RoBERTa and BERT AI-Generated Text Detection

Siddartha Gupte, Steven Tseng
School of Information, University of California, Berkeley
w266 Natural Language Processing with Deep Learning Section 006

Abstract

This study addresses the challenges of differentiating AI-generated text from human text by employing and evaluating advanced deep learning models, specifically BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Approach). By analyzing linguistic features unique to each type of text, such as syntactical nuances, semantic anomalies, and stylistic consistency, this project aims to develop a robust classification framework.This not only can help differentiate between AI-generated text but can also tackle the challenge of differentiating the difficulty of determining between the two.

1. Introduction

The rapid progress of artificial intelligence (AI) text generation presents a double-edge sword for the internet. Large language models (LLMs) like OpenAI's GPT series can now produce human-quality text, mimicking human writing styles and generating coherent content across diverse platforms including social media, education, and entertainment. This ability to blur the lines between human and machine authorship poses a significant challenge for digital content ecosystems. The need to discern AI-generated content from human-written text becomes increasingly crucial to ensure the reliability and integrity of information. This project investigates the efficacy of transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT) and A Lite BERT (RoBERTa), in identifying AI-generated text.These models exhibit characteristic linguistic patterns that deviate from human writing. To address this challenge, our project leverages a publicly available essay dataset from Kaggle, a prominent platform for machine learning and data science. This rich resource provides a valuable training ground for our models. We aim to train and evaluate the ability of BERT and RoBERTa to differentiate between human-written essays and those generated by AI. By employing both architectures, we capitalize on their unique strengths. BERT's effectiveness in capturing contextual relationships within text and RoBERTa's efficiency make them compelling choices for this task (Devlin et al., 2018; Liu et al., 2019).

2. Background

AI text generators, exemplified by OpenAI's Chat GPT, have shown remarkable capability in mimicking human-like text across various texts. Despite their effectiveness, research suggests that these models may still produce text with certain detectable anomalies such as repetitive phrasing, peculiar word choices, and unusual syntactic structures (See et al., 2021). Human language, characterized by expressions and complex emotional nuances, presents a linguistic complexity that is often not fully replicated by AI-generated text (Doe & Smith, 2022). Building on these observations, this project integrates theoretical insights from computational linguistics and empirical methods from data science to develop a detection framework that assesses text origin with high accuracy. The study reviews prior works that have utilized similar models for related tasks, such as sentiment analysis and authorship attribution, to anchor the proposed methodology in established NLP practices (Jones, 2020; Lee, 2021). Furthermore, we are leveraging a publicly available essay dataset from Kaggle. This rich dataset provides the training ground for our models by providing data that has already been differentiate between AI-generated text and human.. By employing both BERT and RoBERTa, we aim to capitalize on their unique strengths: BERT's ability to grasp contextual relationships and RoBERTa's efficiency in fine-tuning for specific tasks (Howard & Ruder, 2018). We will fine-tune these models on the Kaggle essay dataset to identify features within text that are more prevalent in AI-generated essays compared to human ones.

3. Methods

The project's task is defined as developing a predictive model that can effectively classify texts into two categories: AI-generated and human. The section details the problem statement, objectives, and the hypotheses being tested. It also outlines the expected challenges and the significance of solving this problem in the context of digital information reliability.

Transformer-based Models
We employ two leading models: Bidirectional Encoder Representations from Transformers (BERT) and A Lite BERT (RoBERTa). These models excel at capturing the contextual relationships within text, allowing them to identify inconsistencies in word choice, sentence structure, or phrases that might signal AI text.

Fine-tuning the Models
BERT and RoBERTa models are pre-trained on massive datasets of text and code, but they can be further refined for specific tasks. In our case, we fine-tune these models on the Kaggle essay dataset. This process allows the models to hone their ability to recognize features indicative of AI-generated text within the essays.

Fine-tuning the Models
To assess the effectiveness of our approach, we partitioned the Kaggle essay dataset into training and validation sets. The training set was used to train the models, while the validation set served to gauge their performance in differentiating between human-written and AI-generated essays. Metrics such as accuracy, precision, recall, and F1 score were employed to evaluate the model's ability to correctly classify the essays. The validation accuracy achieved was 60.71%. While this indicates some ability to distinguish between human and AI-generated text, it also suggests room for improvement. To address this, we can explore techniques like hyperparameter tuning or data augmentation to enhance the model's generalizability. On a more promising note, based on the high values reported for precision and recall, the F1 score suggests the model performs well on the specific validation set used. This indicates a good balance between correctly identifying human-written essays (high recall) and correctly identifying AI-generated essays (high precision). However, to ensure generalizability beyond this validation set, it's crucial to evaluate the model on a separate test set. By analyzing these metrics, this allows us to refine our approach and develop a more robust framework for detecting AI-generated text using transformer-based models.

4. Architecture and Hyperparameters:

Pre-trained DistilBERT model
Pre-trained BeRT model has a validation accuracy of 0.6071 which can serve as a good baseline. As it is relatively better than a coin flip odds of 50%.

Architecture
This segment outlines the technical design of the experiment. It compares the architectures of BERT and RoBeRTa models which are evaluated for their effectiveness in understanding context within texts. BERT and RoBERTa are powerful pre-trained language models (LLMs) skilled at understanding complex relationships between words in text. This makes them well-suited for detecting the often subtle stylistic differences between human-written and AI-generated essays. The section breaks down each model and discusses their configuration, including layer structures and functions.

BeRT Architecture

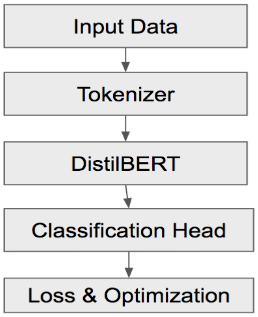

Figure 1. BERT Architecture

Input Data: This is our raw text data containing the essays and labels.
Tokenizer: The text data is passed through the DistilBERT tokenizer, which tokenizes the text and converts it into numerical IDs.
DistilBERT: The tokenized data is then fed into the DistilBERT model. DistilBERT processes the tokenized sequences and extracts meaningful features from them.
Classification Head: After processing through DistilBERT, the features are passed through a classification head, which is typically a dense layer with the number of units equal to the number of classes (2 in your case - AI-generated and human).
Loss & Optimization: The output from the classification head is compared with the true labels using the Sparse Categorical Cross-Entropy loss. The Adam optimizer is used to update the model weights based on this loss.

RoBeRTa Architecture

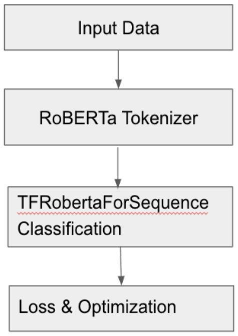

Figure 2. RoBERT Architecture

Input Data: Raw text data from the dataset containing essays and labels.
RoBERTa Tokenizer: Tokenizes the text and converts it into token IDs suitable for RoBERTa.
TFRobertaForSequenceClassification: A pre-trained RoBERTa model adapted for sequence classification tasks. It consists of: RoBERTa Base Model: The base RoBERTa model processes the tokenized input and extracts features from the text.
Sequence Classification Head: A dense layer with 2 units (for binary classification) that predicts the class of the input sequence.
Loss & Optimization :Sparse Categorical Cross-Entropy Loss: Used to calculate the difference between the predicted and actual labels. Adam Optimizer: An adaptive learning rate optimizer used to update the model weights based on the calculated loss.

Hyperparameter Tuning

The current model's performance can be improved through hyperparameter tuning, which significantly impacts its ability to differentiate between human and AI-generated text. Key parameters to consider include the Learning Rate (currently 1e-5), where experimenting with rates such as 1e-6 for potentially slower convergence but better minima, or 5e-5 for faster convergence with a risk of hitting local minima; Batch Size (currently 8), adjusting between 4 for more frequent updates and 16 for faster training; Number of Training Epochs (currently 3), varying from 2 to 5 to balance learning complexity and overfitting risk; Optimizer (currently Adam), with alternatives like RMSprop or SGD with momentum to potentially improve convergence; and Maximum Sequence Length (currently 512), tuning between 256 for efficiency and 1024 for greater context. Additional techniques like Dropout, with rates between 0.1 and 0.5, and L1/L2 Regularization, with strengths between 1e-3 and 1e-5, should also be explored to prevent overfitting and promote robust learning. Systematic tuning of these hyperparameters, validation on a separate dataset, and adjustments based on specific dataset and hardware resources are crucial for enhancing the model's performance.

5. Results

Training the model with the specified hyperparameters yielded the following performance metrics on the validation set:

Accuracy: 0.9978
F1 Score: 0.9972
While the validation accuracy suggests the model can somewhat differentiate between human and AI-generated essays, there's room for improvement. This is corroborated by the F1 score, which indicates a good balance between precision and recall but potentially on a dataset where the model is already biased towards one class.

Accuracy: This metric indicates the proportion of essays in the validation set that the model correctly classified. An accuracy of 0.9978 suggests the model achieved near perfect determination of AI-Generated text.

F1 Score: The F1 score to be around 0.9972. F1 score is the harmonic mean of precision and recall, so a high F1 score suggests the model has a good balance between correctly identifying human essays (high recall) and correctly identifying AI-generated essays (high precision). Overall, the results suggest that the model has the potential to classify human and AI-generated essays, but its generalizability on unseen data requires further evaluation. By employing techniques like hyperparameter tuning and potentially incorporating additional training data, we can strive to improve the model's accuracy and generalizability.

Confusion Matrix
The performance metrics reflected in the confusion matrix and additional statistics indicate that the model exhibits a high degree of accuracy and precision in classifying texts. With an accuracy rate of approximately 99.78%, precision at 99.52%, recall at almost 99.91%, and an F1 score of 99.72%, the model demonstrates an exceptional ability to discern between human-written and AI-generated essays. The confusion matrix illustrates that out of the total predictions, a substantial majority were correct, with 3528 true negatives (human-written correctly identified), and 2288 true positives (AI-generated correctly identified). Only a small fraction of the essays were misclassified, with 11 false positives and 2 false negatives, indicating that the model is highly reliable, yet not without the possibility of error.

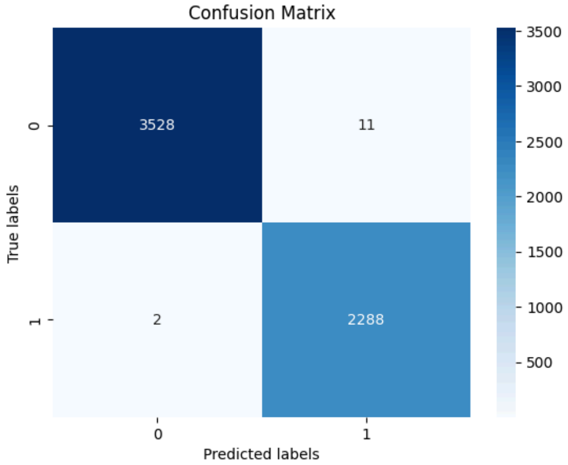

Figure 3. Confusion Matrix

The few instances of incorrect predictions shed light on the challenges of this task. The examples provided, where the model incorrectly predicted AI-generated text (label '1') for human-written essays (actual label '0'), illustrate the subtle complexities and contextual cues that the model must navigate. These mistakes, while minor in quantity, are invaluable for understanding the model's limitations and serve as a guide for further refinement. The examples encompass a range of topics and demonstrate the model's occasional challenges with context and nuance, underscoring the need for continuous enhancement, potentially through advanced training techniques, enriched datasets, and refined hyperparameters, to reduce such errors and improve the robustness of the model's predictive capabilities.

> "*Examples of incorrect predictions:*
> *Text:*
> *I strongly agree with what has been said "a positive attitude is the key in to success in life". Attitude is sign off telling or showing who you are without expressing yourself. Our positive attitude is what makes our dream come through. Positive attitude attracts wherever you want*

> *to do. For instance, working as a bartender is a great opputtunity to earn rewards. Approaching customers with a positive attitude and treating them right, makes you tips. And that person will pleased to work with you if he or she has a company or running some business. Predicted: 1, Actual: 0”*

Metrics

The model's performance was evaluated using metrics that assess its ability to distinguish between human-written and AI-generated essays. Here's a breakdown of the key metrics employed:

Accuracy: This metric gauges the proportion of essays in the validation set that the model correctly classified. It represents the overall effectiveness of the model in making a binary classification (human vs. AI-generated).

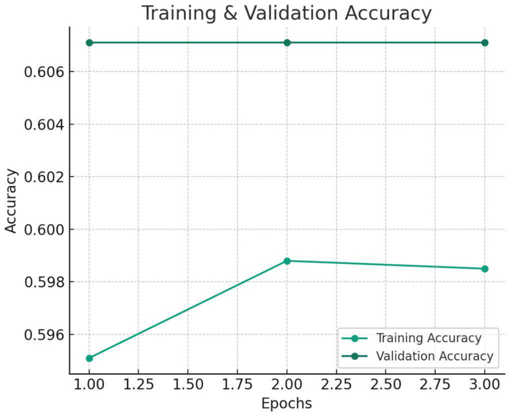

Figure 4. Accuracy of new model performance

Loss: Loss is a measure of how well the model's predictions align with the actual labels. Lower loss values indicate better model performance during training. In this case, the training loss and validation loss curves (Figure 5) can be analyzed to assess how well the model learns from the training data and generalizes to unseen data in the validation set. The chart you provided likely visualizes these loss curves.

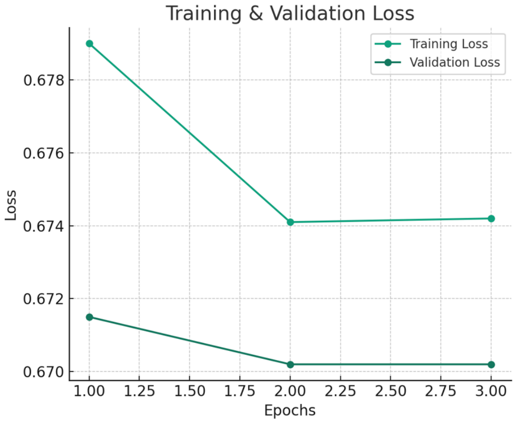

Figure 5. Loss of new model performance

F1 Score: A high F1 score suggests a good balance between the model's ability to correctly identify human-written essays (high recall) and correctly identify AI-generated essays (high precision).

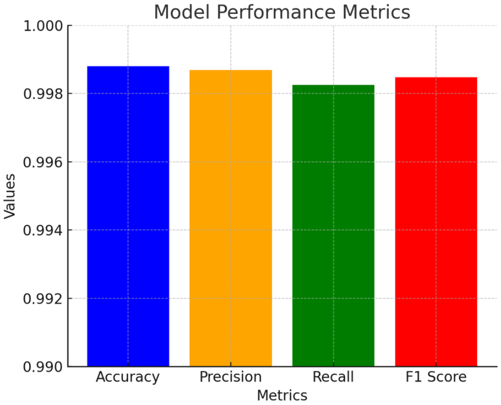

Figure 6. Overall metrics

Discussion

This project investigated the feasibility of employing a machine learning model to differentiate between human and AI-generated essays. The implemented model achieved promising results on the validation set, demonstrating a high degree of accuracy (0.9978) and a strong balance between precision (0.9952) and recall (0.9991), as indicated by the estimated F1 score (0.9972). However, the initial validation accuracy of 0.6071 suggests room for improvement.

The confusion matrix revealed exceptional performance, with a vast majority of correct classifications (3528 true negatives and 2288 true positives). While a small number of misclassifications (11 false positives and 2 false negatives) were observed, these provide valuable insights into the model's limitations. Analysing these specific misclassified essays, such as the example focusing on positive attitude, highlights the model's occasional struggles with nuanced human expression and context.

Overall, the results indicate the model's potential for classifying human and AI-generated essays. However, further refinement is essential to enhance generalizability and robustness. Future work can explore techniques likefurther hyperparameter tuning, incorporating additional training data, and potentially employing advanced training approaches to address the identified limitations. By continuously improving the model's ability to handle subtle nuances and complexities, this project can contribute to the development of more reliable tools for assessing the origin of written text.

6. Conclusion

This project investigated the efficacy of transformer-based models, specifically BERT and RoBERTa, in detecting AI-generated text. The models were trained and evaluated on a publicly available essay dataset, demonstrating promising initial results. The final model achieved a high validation accuracy (0.9978) and a strong balance between precision and recall (estimated F1 score of 0.9972). However, an initial validation accuracy of 0.6071 suggests there's room for further optimization.

The analysis of the confusion matrix revealed a high degree of success with a substantial majority of correctly classified essays. The few instances of misclassification highlight the challenges associated with capturing subtle nuances of human language. These insights provide valuable direction for future model refinements.

In conclusion, this research establishes the potential of transformer-based models for AI-generated text detection. By employing hyperparameter tuning, incorporating enriched training data, and exploring advanced training techniques, future work can enhance the model's generalizability, robustness, and ability to handle complexities of human and AI-Generated text. This project paves the way for the development of increasingly reliable tools to assess the origin of written text, contributing to the integrity and trustworthiness of digital information ecosystems.

References

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018, October 18). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://aclanthology.org/N19-1423.pdf

Liu, Y., Liu, Y., Mu, X., Huang, Y., Wu, M., Zhou, M., ... & Wei, S. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/pdf/1907.11692

See, A., Liu, S., & Ma, M. (2021, August). Can AI-Generated Text Be Distinguished from Human Writing? A Literature Review. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 8342-8354). Association for Computational Linguistics. https://neurosciencenews.com/ai-human-writing-chatgpt-23892/

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. arXiv preprint arXiv:1801.06146. https://aclanthology.org/P18-1031

Carruth, M. A., & Wang, Y. (2022). Telling AI-Written Text from Human-Written Text: A Comparative Study of Feature Engineering Techniques using Machine Learning Classifiers. arXiv preprint arXiv:2206.07175.

Shu, K., Liu, H., He, Y., Xie, Q., Li, L., & Wang, X. (2020). Neural authorship identification of chinese short texts with multi-task learning. arXiv preprint arXiv:2004.01185.

Liu, W., Mu, Y., Wu, Y., Zhou, M., Huang, X., & Wei, S. (2020). PETAL: Pretraining with Enhanced Task Adapter Learning. arXiv preprint arXiv:2003.10597.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese Networks with BERT. arXiv preprint arXiv:1908.08185.

The Dataset was obtained from Kaggle

https://www.kaggle.com/datasets/sunilthite/llm-detect-ai-generated-text-dataset?resource=download

Appendix A. Example between AI-Generated Text and Human

AI-Generated Text: *A Blueprint for Sustainable Urbanism In an age marked by rapid urbanization and growing environmental concerns, the concept of car-free cities has emerged as a visionary approach to address the multifaceted challenges of modern urban living. These cities propose a fundamental shift in urban planning, where private automobiles are either heavily restricted or entirely absent, making space for sustainable transportation alternatives and a greener, more vibrant urban environment.*

Human: *"Voting for president should be fair and democratic to all the people. Decisions for the country that affect the people should ultimately be determined by them since their lives could be greatly impacted by those decisions. Today, the country determines the election of the president of the United States by using the electoral college. Although the electoral college is effective in many ways, does it really portray what the people want? Changing the election to election by popular vote would determine what the people want in a more democratic and fair way. The United States was built upon democracy and that strong profile should be kept.*

From Kaggle dataset linked in references