Received Date:19th December, 2024
Revision Date: 29th December, 2024
Accepted Date:9th January, 2025

AI Content Detection

Prashjeev Rai[1*], Sadikshya Gyawali[2], Shuvechhya Bajracharya[3], Sparsh Nidhi[4], Sudeep Shakya[5]

[1]Dept. of Computer Engineering, Kathmandu Engineering College, E-mail: prashjeevrai@gmail.com
[2]Dept. of Computer Engineering, Kathmandu Engineering College, E-mail: sadikshya222@gmail.com
[3]Dept. of Computer Engineering, Kathmandu Engineering College, E-mail: shuvechhya206@gmail.com
[4]Dept. of Computer Engineering, Kathmandu Engineering College, E-mail: sparsh123nidhi@gmail.com
[5]Assoc. Professor, Computer Engineering, Kathmandu Engineering College, E-mail: sudeep.shakya@kecktm.edu.np

Abstract — AI (Artificial Intelligence) content detection is the task of predicting if the given content is written by humans or AI. This project is a detection tool aimed at eliminating issues created by AI-generated text content such as fake academic reports and papers, articles, news, misinformation, and propaganda by combining multiple detection methods. Three models, LSTM (Long short-term memory), BERT (Bidirectional Encoder Representations from Transformers), and distilBERT (distilled Bidirectional Encoder Representations from Transformers) were fine-tuned on a small labelled dataset of 2492 rows. After comparing their performances, distilBERT was selected for further refinement. Then, a pre-trained distilBERT model was finetuned with 24034 rows of collected datasets to get results specific to the intended application. The language models in AI text generators (e.g. GPT-2) often plagiarize from the training datasets. So, to increase the accuracy BERT Classifier-based plagiarism detector was integrated into the system to determine the originality of input text and predict the likelihood of plagiarism or AI generation. The final model had an overall accuracy of 95% on the unseen data with it being able to detect 100% of all AI content in the unseen dataset and correctly classifying 90% of AI text in the unseen dataset.

Keywords — AI-generated text content, BERT, DistilBERT, detection tools, GPT-2, LSTM, plagiarism.

Introduction

The advent of advanced AI language models, exemplified by technologies such as GPT-3, has brought about a transformative era in Natural Language Processing (NLP). However, the remarkable capabilities of these models raise pressing concerns regarding the authenticity and provenance of text generated by AI systems. Distinguishing between human-generated and AI-generated text has become a critical challenge, significantly impacting information credibility and source verification.

The growing necessity for AI text detection models stems from the convergence of sophisticated AI technologies and the increasing difficulty in differentiating AI-generated content from human-written text. AI-generated text, particularly from advanced language models, closely mimics human writing styles by learning from vast corpora of textual data. This convergence necessitates the development of robust methods for discriminating between human and AI-authored content, ensuring the integrity and reliability of textual information in various applications.

In this context, our research focuses on the design and implementation of AI text detection models to discern between human generated and AI-generated text. We explore statistical analysis, machine learning algorithms, and deep learning architectures as foundational methodologies for constructing these detection systems. Our work contributes to an evolving research area aimed at addressing the challenges posed by the proliferation of AI-generated content and its implications for information trustworthiness.

Furthermore, we highlight the critical issue of plagiarism, which is exacerbated by the rise of AI-generated content. Plagiarism detection is essential for upholding the principles of intellectual integrity, ensuring that credit is given to original authors and preventing the unauthorized use of intellectual property. Leveraging AI capabilities for plagiarism detection alongside AI text detection enriches our system's functionality, enabling comprehensive assessments of text originality and facilitating insights into its potential origins, whether from human sources or AI models trained on plagiarized content.

In this paper, we present our AI content detection system, integrating AI text detection and plagiarism detection

* *Corresponding Author*

functionalities. We showcase the design, implementation, and evaluation of our system, emphasizing its significance in addressing contemporary challenges in NLP and information integrity.

Problem Statement

Limited Training Data: Building accurate AI text detection models requires large and diverse training datasets. However, obtaining labeled data that covers the full range of human and AI-generated text can be challenging. As a result, it lowers the performance of the system.

Multilingual and Cross-lingual Detection: Accurately discriminating between human and AI generated texts across many languages can be difficult due to language-specific traits, variations in writing styles, and cultural differences.

Single Detection Mechanism: Current AI content detection algorithms only have a single method of detecting AI content; they do not have a multi-pronged approach.

A. Adversarial Attack: Adversarial attacks attempt to mislead AI text detection systems by making slight modifications to the text that could lead the model to classify things incorrectly.

Objective

The objectives of this project are structured to address key challenges in Natural Language Processing (NLP) by developing an integrated system for text classification and plagiarism detection. The primary objectives are delineated as follows:

System Development for Text Classification

Create a robust system capable of accurately distinguishing between human-written and AI-generated text. This involves the implementation and optimization of a suitable AI model to achieve high-performance text classification.

Selection of AI Model

Evaluate and select the most effective AI model for text classification from among BERT, DistilBERT, or LSTM based on performance metrics such as accuracy, precision, recall, and F1-score. The selected model should demonstrate superior capabilities in discriminating between human and AI-generated text.

Integration of Plagiarism Detection:

Integrate a plagiarism detection mechanism utilizing a BERT-based classifier to assess the originality of input text. This component aims to enhance the overall accuracy and reliability of the system in identifying instances of plagiarized or nonoriginal content.

These objectives collectively contribute to the development of an advanced AI content detection system tailored to address contemporary challenges associated with AI-generated text and plagiarism in digital contexts. By leveraging cutting-edge AI methodologies and models, our system aims to advance the field of NLP, promoting information integrity and trustworthiness.

In subsequent sections, we will elaborate on the methodology employed for system development, encompassing model selection, training, and evaluation processes. Furthermore, we will present experimental results and insights derived from the implementation of our integrated AI content detection system, highlighting its practical applications and implications within the realm of Natural Language Processing.

Solution provided by our system

Improves the accuracy of AI text detection by combining model-based predictive analysis (using DistilBERT), statistical (or formula-based) analysis, and plagiarism detection (using BERT Classifier and cosine similarity).

Methodology

System Architecture

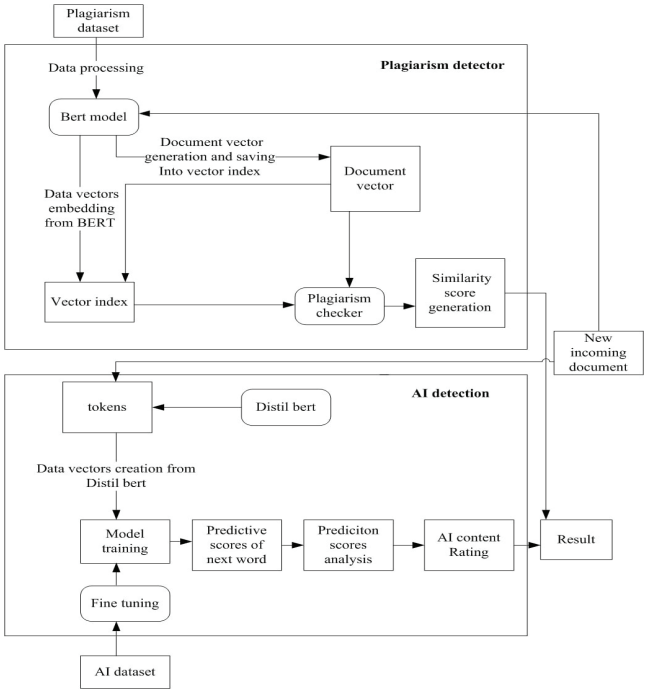

Figure 1. System Block Diagram

Our system comprises of a Plagiarism Detector and an AI content detector. The plagiarism detector tags the input text which are then used to preprocess and analyze the text. The

processed text is subsequently searched for in the database. If a match is found, the plagiarism score is displayed, and the similar article is displayed. The Plagiarism Detector enables us to detect source materials that are used to train an AI model. This source material, which shows up in the final AI-generated text, helps us determine if the text is original or not. The system then provides a plagiarism score along with a list of similar articles found in the input text. The AI content detector uses a transformer architecture-based Large Language Model to find AI-generated text using predictive analysis based on the prediction scores generated by our model. This provides us with a probabilistic measure of whether the text was generated by AI or not.

Dataset Compilation and Preparation

To facilitate the fine-tuning of our models and subsequent testing, we meticulously curated datasets from diverse sources. The compilation and preparation of these datasets were crucial for ensuring the effectiveness and reliability of our machine-learning models.

Dataset Compilation

We sourced data from various platforms and repositories, as detailed below:

1) *LLM - Detect AI Generated Text Dataset on Kaggle:* This dataset comprised 27,340 rows specifically designed for detecting AI-generated text [1].
2) *News Article Dataset on Kaggle:* A collection of 1,173 news articles obtained from Kaggle [2].
3) *Chat-GPT Prompts Dataset on Hugging Face:* Consists of 360 prompts for Chat-GPT, sourced from Hugging Face [3].
4) *Chat-GPT Detector Bias Dataset on Hugging Face:* This dataset included 749 instances aimed at detecting biases in Chat-GPT [4].
5) *Custom Testing Dataset:* A custom dataset of 30 rows was compiled using Google Forms and ChatGPT (refer to Appendix I) to evaluate the performance of fine-tuned models.

Data Tokenization

For the tokenization process, we utilized the distilBERT base-uncased model [5] to prepare the data for training and testing purposes.

Training Data for Plagiarism Detection

In preparation for the plagiarism detection component, we employed the following data sources and techniques:

1) *Tokenization:* We utilized the BERT uncased model pre-trained on a vast corpus of 3.3 billion words, sourced from Wikipedia and BooksCorpus [6].
2) *arXiv Dataset:* This dataset, obtained from Kaggle and maintained by Cornell University, comprised abstracts and metadata from over 1.7 million scholarly papers across STEM fields [7].

Data Preparation and Cleaning

Given the curated datasets, it was imperative to ensure data cleanliness and consistency to optimize their utilization in machine learning models. We undertook the following steps and techniques for data preparation and cleaning:

1) *Identification of Missing or Incomplete Data:* Each dataset column was scrutinized to identify and address missing or incomplete values.
2) *Duplicate Removal:* Duplicate entries were identified and eliminated to maintain data integrity and consistency.
3) *Standardization and Normalization:* Data underwent standardization and normalization processes to enhance model *performance*.
4) *Outlier Removal:* Outliers were identified and removed to mitigate their potential impact on model performance.
5) *Data Splitting:* The dataset was partitioned into training, validation, and test sets to prevent overfitting and assess model performance effectively.

Model Selection

From a selection of LSTM, BERT, and DistilBERT models, we opted to use DistilBERT for AI content detection based on comparative analysis (refer to Section 5.2.4 for details on model selection).

Model Training

We fine-tuned a pre-trained DistilBERT model using 24,043 rows of curated data, encompassing articles from the web (human-written) and text generated by AI models such as GPT-2 and ChatGPT. The final DistilBERT model was trained with specific hyperparameters as outlined in Table 3.1.

This meticulous approach to dataset compilation, preparation, and model selection underscores the robustness and efficacy of our AI content detection system, ensuring accurate differentiation between human-authored and AI-generated text, alongside robust plagiarism detection capabilities. Subsequent sections will delve into experimental results, performance evaluations, and insights derived from the implementation of our system.

Table I

Hyperparameters for training

Parameters	Values
learning_rate	1.84e-6
train_batch_size	16
test_batch_size	16
train_size	19227
test_size	4807
num_epochs	8

Findings

Verification and Validation

1) *Training and Validation Result of DistilBERT:* In our study, we leveraged DistilBERT, a compact variant of BERT, to develop an AI content detection system. DistilBERT was created using knowledge distillation, a technique that transfers knowledge from a larger model (BERT) to a smaller, more efficient model, resulting in approximately 66 million parameters and exceptional performance for its size.

- Dataset Preparation We curated a balanced dataset comprising 2,492 articles categorized as either "Human" or "AI". Each article was represented by a row in the dataset, with corresponding labels indicating its origin. The dataset was randomly split into training (80%) and testing (20%) sets to facilitate model training and evaluation.
- Pre-processing Prior to model training, the dataset underwent pre-processing steps, including tokenization and normalization. Tokenization involved breaking down the text into individual tokens, which were then converted into numerical representations suitable for input to the DistilBERT model. Normalization techniques were applied to standardize the textual data, enhancing model performance and convergence during training.
- Model Training and Fine-tuning We employed the DistilBERT model to perform AI content detection, fine-tuning its parameters using the prepared dataset. The training process spanned 30 epochs, during which the model learned to distinguish between human-written and AI-generated content.
- Optimization and Evaluation To address potential overfitting, we optimized the learning rate, identifying an optimal value of 1.84×10^{-6} that stabilized the training process. The model's performance was monitored using training and validation loss metrics, ensuring convergence and generalization capability.
- Results Following fine-tuning and evaluation, the trained DistilBERT model demonstrated robust performance:
- Validation Loss: 0.0903
- Validation Accuracy: 0.9769

The achieved validation accuracy underscores the effectiveness of the AI content detection system, validating its utility in distinguishing between human-authored and AI-generated text.

Through this methodology, we have showcased the efficacy of leveraging DistilBERT for AI content detection, emphasizing its compactness, efficiency, and remarkable performance in real-world applications.

In the confusion matrix above, each row represents the instances in an actual class whereas each column represents the instances in a predicted class. In the above matrix:

- Instances correctly predicted as "AI": 238
- Instances incorrectly predicted as "AI" when they are actually "Human": 11
- Instances correctly predicted as "Human": 245
- Instances incorrectly predicted as "Human" when they are actually "AI": 4

Table ii

Classification report for distilbert

Class	Precision	Recall	F1-Score	Support
AI	0.98	0.96	0.97	249
Human	0.96	0.98	0.97	249
Accuracy			0.97	498

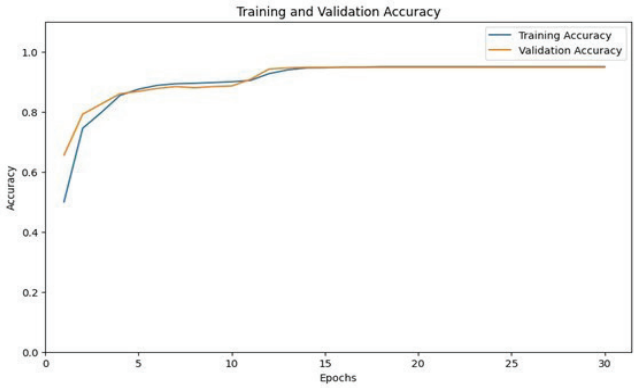

Figure 2. Accuracy of DistilBERT

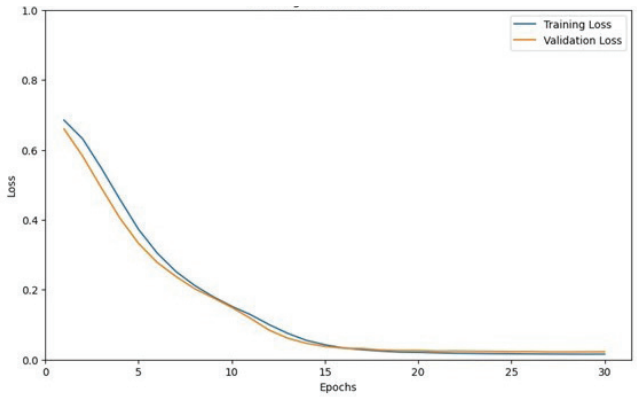

Figure 3. Loss of DistilBERT

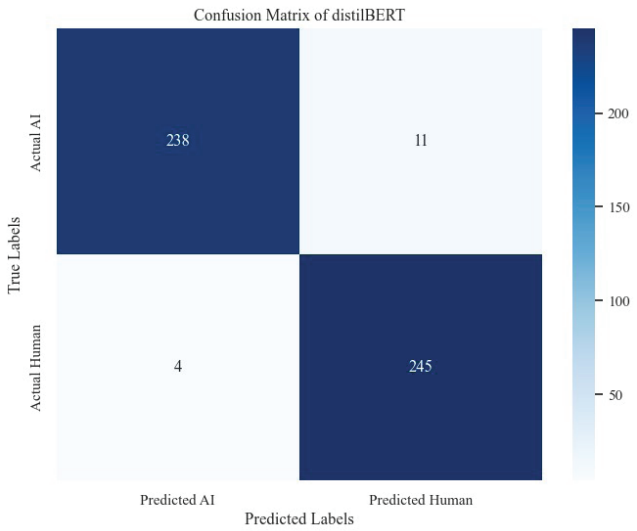

Figure 4. Confusion Matrix of DistilBERT

2) Training and Validation Result of BERT: BERT stands for Bidirectional Encoder Representations from Transformers and is a family of language models suitable for Natural Language Processing (NLP) introduced in 2018 by researchers at Google. With the use of surrounding text to create context, BERT is intended to assist computers in understanding meaning-ambiguous language in text.

We used the pre-trained BERT base model (uncased) for the purpose of AI-generated text detection. The BERT base model has 109,531,521 total parameters, all of which are trainable. We used a dataset with 2492 data rows and labels of '0' for Human content and '1' for fine-tuning the BERT base model for our specific purpose. The updated metrics for this fine-tuned model upon evaluation were found to be:

- Validation Loss: 0.1371
- Validation Accuracy: 0.948

The training was done for 30 epochs using a train-test split ratio of 4:1 on the training dataset. The model achieved stability in both the training and validation datasets

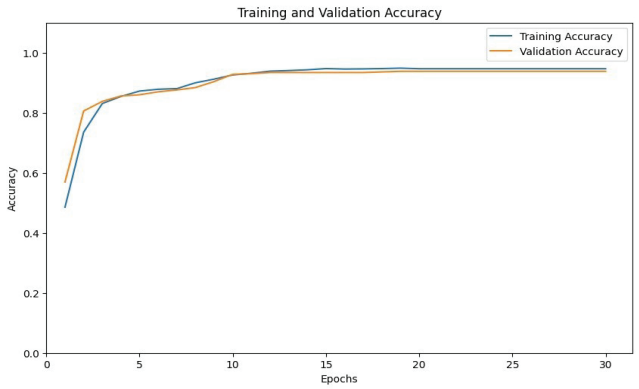

Figure 5. History of Model Accuracy of BERT

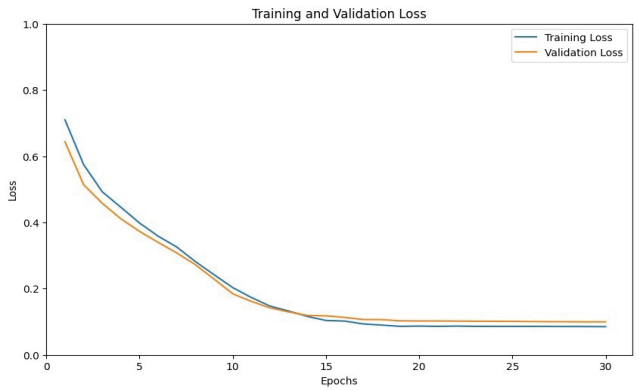

Figure 6. History of Model Loss of BERT

In the confusion matrix above, each row represents the instances in an actual class whereas each column represents the instances in a predicted class. In the above matrix:

- Instances correctly predicted as "AI": 230
- Instances incorrectly predicted as "AI" when they are actually "Human": 6
- Instances correctly predicted as "Human": 243
- Instances incorrectly predicted as "Human" when they are actually "AI": 19

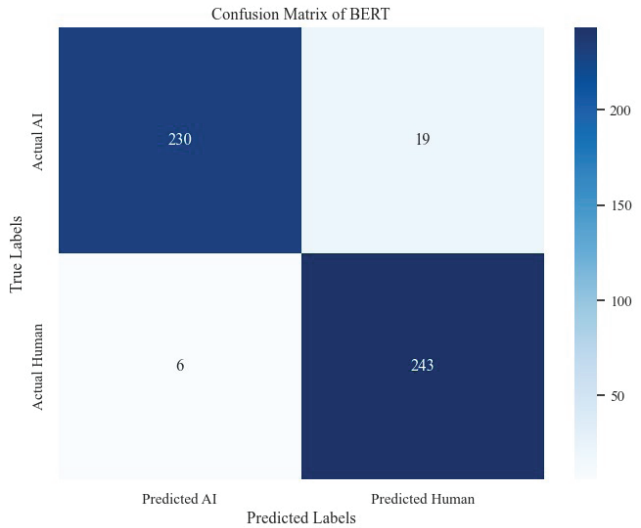

Figure 7. Confusion Matrix of BERT

Table II

CLASSIFICATION REPORT FOR BERT

Class	Precision	Recall	F1-Score	Support
AI	0.97	0.92	0.94	249
Human	0.93	0.97	0.95	249
Accuracy			0.95	498

From the above metrics, it's clear that the model produced correct predictions for 498 test cases, while having 25 False Positives and 50 True Negatives. This proves that the model is rather accurate and can be used for our project.

3) Training and Validation Result of LSTM: LSTMs are long short-term memory networks that use artificial neural networks (ANN) in the field of artificial intelligence (AI) and deep learning. Unlike normal feed-forward neural networks, also known as recurrent neural networks, these networks feature feedback connections. Applications of LSTM include unsegmented, connected handwriting recognition, robot control, video gaming, speech recognition, machine translation, and healthcare[8].

The model was trained using the same dataset as the above models. The dataset consisted of labeled data with 2492 rows labeled as either human or AI-generated. The LSTM architecture consisted of 6 layers, with the first layer being the Embedding layer, followed by four LSTM layers with 128, 128, 64, and 32 units respectively. The last layer was a dense layer for classification with a single unit.

After the evaluation, the final accuracy and loss are as follows:

- Validation Loss: 0.48
- Validation Accuracy: 0.55

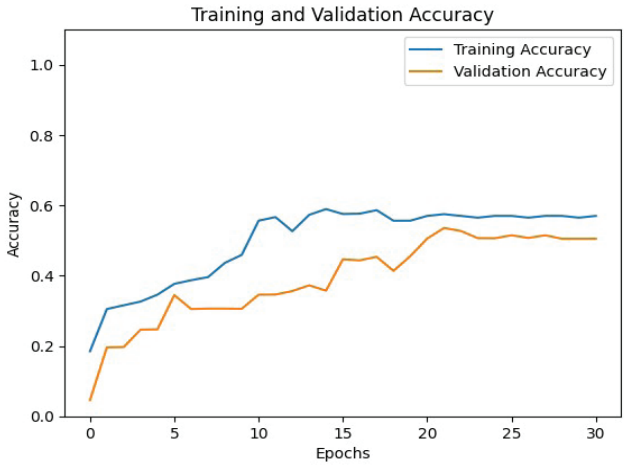

Figure 8. History of Model Accuracy of LSTM

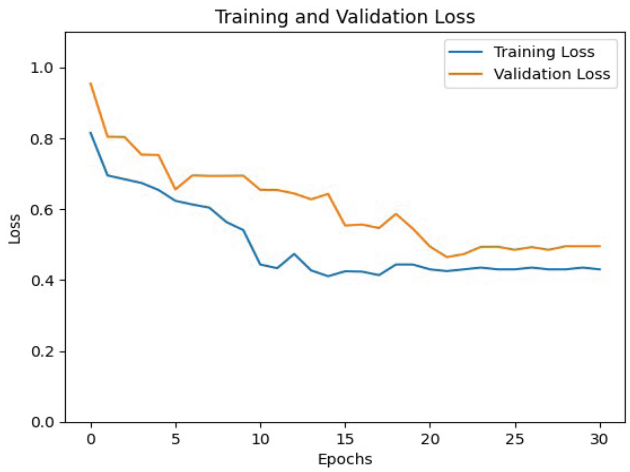

Figure 9. History of Model Loss of LSTM

In the above confusion matrix:

- Instances correctly predicted as "AI": 75
- Instances incorrectly predicted as "AI" when they are actually "Human": 50
- Instances correctly predicted as "Human": 199
- Instances incorrectly predicted as "Human" when they are actually "AI": 174

Table iii

Classification report for lstm

Class	Precision	Recall	F1-Score	Support
AI	0.6	0.3	0.4	249
Human	0.53	0.8	0.64	249
Accuracy			0.55	498

Selection of best model for AI Content Detection

Initially three models, LSTM, BERT and ditilBERT were trained on a balanced data (equal number of Human and AI text) of randomly selected data from the total dataset. This balanced dataset was compiled by combining the News Article Dataset on Kaggle [2], Chat-GPT prompts dataset on Hugging Face [3] and Chat-GPT detector bias on Hugging Face [4] and randomly selecting 1246 rows of Human text and 1246 rows of AI generated text totalling to 2492 rows of data. The BERT model, initially showing moderate performance, underwent fine-tuning and emerged with exceptional results. Its validation accuracy surged to 94.8%, accompanied by a substantial reduction in test loss. On the unseen dataset it had a 52 precision of 97% for AI content and 93% for Human content, it is slightly better at identifying AI content and with a recall of 97% for Human content and 92% for AI content it is more capable of detecting Human content. The LSTM approach demonstrated potential in handling sequential data like text but lagged behind BERT's performance. With a validation accuracy of 51had an overall accuracy of 55% on the unseen test dataset. The LSTM was not a pretrained model hence it underperformed. This highlighted the efficiency of pretrained models in NLP tasks. DistilBERT with a precision of 98% for AI content and 96% for human content, it is slightly better at identifying AI content, and with a recall of 96% for AI content and 98% for human content it is slightly better at detecting Human content. With a validation accuracy of 97.7%, the distilBERT model generalized well. In summary, this project illuminated the strengths of fine-tuned BERT and distilBERT models in comprehending and categorizing intricate text content. This showcased the vast improvement transformers bring to NLP in terms of LSTM. This underscores the transformative impact of pre-trained transformer models in advancing natural language processing.

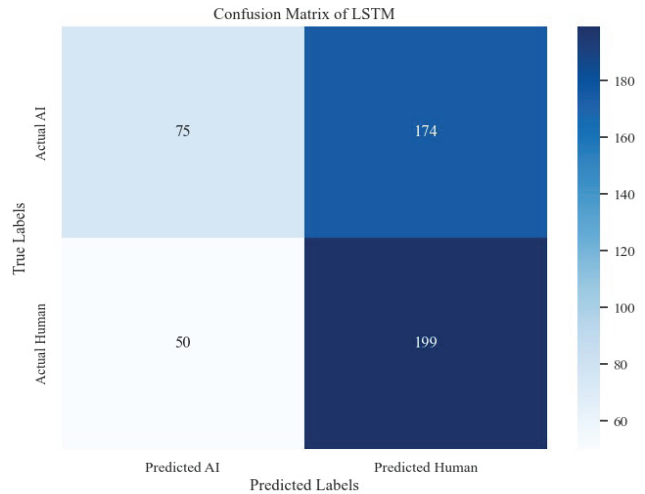

Figure 10. Confusion Matrix of LSTM

Table iv

Comparison of bert, lstm, and distilbert based on their ai classification reports

LLM	Precision	Recall	F1-Score	Accuracy
BERT	0.97	0.92	0.95	0.95
LSTM	0.6	0.3	0.4	0.55
DistilBERT	0.98	0.96	0.97	0.97

Final DistilBERT Model

After selecting the DistilBERT model, the DistilBERT model was finetuned with the final prepared dataset [1] [2] [3] [4] of 24034 rows and 2 columns. This dataset consisted of equal division between the 'Human' and 'AI' dataset; 12017 were labelled 'Human' and 12017 were labelled 'AI'. The dataset was split randomly into a training dataset and a testing dataset with the training dataset containing 80% and the validation dataset containing 20% of the total dataset. Then the DistilBERT model was finetuned using the training dataset up to 8 epochs, where it achieved a validation accuracy of 96.2% and this model's finetuned parameters were saved. The model was then evaluated with the unseen dataset to calculate the confusion matrix below.

The result of finetuning was:

Validation Accuracy: 0.962

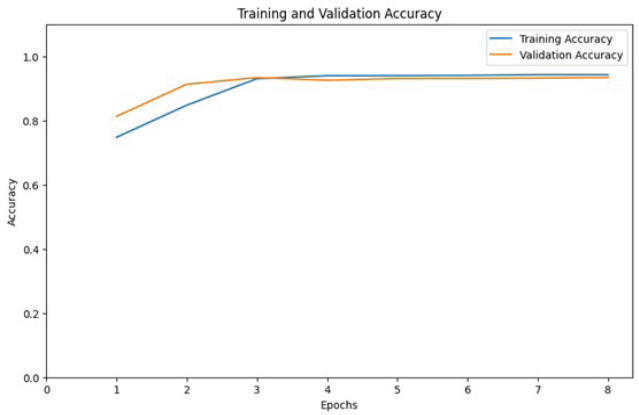

Figure 11. Accuracy graph of final DistilBERT model

Validation Loss: 0.0593

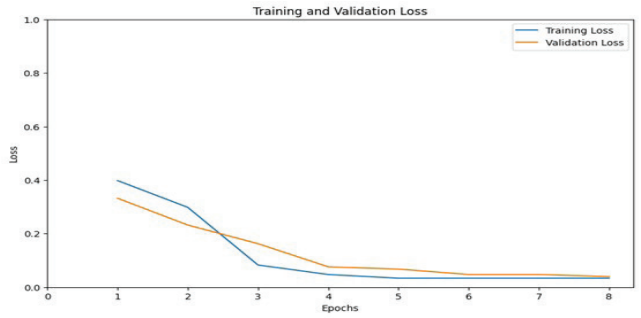

Figure 12. Loss graph of final DistilBERT model

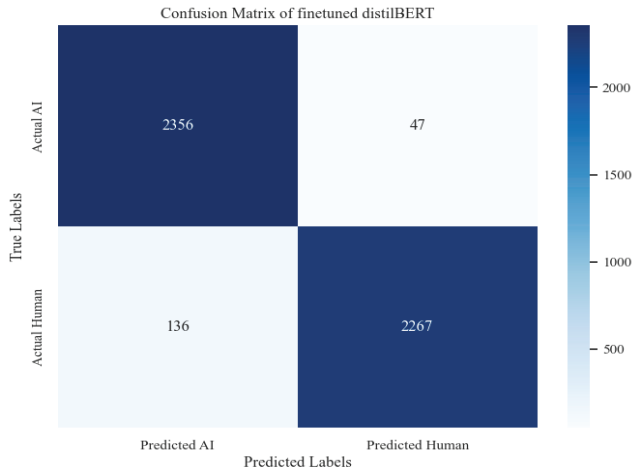

Figure 13. Confusion Matrix of finetuned DistilBERT

In the confusion matrix above, each row represents the instances in an actual class whereas each column represents the instances in a predicted class. In the above matrix:

- Instances correctly predicted as "AI": 2356
- Instances incorrectly predicted as "AI" when they are actually "Human": 47
- Instances correctly predicted as "Human": 22267
- Instances incorrectly predicted as "Human" when they are actually "AI": 136

Table v
Classification report for distilbert

Class	Precision	Recall	F1-Score	Support
AI	0.94	0.98	0.95	2403
Human	0.98	0.94	0.95	2403
Accuracy			0.95	4806

Test Result on Custom Dataset

To evaluate the performance of our system in text generation, we compiled a dataset (see Appendix I) consisting of 10 samples each of AI-generated text, human-generated text, and mixed text (with AI and Human Content mixed in some proportion). The test was performed on the fine-tuned distilBERT model.

The human dataset is a collection of essays from our acquaintances and colleagues. We compared our system against two existing systems: Winston.ai and Undetectable.ai. The aim was to assess the strengths and weaknesses of each system in producing text. The text column contains the text that is being evaluated, the Actual column is the actual origin of the text, which is either purely AI, purely Human or a mix of AI-generated text and Human-generated text. Winston AI and undetectable.ai column are the results of putting the texts through the respective AI content checker and finally, the Our Model column is the result of testing the text with our final fine-tuned model. The results are as follows:

Table vi
Test results for distilbert vs winston ai vs undetectable.Ai

TEXT NO	Actual Source	Prediction of Winston AI	Prediction of Undetectable.ai	Prediction of Our System
TEXT 1	AI	AI	AI	AI
TEXT 2	AI	AI	AI	AI
TEXT 3	AI	AI	AI	AI
TEXT 4	AI	AI	AI	Human
TEXT 5	AI	AI	AI	AI
TEXT 6	AI	AI	AI	AI
TEXT 7	AI	AI	AI	AI
TEXT 8	AI	AI	AI	AI
TEXT 9	AI	AI	AI	AI
TEXT 10	AI	AI	AI	AI
TEXT 11	Human	Human	Human	Human
TEXT 12	Human	Human	AI	Human
TEXT 13	Human	Human	Apparent human	Human
TEXT 14	Human	Human	AI	Human
TEXT 15	Human	AI	Apparent human	Human
TEXT 16	Human	AI	Apparent human	Human
TEXT 17	Human	Human	Apparent human	Human
TEXT 18	Human	AI	Apparent human	Human
TEXT 19	Human	Human	human	Human
TEXT 20	Human	Human	AI	Human
TEXT 21	AI + Human	100% AI	Apparent human	AI
TEXT 22	AI + Human	100% AI	AI	AI
TEXT 23	AI + Human	100% AI	AI	AI
TEXT 24	AI + Human	100% AI	Apparent human	Human
TEXT 25	AI + Human	100% Human	100% AI	Human
TEXT 26	AI + Human	100% AI	100% AI	AI
TEXT 27	AI + Human	100% AI	100% AI	AI
TEXT 28	AI + Human	100% AI	100% AI	Human
TEXT 29	AI + Human	100% Human	100% AI	Human
TEXT 30	AI + Human	100% AI	100% AI	Human

Confusion Matrix for Testing Custom Dataset

This confusion matrix is the result of testing the model on the custom dataset of 30(excluding mixed data with both AI and Human text) created by collecting responses through Google Forms and from prompts given to ChatGPT (see Appendix I):

Table vii
Confusion matrix for distilbert (custom dataset)

	Predicted AI	**Predicted Human**
Actual AI	9	1
Actual Human	0	10

In the above confusion matrix:

- Instances correctly predicted as "AI": 9
- Instances incorrectly predicted as "AI" when they are actually "Human": 0
- Instances correctly predicted as "Human": 10
- Instances incorrectly predicted as "Human" when they are actually "AI": 1

Classification Report for Testing Custom Dataset

Table viii
Classification report for distilbert (custom dataset)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
AI	1.00	0.90	0.95	10
Human	0.91	1.00	0.95	10
Accuracy				0.95

Discussion

Upon inspection and analysis of the verification and validation test results we could outline the following points about our system:

- This system demonstrated a commendable accuracy rate of 95%, accurately classifying text types except in a few instances.
- On the mixed dataset, it generally predicted AI or Human depending on which content was more present.

Evaluating Winston.ai test results:

- Winston.ai struggled with mixed text samples, displaying inconsistencies and coherence issues when blending AI-generated and human-generated content.
- Additionally, Winston.ai imposes a constraint of 600 characters, which may limit the complexity and depth of generated text. Its credit hour system further restricts usage, potentially hindering continuous text generation.

Contrasting Undetectable.ai test results:

- While Undetectable.ai exhibited a high degree of creativity and variability in language use, it often failed to accurately predict human-generated text, leading to inconsistencies.
- The system's unpredictability sometimes resulted in disjointed predictions, giving an overall accuracy of 95%.

The comparison highlighted distinct characteristics and performance metrics for each text generation system. While our system achieved a balance of AI and human text detection, Winston.ai struggled with mixed text samples and imposed constraints on character count and usage. Undetectable.ai showcased good accuracy but faced challenges in accurately predicting human generated text.

Acknowledgment

First and foremost, we would like to express our gratitude to the Institute of Engineering for including this major project as part of our course's syllabus, as it has allowed us to gain a better understanding of machine learning and deep learning. Similarly, we must express our gratitude to the Department of Computer Engineering, Kathmandu Engineering College for assisting us in the development of this project. We are extremely grateful to our teachers for providing us with all the necessary project guidance and supervision.

We owe a debt of gratitude to Er. Sharad Chandra Joshi, our project and year coordinator, for his invaluable advice and suggestions in the development of our project. We also are thankful to Er. Dhawa Sang Dong for his input on the project idea. We are obliged to Associate Prof Er. Sudeep Shakya, Head of the Department of Computer Engineering, for all his help and support so far.

Our lecturers have been helpful in giving us all the necessary project advice and monitoring. We are indebted to our parents and friends for their unwavering support, cooperation, and encouragement, which aided us much during our endeavor.

References

[1] S. Thite. *LLM - Detect AI Generated Text Dataset*. Accessed: Mar. 08, 2024. Kaggle.com. URL: https://www.kaggle.com/datasets/sunilthite/llm-detect-ai-generated-text-dataset.

[2] Asadmahmood. *News Articles*. Accessed Mar. 08, 2024. Kaggle.com. URL: https://www.kaggle.com/datasets/asad1m9a9h6mood/ news-articles/data.

[3] M. Rashad. *ChatGPT-prompts*. Accessed Mar. 08, 2024.

Hugging Face. URL: https://huggingface.co/datasets/MohamedRashad/ ChatGPT-prompts.

[4] W. Liang et al. *ChatGPT-Detector-Bias*. Accessed Mar. 08, 2024. Hugging Face. URL: https://huggingface.co/datasets/WxWx/ChatGPT-Detector-Bias.

[5] V. JAYANT. *distilbert-base-uncased*. Accessed Mar. 08, 2024. Kaggle.com. URL: https://www.kaggle.com/datasets/virajjayant/distilbertbaseuncased.

[6] J. Devlin et al. "Bert: Pre-training of deep bidirectional Transformers for language understanding". In: *arXiv.org* (). Accessed: 13 February 2024. URL: https://arxiv.org/abs/1810.04805.

[7] *arXiv Dataset*. Accessed Mar. 08, 2024. Kaggle.com. URL: https://www.kaggle.com/datasets/Cornell-University/arxiv/data.

[8] Mohamed Banoula. *Introduction to long short-term memory (LSTM)*. Online. Accessed: 13 February 2024. Simplilearn, 2023. URL: https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/lstm.

Appendix i- testing data for ai content detection

Self-curated custom dataset through google forms and ChatGPT of 30 rows with 10 rows of human written text, 10 rows of AI generated text and 10 rows with mixed AI and human text used for testing purposes in this project is provided below:

Ai Generated Text	Text No
Gratitude, often described as the antidote to negativity, holds remarkable power in shaping our perspectives and experiences...	TEXT 1
Gratitude is a transformative force that enriches our lives with a deep sense of appreciation, humility, and connection to the world around us...	TEXT 2
In a world that often values conformity over innovation, creativity serves as a beacon of light, illuminating the path to a future filled with possibility and promise...	TEXT 3
Creativity, often heralded as the hallmark of human ingenuity, transcends the realms of art, science, and entrepreneurship...	TEXT 4
In a world characterized by rapid change, relentless demands, and constant connectivity, the pursuit of happiness has emerged as a universal aspiration...	TEXT 5
Happiness is a state of being that transcends mere pleasure or satisfaction; it is a profound sense of contentment, fulfillment, and inner peace...	TEXT 6
Friendship is a cherished bond that enriches our lives with companionship, support, and shared experiences...	TEXT 7
Indeed, the choice of units is fundamental in ensuring the accuracy and relevance of mathematical calculations, particularly in real-world applications...	TEXT 8
Family serves as a sanctuary in times of joy and celebration, providing a safe haven where we can share laughter, create cherished memories, and revel in each other's company...	TEXT 9
Indeed, life is a remarkable journey, a grand tapestry woven from the threads of experiences, emotions, and growth...	TEXT 10
Hi I am Abhigya, I go with the pet name abu. I have a lovely family of 5 and I have a dog whose name is Happy. He is a German shepherd but he looks like simba from The Lion King. ...	TEXT 11
It is proof of his growing stature in Nepali cricket that he is trusted with the leadership of the under-19 team despite being a regular member of the senior team...	TEXT 12
Five years back when Nepal played the T20 World Cup in Bangladesh, millions of kids back home were inspired to hold the bat and ball. One of them was Rohit Kumar Paudel...	TEXT 13
A morning walk is a good exercise for all. It is very helpful for our health. It is as useful for the body as food. ...	TEXT 14
There was once a unique child named Mai. He was a happy child with many older siblings. His father was a respected man in the village, while his mother was very loved. ...	TEXT 15
Ram is the name of my best friend. He has curly hair and is the tallest guy in our class. He excels in all sports events such as running races, football, and other similar activities...	TEXT 16
During our summer holidays, we even went to summer camp together to learn cricket and made a lot of memories. Moreover, we also invented new and our own handshake which only we both knew...	TEXT 17
My bag is pink in colour and on the top of it, the picture of my favourite cartoon Cinderella is there. It is a two-compartment bag in which I keep my notes in the first compartment whereas books in the second compartment...	TEXT 18

Spandan is a female in her 20's who just graduated from her BBA in Kathmandu. Currently she is working as the project associate at Foodmandu, which is her first corporate job...	TEXT 19
My house has a small garden. It contains different types of plants like various flowers such as roses, lilies, sunflowers, and daisies. All flowers have different colors, but I like roses...	TEXT 20
Parks are essential components of urban and rural environments, serving as green oases that provide numerous benefits to communities and the environment...	TEXT 21
Road accidents are very common in big cities. Careless driving causes accidents. Some drivers do not obey the traffic rules...	TEXT 22
"My mother," two words that encapsulate a universe of love, sacrifice, and unwavering support. She is the embodiment of strength, grace, and resilience. She is my confidante, my cheerleader, and my rock...	TEXT 23
A picnic, a delightful escape into nature's embrace, is a cherished tradition that brings together friends, family, and loved ones for an idyllic outdoor gathering filled with laughter, relaxation, and culinary delights...	TEXT 24
Never in my life did I think writing about myself would be this hard. Hi, this is me, AAAA AAA, giving a little reflection about myself...	TEXT 25
Teachers are a special blessing from God to us. They are the ones who build a good nation and make the world a better place. A teacher teaches us the importance of a pen over that of a sword...	TEXT 26
Many people travel for different purposes. Whether it's for a business trip or a holiday, we see people traveling often. Some people prefer hilly areas, while others like to travel to places with beaches...	TEXT 27