved for testing to evaluate the model's performance.

The model was trained with the following specifications:

- Optimizer: Adam with a learning rate of 5e-4, we tested rates from 2e-5 to 5e-5 to find the best balance between speed and accuracy.
- Epochs: Trained for 1 to 4 epochs to observe performance saturation and resulted in training for 1 epoch.
- Batch Size: Experimented with sizes of 16, 32, and 64 to evaluate the impact on training dynamics and memory usage and finalized with 64.

Experiment 1

For our initial experiments, the model was trained on a dataset consisting of 1,000 data points to gauge performance along with the specifications mentioned above. This training session lasted approximately 30 minutes, resulting in an accuracy of around 80%. The results obtained serve as a baseline from which further improvements are sought through like additional training of data to capture complex patterns, mitigate bias and to improve generalization and below are a few reasons why adding more data points could boost performance.

- With only 1,000 data points, the model has limited examples to learn from. BERT and similar models thrive on large datasets, which provide a diverse range of examples and contexts
- Training on a small dataset can lead the model to memorize rather than generalize from the training examples.
- If the initial 1,000 data points aren't representative of the broader range of language usage for our classification task, the model's accuracy can be limited. Expanding the dataset can ensure a more representative sample of the linguistic and contextual diversity.

Experiment 2

The next phase of our experiment involved expanding the dataset with more diverse examples to enhance its accuracy. The expanded training improved accuracy to 92.12% on the test set, demonstrating the model's effectiveness in distinguishing between human and AI-generated texts. We evaluated the model using F1 score and confusion matrix.

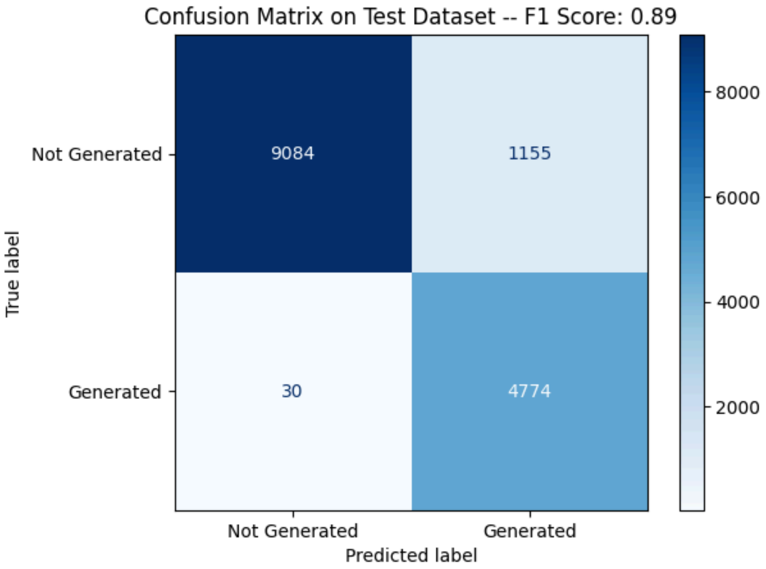

Exhibit 5: Confusion Matrix distinguishing between Human & AI generated essays

The F1 score of 0.89, indicating high precision and recall. The high precision and recall suggest that the model is reliable in its classifications, with minimal false positives and negatives and shows effective learning of the model. These results demonstrate the effectiveness of DistilBert in detecting AI-generated text, with significant potential for applications in educational settings.

We also used the NLTK library to download resources like stopwords and tokenize texts to create frequency distribution of words from AI-generated and human written texts, we further normalized these frequencies and visualized the differences using a word cloud.

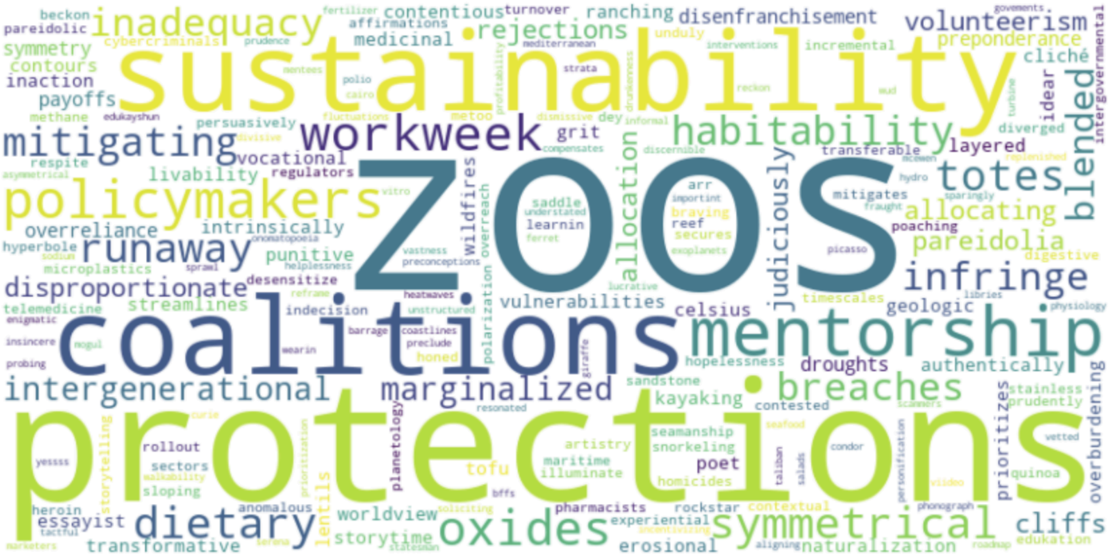

Exhibit 6: Frequently seen words in AI-generated essays

<u>Limitations</u>

While DistilBERT performed well, it is crucial to acknowledge limitations such as:

- **Susceptibility to Intricate Prompt Engineering**: Hard to distinguish the AI generated texts from human-written texts if LLM is prompted to write in simple words
- **Dependency on Data Quality**: Performance is highly contingent on the quality and variety of the training data.
- **Computational Resources**: Although lighter than BERT, DistilBERT still demands substantial computational resources, particularly when training on large datasets.

External Use-Cases

- **Content Moderation**: Social media platforms and online forums could use such tools to identify and filter out AI-generated content that may spread misinformation, spam, or malicious content disguised as genuine user communications.
- **Security**: In cybersecurity, detecting AI-generated phishing emails or other malicious communications can prevent scams that often mimic legitimate user interaction.

V. Conclusion

This project successfully demonstrates the application of DistilBERT model in differentiating between human-written and AI-generated essays, which is pivotal for maintaining academic integrity. The experiments highlighted the effectiveness of this model with a significant improvement in accuracy from the initial training to the expanded dataset experiments. Key findings include the importance of a balanced and diverse dataset and the model's ability to adapt and learn from the nuances of human vs. AI-generated text. The project not only confirms the robustness of DistilBERT in an educational context but also underscores the ongoing need for substantial computational resources and high-quality data to optimize performance.

VI. References

Manojvd. (n.d.). *MANOJVD/detect-ai-generated-text: Classified essays as AI generated (chatgpt) or human written using 5000 essay dataset. • requirement analysis • collected 2500 human written essays and 2500 AI generated essays (from API call). • finetuned bert and LSTM models for the supervised classification task.* GitHub. https://github.com/manojvd/Detect-AI-Generated-Text

NLP2CT. (n.d.). *NLP2CT/LLM-generated-text-detection: A survey and reflection on the latest research breakthroughs in LLM-generated text detection, including data, detectors, metrics, current issues and future directions.* GitHub. https://github.com/NLP2CT/LLM-generated-Text-Detection

Reboul, R. O. (2021, December 10). *Distillation of bert-like models: The theory*. Medium. https://towardsdatascience.com/distillation-of-bert-like-models-the-theory-32e19a02641f

Rejeb, A., Rejeb, K., Appolloni, A., Treiblmaier, H., & Iranmanesh, M. (2024). Exploring the impact of CHATGPT on education: A web mining and machine learning approach. *The International Journal of Management Education*, *22*(1), 100932. https://doi.org/10.1016/j.ijme.2024.100932

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv, abs/1910.01108*.

Team, K. (n.d.). *Keras Documentation: Getting started with Kerasnlp*. https://keras.io/guides/keras_nlp/getting_started/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vidhi1290. (n.d.). *VIDHI1290/LLM---detect-ai-generated-text: AI-generated text detection: A bert-powered solution for accurately identifying AI-generated text. seamlessly integrated, highly accurate, and user-friendly.*🚀. GitHub. https://github.com/Vidhi1290/LLM---Detect-AI-Generated-Text