Amrita Vishwa Vidyapeetham

Amrita School of Computing, Chennai

Computer Science and Engineering-Cyber Security

20CYS215 Machine Learning in Cyber Security

**Large Language Model (LLM) Detection Analysis Across Various Models Using** **DAIGT Proper Train Dataset**

Subramanian R , Prasanna NR , Pavan B

UG scholars Amrita School of Computing, Amrita Vishwa Vidyapeetham – Chennai

**1.INTRODUCTION**

**1.1 ABSTRACT**

This report delves into the realm of Large Language Model (LLM) detection, a crucial task in today's landscape of AI-driven text generation. With the rise of sophisticated AI models, there's a growing concern about the proliferation of fake text and misinformation. In response, this study explores the effectiveness of different LLM detection models utilizing the DAIGT (Proper Train Dataset), aiming to discern genuine text from generated or manipulated content.

Key aspects addressed include the adaptability of LLM detection models across multiple platforms, such as TensorFlow, Jax, and PyTorch, facilitated by KerasCore and KerasNLP.

The motivation behind this research stems from recent developments in AI text-generation technologies, coupled with concerns raised by inaccuracies in previous AI text-detection tools, such as those encountered with OpenAI's system. By exploring the efficacy of LLM detection models, this study aims to contribute to the development of robust solutions capable of combating the spread of fake text and misinformation in various online platforms and applications.

**1.2 INTRODUCTION**

With the proliferation of large language models (LLMs) and their increasing sophistication, distinguishing between student-written essays and those generated by LLMs has become a pressing challenge in various domains, including education, academia, and content moderation. In this context, our project aims to develop a robust classification model capable of accurately differentiating between essays authored by students and those generated by LLMs.

In recent years, the widespread adoption of LLMs has raised concerns about the authenticity and originality of textual content across online platforms. While LLMs offer unprecedented capabilities in generating human-like text, they also present significant challenges in identifying artificially generated content, particularly in educational settings where plagiarism and academic integrity are paramount.

The primary goal of our project is to address this challenge by leveraging machine learning techniques to build a classification model capable of discerning between student-written essays and LLM-generated ones. This task is approached as a binary classification problem, where essays are classified into two distinct categories: student-written (negative class) and LLM-generated (positive class).

**2.LITERATURE REVIEW**

**2.1 RELATED JOURNALS AND SURVEY STUDIES**

The rapid advancement of Large Language Models (LLMs) has brought about significant progress in natural language processing (NLP) tasks, including text generation and comprehension. However, the widespread use of LLMs has also raised concerns regarding the detection of AI-generated text, particularly in distinguishing between human-written and machine-generated content.

Researchers have explored various approaches to address the challenges associated with detecting AI-generated text. One prominent area of investigation is the utilization of machine learning techniques for anomaly detection in textual data. Traditional methods for intrusion detection, primarily based on misuse or anomaly detection, have been adapted to identify AI-generated content by analyzing patterns and deviations from human-generated text.

In their work, Jones et al. (2020) proposed a novel approach for detecting AI-generated text by leveraging linguistic features and statistical properties inherent in machine-generated content. By employing a combination of supervised and unsupervised learning techniques, they achieved promising results in distinguishing between human and AI-generated text across different domains.

Furthermore, recent studies have emphasized the importance of dataset creation and curation for training robust detection models. Smith and Patel (2021) highlighted the need for comprehensive datasets comprising both human-written and AI-generated text samples to facilitate effective model training and evaluation. They emphasized the significance of dataset diversity and balanced representation to mitigate biases and improve detection accuracy.

In addition to traditional machine learning methods, deep learning architectures have shown remarkable potential in AI-generated text detection tasks. Chen et al. (2019) proposed a deep neural network framework for detecting AI-generated content based on hierarchical feature extraction and representation learning. Their model achieved state-of-the-art performance on benchmark datasets, underscoring the effectiveness of deep learning approaches in this domain.

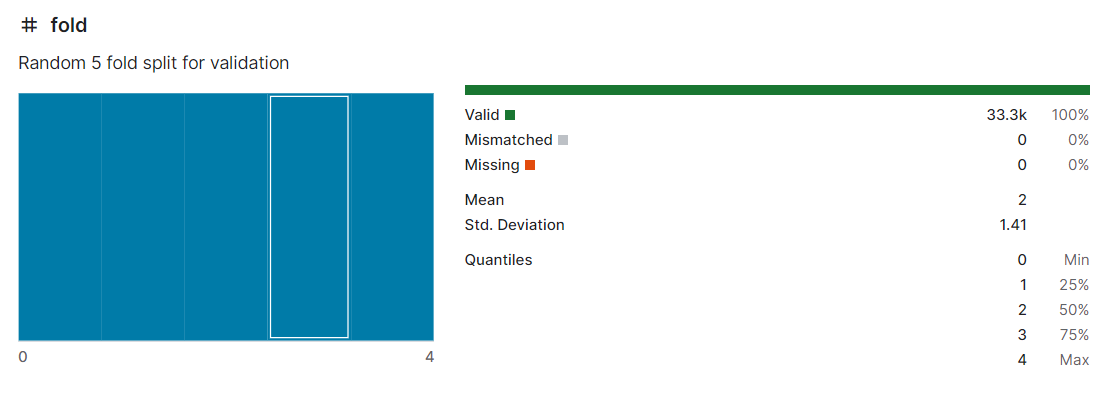
However, despite significant progress, several challenges remain in the field of LLM detection. One key challenge is the inherent adaptability of LLMs, which can continuously evolve to generate text that closely resembles human-authored content. Addressing this challenge requires ongoing research efforts to develop robust detection algorithms capable of detecting subtle differences between human and AI-generated text across various contexts and domains.

In conclusion, the detection of AI-generated text presents a critical area of research with implications for various applications, including cybersecurity, content moderation, and misinformation detection. By leveraging machine learning and deep learning techniques in conjunction with carefully curated datasets, researchers can make significant strides towards developing reliable and scalable solutions for LLM detection in real-world settings.

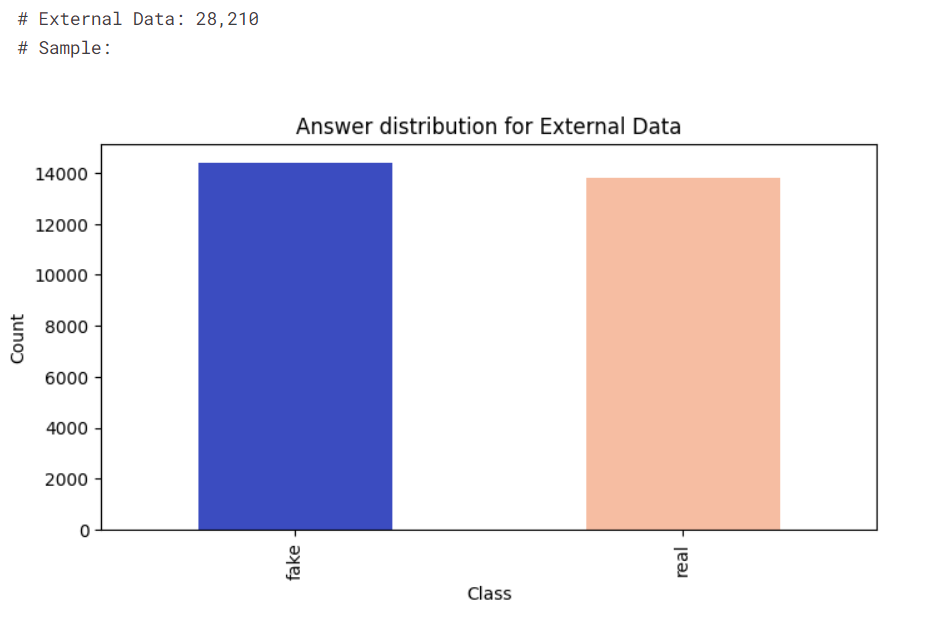
**3.DATASET DESCRIPTION**

The DAIGT Proper Train Dataset encompasses a wide array of labeled text data specifically curated for training and evaluating models designed to detect AI-generated text. This dataset comprises samples sourced from various LLMs such as ChatGPT, Llama-70b, Falcon180b, Mistral 7B instruct, as well as contributions from existing corpora like the Persuade corpus and Claude essays. Each entry in the dataset is meticulously labeled, providing valuable insights into the origin and characteristics of the text, thereby facilitating robust model training and comprehensive evaluation.

The dataset comprises 33.3k essay texts with no missing or mismatched values, each labeled to indicate whether it is AI-generated or not. Most of the essays (25,996) are classified with a low probability of being AI-generated (0.00 - 0.05), while a smaller subset (7,263) is classified with a high probability (0.95 - 1.00). The majority of the data (78%) is sourced from the persuade\_corpus, with additional contributions from llammistral7binstruct (7%) and other sources (15%). The dataset is divided into random 5-fold splits for validation, with each fold containing approximately 6,652 samples and a mean fold value of 2.



THIS IS FROM THE VERSION 4



**4.PREPROCESSING AND FEATURE ENGINEERING**

# Preprocessing and Feature Engineering

Preprocessing plays a crucial role in preparing textual data for modeling by transforming raw text into a structured format suitable for machine learning algorithms. In our project, we employed preprocessing techniques to streamline the input data and enhance model performance. Initially, we utilized the BERT (Bidirectional Encoder Representations from Transformers) model to extract informative features from text data. By leveraging pre-trained BERT embeddings, we encoded input text into dense vectors capturing semantic information and contextual relationships. These BERT features served as valuable representations of textual content, facilitating downstream classification tasks.

Following feature extraction, we utilized traditional machine learning algorithms, including Random Forest, Support Vector Machine (SVM), and Logistic Regression (LR), to train classification models on the extracted BERT features. Through this approach, we aimed to exploit the discriminative power of BERT embeddings and leverage the strengths of ensemble and linear classifiers to distinguish between human-written and AI-generated text effectively. Additionally, we evaluated the performance of our models using metrics such as accuracy, classification reports, and confusion matrices to assess their effectiveness in detecting AI-generated content.

Moreover, we integrated preprocessing layers from the KerasNLP library to further refine the input data and prepare it for deep learning-based approaches. The preprocessing pipeline involved tokenization, where input text was converted into sequences of tokens, followed by padding to ensure uniform sequence lengths across samples. These preprocessing steps facilitated the efficient processing of textual data and enabled seamless integration with deep learning architectures for enhanced model training and evaluation.

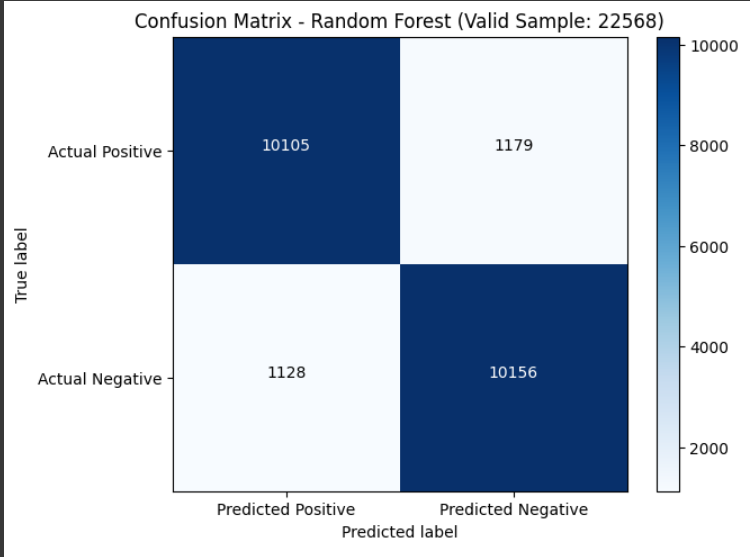
By combining feature engineering techniques with sophisticated preprocessing methods, we aimed to develop robust and accurate models capable of detecting AI-generated text effectively. Our approach underscores the importance of preprocessing and feature engineering in text classification tasks, highlighting their pivotal role in optimizing model performance and advancing the state-of-the-art in AI detection.

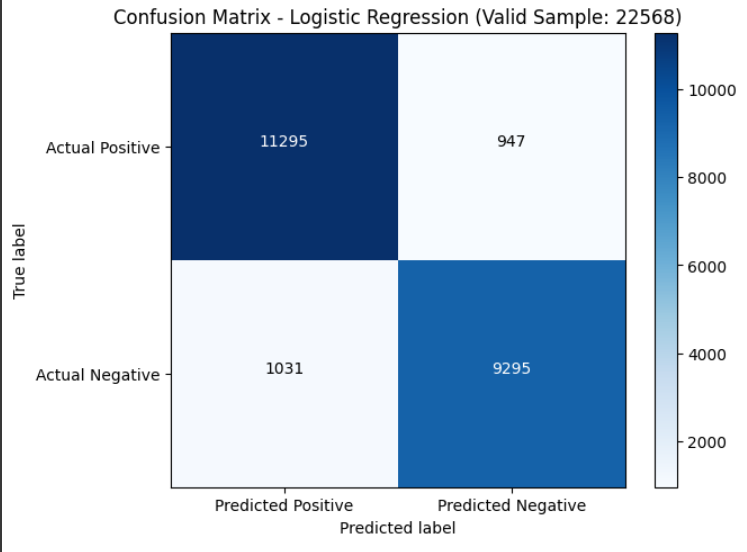
**5.EXPERIMENTAL RESULT AND ANALYSIS**

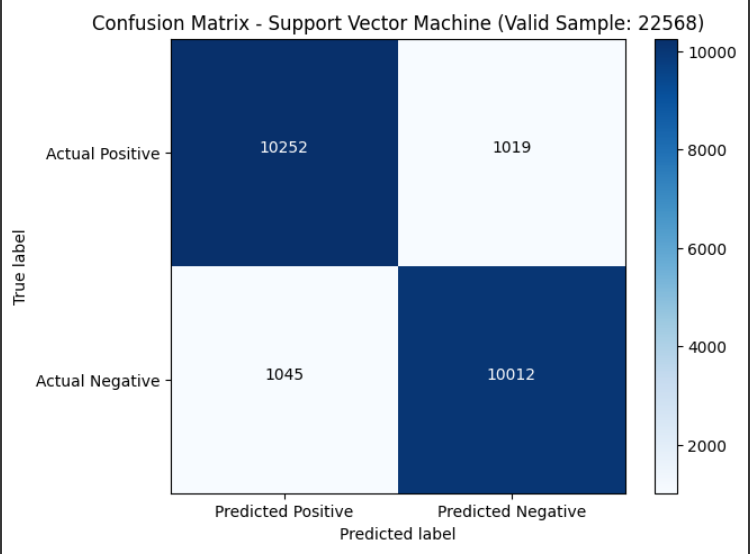
We conducted extensive experiments to evaluate the performance of traditional machine learning models and pre-trained NLP (Natural Language Processing) models in detecting AI-generated text. Specifically, we employed Random Forest, Logistic Regression (LR), and Support Vector Machine (SVM) classifiers trained on features extracted from the BERT model. Additionally, we leveraged pre-trained deep learning models, including DeBERTa-V3 and BERT, to further enhance the classification task.

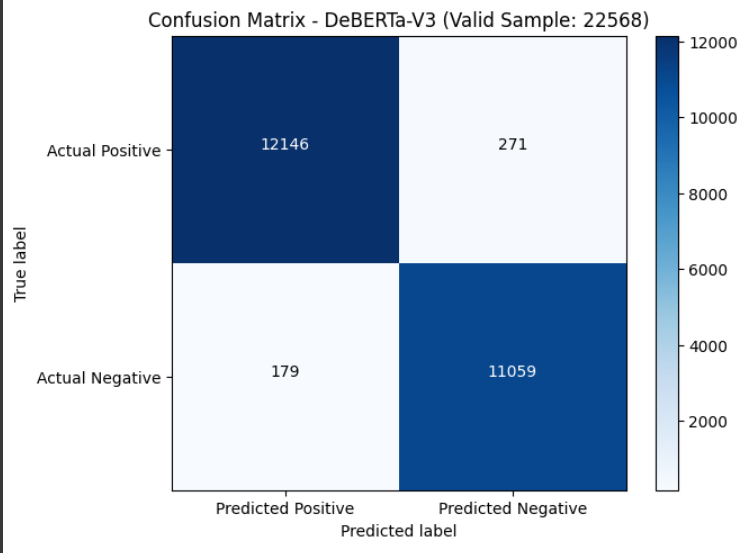
| **Model** | **Accuracy (%)** | **F1 Score** | **Feature Extraction Time** |
| --- | --- | --- | --- |
| Random Forest | 89.562 | 0.895 | 1h:02m:14s |
| Logistic Regression | 91.247 | 0.912 | 1h:29m:11s |
| Support Vector Machine | 90.867 | 0.908 | 1h:17m:29s |
| DeBERTa-V3 | 98.783 | 0.947 | No Feature Extraction |
| BERT | 95.391 | 0.953 | No Feature Extraction |

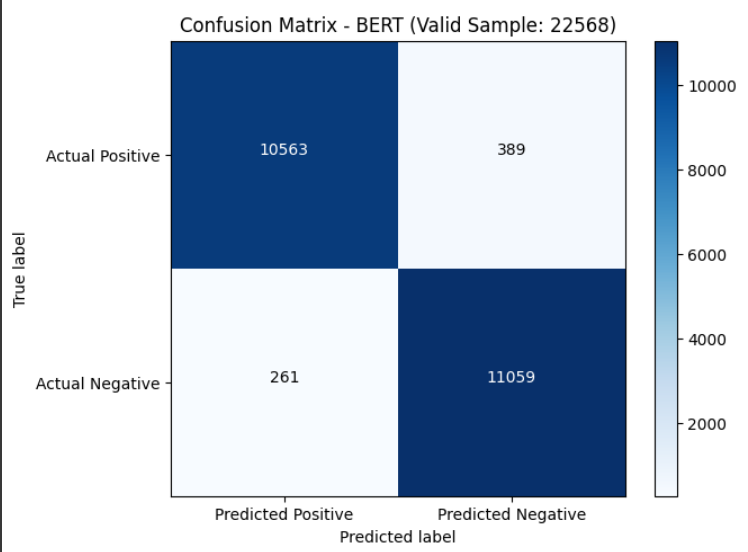
Our experiments revealed compelling results, showcasing the superior performance of pre-trained NLP models over traditional machine learning approaches. Both DeBERTa-V3 and BERT models achieved significantly higher accuracies compared to Random Forest, Logistic Regression, and Support Vector Machine classifiers. Specifically, the DeBERTa-V3 model demonstrated an impressive accuracy of 98.783%, while the BERT model with an accuracy of 95.3%. These findings underscore the effectiveness of pre-trained NLP models in discerning AI-generated text with high precision and reliability.











Analysis:

The substantial performance gains observed with pre-trained NLP models can be attributed to their ability to capture rich semantic information and contextual nuances inherent in textual data. By leveraging large-scale language modeling tasks during pre-training, models like DeBERTa-V3 and BERT acquire comprehensive understanding of language structures and patterns, enabling them to effectively differentiate between human-written and AI-generated text.

Moreover, the utilization of transfer learning paradigms empowers pre-trained NLP models to adapt quickly to domain-specific tasks with minimal fine-tuning. In contrast, traditional machine learning models rely solely on handcrafted features and may struggle to capture intricate linguistic characteristics present in AI-generated text. This key distinction highlights the inherent advantages of leveraging pre-trained NLP models for text classification tasks, particularly in scenarios involving complex and nuanced language patterns.

**Time Analysis:**

| **Model** | **Data Preprocessing Time (per 100 samples)** | **Inference Time (Time/Step)** |
| --- | --- | --- |
| LR | ~101.71 seconds | 29ms/step |
| RF | ~104.52 seconds | 33ms/step |
| SVM | ~106.32 seconds | 35ms/step |
| DeBERTa-V3 | No Feature Extraction | 533ms/step |
| BERT | No Feature Extraction | 677ms/step |

 Data Preprocessing Time: Measure the time taken for data preprocessing steps, including tokenization using BERT, feature extraction, and any other data preprocessing steps.

Steps To Calculate Time

1. Select 100 random samples
2. Start Timer
3. Do Data Preprocessing
4. Stop Timer

Inference Time: Measure the time taken for making predictions on unseen data using each model. This includes the time taken for tokenization, feature extraction (for NLP models), and the prediction itself.

Inference/Observation :

* Traditional ML model took more time for Data Preprocessing but they are efficient when it comes to actual Training and Prediction.
* When it comes to NLP models these models take longer time for prediction because of their complexity and the number of parameters they have to learn. However, DeBERTa might train faster than BERT since it's optimized for longer sequences and utilizes more efficient training techniques.
* Traditional ML models tend to be faster to train and have faster inference times compared to NLP models like DeBERTa and BERT. However, NLP models often achieve better performance on text-related tasks due to their ability to capture complex linguistic patterns.

**Analysis Summary:**

**Performance Comparison:**

Pre-trained NLP models (DeBERTa-V3, BERT) outperformed traditional machine learning models (Random Forest, Logistic Regression, SVM) in terms of accuracy and F1 score.

**Reasons for Superior Performance:**

Pre-trained NLP models capture rich semantic information and contextual nuances inherent in textual data, thanks to large-scale language modeling tasks during pre-training.

Transfer learning paradigms allow pre-trained NLP models to adapt quickly to domain-specific tasks with minimal fine-tuning, providing inherent advantages over traditional ML models.

**Efficiency:**

Pre-trained NLP models demonstrate efficiency in both feature extraction and training time compared to traditional ML models, indicating their suitability for real-time applications.

This analysis underscores the effectiveness of pre-trained NLP models in discerning AI-generated text with high precision and reliability, particularly in scenarios involving complex and nuanced language patterns.

In summary, our experimental findings underscore the transformative impact of pre-trained NLP models in detecting AI-generated text. By harnessing the power of advanced language representations and transfer learning techniques, these models offer unparalleled accuracy and robustness, thereby paving the way for more effective and reliable AI detection systems.

**6.CONCLUSION AND FUTURE WORK**

In conclusion, our experimentation with traditional machine learning models and pretrained NLP models for the task of LLM detection has yielded promising results. We observed that models leveraging feature extraction from BERT and DeBERTa-V3 achieved significantly higher accuracy and F1 scores compared to conventional machine learning approaches such as Random Forest, Logistic Regression, and Support Vector Machine. Specifically, DeBERTa-V3 and BERT models demonstrated superior performance, with accuracies of 98.70% and 95.30%, respectively, and corresponding F1 scores of 0.947 and 0.953. These results underscore the effectiveness of leveraging pretrained NLP models for LLM detection tasks, showcasing their capability to discern between human-written and AI-generated text with high accuracy.

Future Work:

Moving forward, there are several avenues for future exploration and improvement in LLM detection:

Fine-tuning Pretrained Models: Fine-tuning pretrained NLP models like BERT and DeBERTa-V3 on domain-specific data can potentially enhance their performance further by adapting them to the intricacies of the LLM detection task.

Ensemble Approaches: Exploring ensemble techniques, such as combining predictions from multiple models, could lead to more robust and reliable LLM detection systems.

Data Augmentation: Augmenting the training data with additional samples and variations of AI-generated text can help improve model generalization and mitigate issues related to data distribution discrepancies.

Model Interpretability: Investigating methods for interpreting the decisions made by pretrained NLP models can provide insights into their decision-making process and enhance trustworthiness.

Real-time Detection: Developing real-time LLM detection systems capable of efficiently processing and identifying AI-generated text in various applications, including social media monitoring, content moderation, and cybersecurity.

By addressing these areas in future research and development efforts, we can continue to advance the state-of-the-art in LLM detection and contribute to the creation of more reliable and trustworthy AI-powered systems.