SEA 820 NLP Final Project: Detecting AI-Generated Text

By Madhur Saluja and Mohammed Aeraf Khan

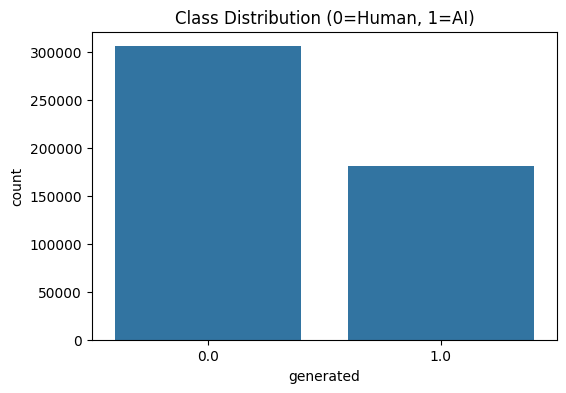
# Methodology

## Exploratory Data Analysis

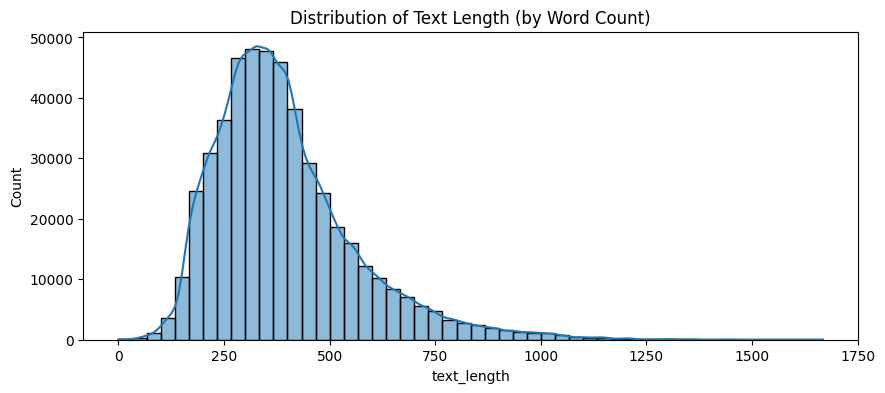
The main objective of this project was first to train a classical machine learning baseline model to classify whether a piece of text is AI-generated (1) or Human-written (0), and then to fine-tune a transformer model to achieve better performance metrics than the baseline model.

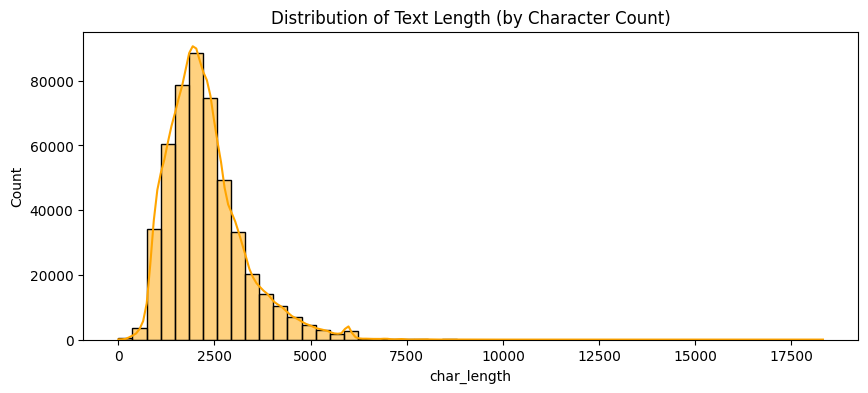
To achieve this, we started by loading the dataset and performing some Exploratory Data Analysis:

1. Class distribution: We found out that Human-written Text is almost double that of AI-generated text. With almost 300000 out of 487235 records being Human-written Text. This indicates that it is quite possible that the model will be more biased towards classifying text as Human-written text.

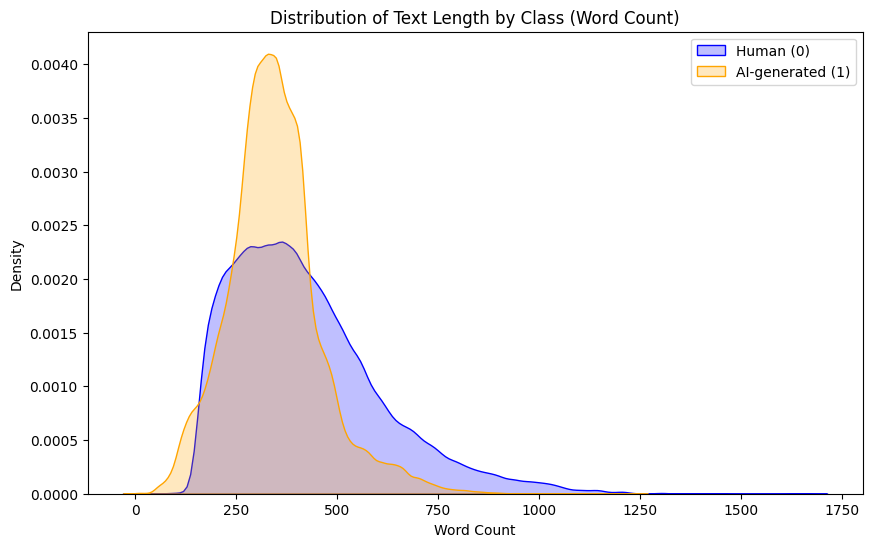


1. Text Length Distribution: Then, we performed some analysis on Text Length Distribution. Our Analysis, however, didn’t prove very valuable as we needed to find out the class-specific Text Length distribution.



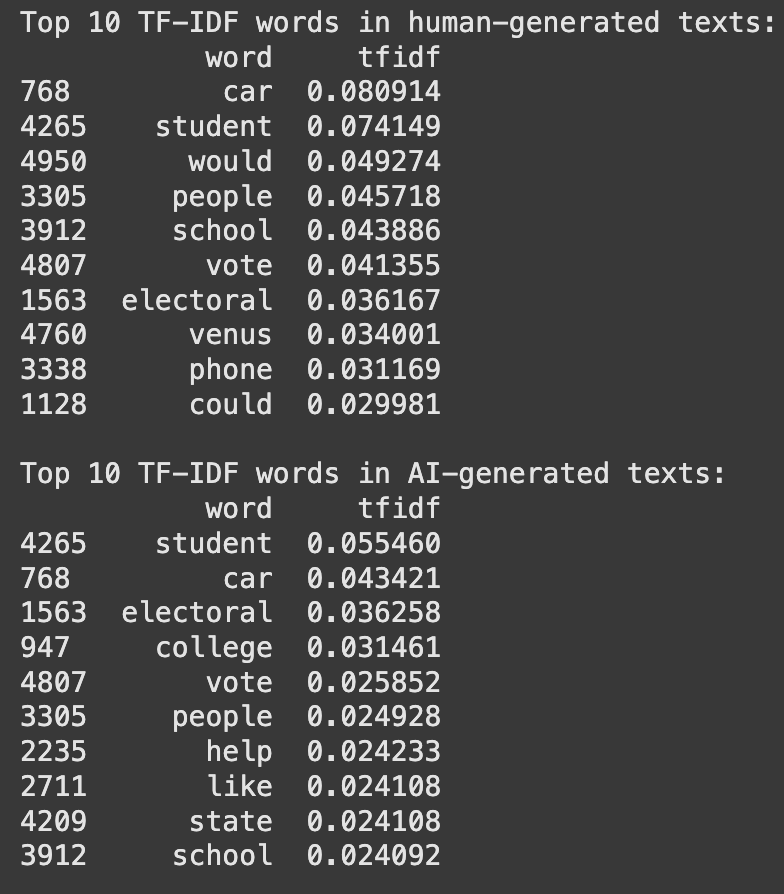


1. Distribution of Text Length by Class (Word Count): We experimented with the DIstribution of Text Length by class and found out that text length could be a distinguishing feature of the two classes. While AI-generated texts are long, many were shorter and more closely clustered than human-generated texts, and this might be a pattern in the distribution that our baseline model might learn.



1. Vocabulary Analysis (TFIDF): Following the preprocessing steps (lowercasing, stopword removal, lemmatization), we applied TF-IDF vectorization for a comparison of important words across human- and AI-generated corpora.

We calculated average TF-IDF scores for each given class and extracted the Top 10 words for each:



Some words were represented in both classes, including student, car, vote, and people. These are words that might be topical or just common across the dataset.

This lexical separation is one reason the classical models performed so highly.

However, this is also a potential problem because it makes the model susceptible to surface cues. For example, if in the future AI-generated texts were to have traded vocabulary (e.g., dubious words like humans sometimes do). The models might not be able to generalize.

## Implementing the Baseline Model

We first created a few performance standards and executed two standard models in logistic regression and multinomial naive bayes with simple TF-IDF features generated from the pre-processed text.

**TF-IDF Vectorization**

To convert our text into numerical features, we used TfidfVectorizer from scikit-learn. Our default configuration used all the unique words that were processed from the corpus — i.e. a very large feature space of 250,593 dimensions across ~487K text samples. While it resulted in the most computationally challenging setup, it also produced the greatest accuracy and F1 scores.

**Model Training & Evaluation**

* We stratified the data and put out 80% as training data and 20% as test data.
* Logistic Regression was trained on max\_iter=1000 for convergence due to the high dimensionality.
* Both models were assessed by precision, recall, F1, and overall accuracy.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | 99.29% | 99% | 99% | 99% |
| **Naive Bayes** | 95.17% | 94–98% | 89–99% | 93–95% |

**Max Features Experimentation**

We experimented with a number of max\_features values in the TF-IDF vectorizer to see how they performed:

**100 features:** Accuracy and precision dropped considerably (~70-75%).

**500 and 1,000:** Some improvement, but still performed poorly (~85-90%).

**5,000:** Performance improved further to ~96-97%.

**10,000**: Almost 99% accuracy and precision. At this point, we were considering setting the max-feature at 10,000 as there was no change in accuracy and precision further on.

**No cap (250,593 features)**: Best performance, but also by far the best memory usage and training time.

**Final Choice**

We chose no cap on features at the final baseline because it provided the best possible classification, though at a greater computational cost. This model will be the baseline for considering more advanced models (such as fine-tuning Transformers).

## Class Balancing and Length Normalization

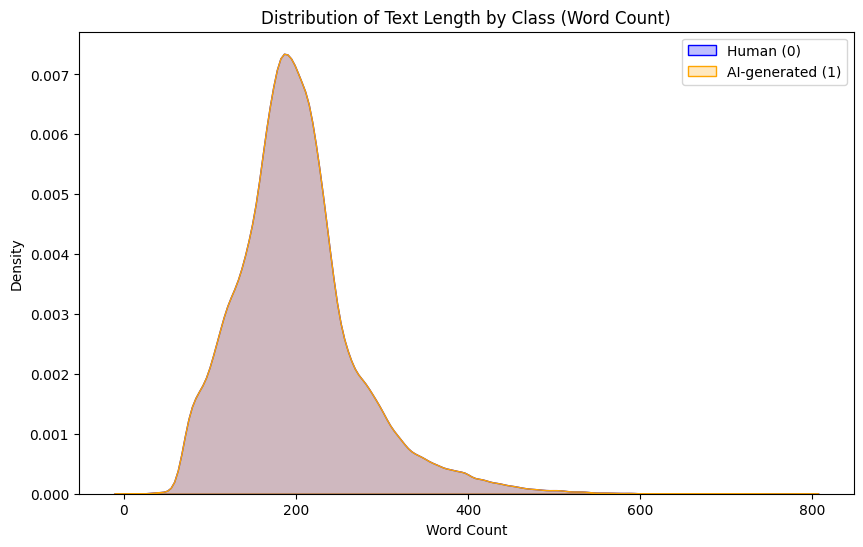
To begin with, we noted that the number of human-written samples exceeded AI samples (roughly 300k to 180k), and that the distributions of text length considerably varied between the two. These characteristics could introduce bias — especially since classical models can pick up more on low-level statistical tendencies (e.g., length).

To account for this,

* We uniformed the class sizes by taking the same number of human and AI samples in each word length cluster.
* This gave a uniform dataset of 352,322 samples — 50% human, 50% AI



* We also ensured the distribution of text length across classes was nearly the same, eliminating text length as a good learning signal.



We subsequently retrained the same models (Logistic Regression and Naive Bayes) on this balanced dataset using the complete TF-IDF vocabulary.

**Results After Balancing**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | 0.9917 | 0.99 | 0.99 | 0.99 |
| **Naive Bayes** | 0.9566 | 0.95 | 0.96 | 0.96 |

**Interpretation:**

* Even when removing any biases from class imbalance and document length, the models—Logistic Regression in particular—still had a performance level that was very high.
* This suggests that the models were learning more than just surface rules; they were picking up semantic or syntactical patterns that could reliably distinguish AI-generated from human-generated documents.
* Consistent performance establishes that our preprocessing, TF-IDF representation, and model choice for this stage are sound.

# Ethical Considerations

As AI-generated text becomes harder to identify as humanly written, tools that identify AI-generated text must balance important beneficial uses and potential ethical implications.

1. Who benefits?
   1. Educators: Preserve academic integrity by detecting AI-generated papers or essays.
   2. Content platforms: Important for content moderation of AI-generated spam or misinformation.
   3. Researchers could help with studying AI language models and their linguistic traces.
2. Who could be harmed?
   1. Non-native English speakers: Might write in mechanically rigid or simple prose, and be mistakenly identified.
   2. Neurodivergent or disabled authors: Could be unfairly identified as AI authors if they used assistive tools or offered a different way of typical expression.
   3. Creative writers: Poets or experimental writers may produce text that is outside normal language patterns and confuse classifiers.
3. Dataset & Model Biases
   1. Dataset Imbalance: Because there were initially many more human-written examples than AI-generated ones, if they are not balanced, the model would be biased toward over-predicting the majority class (human).
   2. Topic & Genre bias: If some topics are over-represented in one class, the model might merge the content type with authorship.
   3. Lexical bias: One of the classifiers relies solely on surface features, like word length or vocabulary richness, which may not tell us anything useful.
4. Risks of Misuse:
   1. Organizations could use these detectors without proper attention to predict and act upon -possibly punitive- decisions only by relying on predicted results; there will be no rules or support decision process for human verification.
   2. Students or individuals falsely predicted could be unfairly punished, stressing the real need for explainable AI and Human-In-The-Loop moderation.
5. Recommendations:
   1. Always co-locate detection systems with explainability and human verification.
   2. When developing tools, develop tools to flag uncertainty or display confidence levels (versus simply yes/no).
   3. Re-train and audit regularly, especially because the AI field is evolving and new AI-generated content may fool current detectors (e.g., new LLMs).