# "AI vs. Human Text: Plagiarism Detection"

**Mauryan Uppalapati**
muppalapati@tulane.edu

**Russell George**
rgeorge7@tulane.edu

## 1 Problem Overview

The proliferation of AI text generators has introduced a novel challenge in distinguishing between text authored by humans and that generated by artificial intelligence. This distinction holds significant implications for academic integrity, copyright laws, and content originality verification. At the heart of the problem lies the need to detect subtle patterns and sequences in text that differentiate human creativity and language use from the programmed randomness and learning models inherent in AI-generated text. This project aims to delve into these differences, which are critical for maintaining authenticity and accountability in digital content creation. The dataset used in this project was obtained from an ongoing Kaggle competition titled "AI vs. Human Text: Plagiarism Detection." This dataset comprises 500,000 essays, which include text generated by artificial intelligence (AI) models as well as essays written by humans.

## 2 Data

The dataset used in this project was obtained from an ongoing Kaggle competition titled "AI vs. Human Text" This dataset comprises 500,000 essays, which include text generated by artificial intelligence (AI) models as well as essays written by humans. Upon accessing the Kaggle dataset repository, the data was downloaded and utilized for the project. The dataset consists of a collection of essays, each labeled to indicate whether it was authored by a human or generated by AI. This labeling enables supervised learning techniques to be applied for text classification tasks. Prior to model training, the dataset underwent preprocessing steps, including data cleaning, tokenization, and normalization, to ensure consistency and compatibility with the chosen machine learning and deep learning algorithms. Additionally, the ongoing nature of the competition ensures the relevance and currency of the dataset.

## 3 Methods

The project will utilize a combination of traditional machine learning algorithms and deep learning architectures for text classification and plagiarism detection. Initially, standard classifiers such as Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, and Random Forests were trained on the dataset to establish baseline performance. Subsequently, a Long Short-Term Memory (LSTM) network was explored for its ability to capture complex patterns in text data.

Evaluation metrics such as accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curve analysis were used to assess the performance of the classifiers and deep learning models. Additionally, qualitative analysis will be conducted to understand the models' ability to detect subtle differences between human-written and AI-generated text. Subsequently, there will be a transformation phase, where upon receiving input identified as AI-generated, the project will aim to iteratively modify the text until it can be classified as human-written. This iterative transformation process will involve techniques such as paraphrasing, grammatical adjustments, and semantic restructuring. The goal is to develop a system capable of refining AI-generated text to pass as human-written, thereby highlighting the malleability and susceptibility of current text classification methods.

## 4 Preliminary Experiments & Results

In the initial phase of our analysis, we conducted a comprehensive review of the training data,

Therefore, the dataset at hand was of substantial size and complexity

Subsequently, to establish a baseline for perfor-

```
RangeIndex: 341064 entries, 0 to 341063
Data columns (total 2 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   text       341064 non-null  object
 1   generated  341064 non-null  int64
dtypes: int64(1), object(1)
memory usage: 5.2+ MB
None

First Few Rows:
                                                text  generated
0  Cars. Cars have been around since they became ...          0
1  Transportation is a large necessity in most co...          0
2  "America's love affair with it's vehicles seem...          0
3  How often do you ride in a car? Do you drive a...          0
4  Cars are a wonderful thing. They are perhaps o...          0

Dataset Shape:
(341064, 2)

Column Names:
['text', 'generated']
```

mance, we proceeded to train a suite of standard classifiers on the dataset. This ensemble included Logistic Regression, Naive Bayes, Decision Trees, Extra Trees Classifier, and Random Forests. Each model was rigorously evaluated using a variety of metrics, namely accuracy, precision, recall, and the F1 score, to ensure a holistic assessment of their performance.

```
Logistic Regression has been trained.
Evaluation Metrics for Logistic Regression:
Accuracy: 0.9773
Precision: 0.9839
Recall: 0.9549
F1 Score: 0.9692

Multinomial Naive Bayes has been trained.
Evaluation Metrics for Multinomial Naive Bayes:
Accuracy: 0.9165
Precision: 0.9905
Recall: 0.7847
F1 Score: 0.8757

Decision Tree has been trained.
Evaluation Metrics for Decision Tree:
Accuracy: 0.9043
Precision: 0.8637
Recall: 0.8842
F1 Score: 0.8738

Random Forest has been trained.
Evaluation Metrics for Random Forest:
Accuracy: 0.9653
Precision: 0.9905
Recall: 0.9161
F1 Score: 0.9519

Extra Trees Classifier has been trained.
Evaluation Metrics for Extra Trees Classifier:
Accuracy: 0.9707
Precision: 0.9920
Recall: 0.9293
F1 Score: 0.9596
```

The results of our evaluation indicate that Logistic Regression emerged as the most effective model, demonstrating the highest levels of accuracy and F1 score among the classifiers tested. This was closely followed by the Extra Trees Classifier, which also exhibited commendable performance metrics.This structured approach not only allowed us to gauge the baseline performance of conventional classifiers on the dataset but also provided invaluable insights into the comparative efficacy of these models in handling the data's inherent complexities.

We then proceeded to train a Long Short-Term Memory (LSTM) network

The model achieved a test accuracy of 61.5.The test loss stood at 0.766. Furthermore, the model demonstrated a F1 score of 0.6637

```
Epoch 1/5
853/853 ━━━━━━━━━━━━━━━━━━━━ 573s 671ms/step - accuracy: 0.8697 - loss: 0.3191 - val_accuracy: 0.9598 - val_loss: 0.1366
Epoch 2/5
853/853 ━━━━━━━━━━━━━━━━━━━━ 539s 632ms/step - accuracy: 0.9610 - loss: 0.1309 - val_accuracy: 0.8763 - val_loss: 0.3055
Epoch 3/5
853/853 ━━━━━━━━━━━━━━━━━━━━ 515s 604ms/step - accuracy: 0.9013 - loss: 0.2634 - val_accuracy: 0.8165 - val_loss: 0.4140
Epoch 4/5
853/853 ━━━━━━━━━━━━━━━━━━━━ 525s 615ms/step - accuracy: 0.9128 - loss: 0.2314 - val_accuracy: 0.9774 - val_loss: 0.0696
Epoch 5/5
853/853 ━━━━━━━━━━━━━━━━━━━━ 526s 616ms/step - accuracy: 0.9787 - loss: 0.0704 - val_accuracy: 0.6151 - val_loss: 0.7662
214/214 ━━━━━━━━━━━━━━━━━━━━ 27s 124ms/step - accuracy: 0.6145 - loss: 0.7748
Test Loss: 0.7661918497371814
Test Accuracy: 0.6150488912506103
214/214 ━━━━━━━━━━━━━━━━━━━━ 28s 129ms/step
Precision: 0.4967420467612112
Recall: 1.0
F1 Score: 0.6637644840894751
ROC AUC: 0.9883258726914716
```

## 5 Related Work

*DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature* by Eric Mitchell, Yoonho Lee, Alexander (Sasha) Khazatsky, Christopher D. Manning, and Chelsea Finn. This work focuses on computing log-probabilities in text to determine its AI-generated nature, avoiding the traditional data collection and classifier training approaches.

*Real or Fake? Learning to Discriminate Machine from Human Generated Text* by Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. This study leverages various machine learning models to differentiate AI-generated text, echoing the methodologies we have employed.

*ChatGPT or Human? Detect and explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text* by Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Utilizing a transformer-based model, this research examines the intricacies of distinguishing short AI-generated online reviews.

*MGTBench: Benchmarking Machine-Generated Text Detection* by Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. The introduction of MGTBench, a framework aimed at detecting machine-generated texts against advanced language models, aligns with our objectives of evaluating our model within a structured benchmark.

## 6 Division of Labor

There remain substantial areas for potential improvement and approaches that we intend to explore. Before proceeding with these directional shifts and the implementation of new ideas, we will seek guidance and insights from Dr. Culotta.

To facilitate effective collaboration and ensure the continuity of our project's momentum, we have scheduled bi-weekly in-person meetings. Additionally, to enhance our workflow and project management, we will utilize shared Google Docs for real-time collaboration and maintain an ongoing

dialogue through a dedicated group chat. This will enable us to coordinate our efforts seamlessly.

Task allocation will be conducted with a focus on fairness and aligning responsibilities with individual interests and expertise, ensuring that both of us are engaged and contributing optimally to our collective objectives.

## 7 Timeline

This is a guiding framework for our project's key milestones, as we do not have precise submission dates as of yet: April 9th, Tuesday: Engage in a comprehensive discussion with Dr. Culotta to establish a clearly defined project direction, including the formulation of a detailed outline and plan. April 19th: Completion of the project report. April 20th: Finalize and publish the repository's README.md file. April 25th: Presentation of the project demo April 30th: Delivery of the project presentation.

## 8 References