AI-GENERATED TEXT DETECTION

Authors:
J. Ganga Sai Ram
S. Aprameya
M. Babji
Dr. P Ashok Babu

Abstract

The rise of AI-generated text presents significant challenges in distinguishing human-written text from AI-generated text. We use machine learning algorithm to address this challenge, utilizing the "AI vs. Human Text" dataset, which includes both AI-generated and human-written text. By employing a combination of BERT-based feature extraction and an XGBoost classifier, the system achieves high accuracy in differentiating between the two types of text. The model ensures consistent performance across various contexts. The trained model demonstrates exceptional performance, achieving a 97.7% accuracy rate in classification. This work contributes to the fields of AI ethics and content security by providing a reliable tool to detect AI-generated text, reducing risks across academia, journalism, and social media.

Key words: Text classification, Machine learning, BERT, Extreme Gradient Boosting (XGBoost), AI text detection

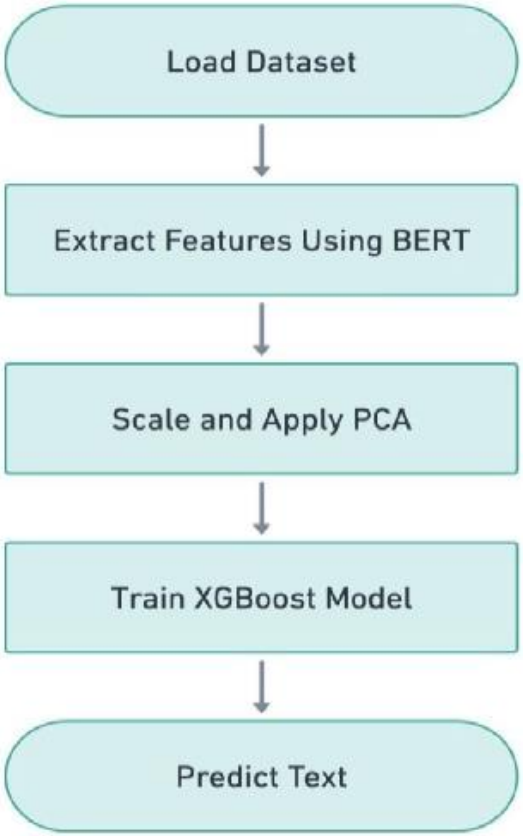
Introduction

- Models like GPT-3, GPT-3.5, and LlaMA have revolutionized NLP tasks (summarization, grammar correction, Q&A).
- AI-generated text is often indistinguishable from human writing, raising ethical, social, and economic concerns across multiple sectors.
- Issues like academic dishonesty and fake news propagation highlight the need for AI text detection methods.
- Standard plagiarism detectors fail to effectively identify AI-generated content, impacting education and research.
- This paper presents a technique using BERT for feature extraction and XGBoost for classification to detect AI-generated text.

Objectives

- To build an effective detection model for distinguishing between AI-generated text and human-written text.
- The model should generalize well across different domains and different types of text.

Methodology



FLOW CHART OF METHODOLOGY

Methodology

Dataset

- We used "AI Vs Human Text" dataset containing 15,000 text samples.
- The dataset contains two labels: Human-written is indicated using 0 and AI-generated is indicated using 1.

Feature Extraction Using BERT (Bidirectional Encoder Representations from Transformers)

- Feature extraction is performed by using the bert-base-uncased model.
- Each input text is tokenized using the BERT tokenizer. The tokenized text is then fed into the BERT model, where the last hidden state of the model's output is extracted.
- The mean pooling technique is applied to aggregate information across all token embeddings, resulting in a fixed-size feature vector.

Methodology

Dimensionality Reduction with PCA

- The features extracted using BERT are standardized using StandardScaler, which ensures that each feature has a mean of zero and a standard deviation of one.
- Next, PCA is applied with 100 components to transform the data into a lower-dimensional space while retaining as much variance as possible.

Model Training with XGBoost

- The XGBClassifier is initialized and trained on the PCA-transformed training data and its corresponding labels.

Results And Discussion

	Precision	Recall	F1-score	Support
Human	0.96	0.97	0.96	1521
AI	0.97	0.95	0.96	1479
Accuracy	0.97	0.95	0.98	3000
Macro avg	0.96	0.96	0.96	3000
Weighted avg	0.96	0.96	0.96	3000

Random Forest model Classification Report

	Precision	Recall	F1-score	Support
Human	0.97	0.98	0.98	1521
AI	0.98	0.97	0.98	1479
Accuracy	0.98	0.97	0.98	3000
Macro avg	0.98	0.98	0.98	3000
Weighted avg	0.98	0.98	0.98	3000

XGBoost model classification report

Metrics	XGBoost	Random Forest
Accuracy	0.97	0.96
Precision	0.98	0.97
Recall	0.97	0.95
F1 Score	0.97	0.96

Comparision of Model Performance Metrics

Conclusion

- The use of AI-generated text has introduced significant challenges in distinguishing between human and AI-generated content.
- Our paper has addressed this issue by developing an effective detection model using BERT for feature extraction and XGBoost algorithm for classification.
- The model we developed demonstrates an accuracy of 97% in identifying AI generated texts, providing a robust tool to reduce the risks associated with AI-generated content.
- As large language models continue to evolve, our approach offers a necessary safeguard to maintain the integrity and authenticity of written content.

THANK YOU