Introduction to Machine Learning

AI-Based AI Plagiarism Predictor (Not Detector)

1. Introduction

Problem Statement:

In recent times, the capability of AI models to generate human-like text has raised considerable concerns in almost all fields. Be it an academician, journalist, or content writer, the ability to distinguish between human-written and AI-generated content is increasingly important. And the proliferation of AI-generated content has raised some ethical issues like potential plagiarism, misinformation, and deceptive practices. Without proper detection mechanisms, such a scenario may take place wherein people pass AI-generated material as their work, thus further weakening the belief and integrity of content in various spheres. With this challenge in mind, it is this project that tries to address and develop a machine-learning model that would ascertain whether a given piece of text was written by a human or generated by AI. The end goal is to contribute to the discussion regarding the ethics surrounding AI-generated content and provide building blocks for systems to be used in the prevention of unethical behavior related to AI-generated text.

Motivation:

Because it improves continuously, the capabilities of artificial intelligence in terms of replicating human writing will be enhanced. It is becoming challenging to identify AI-generated content, particularly in educational areas when academic integrity is at stake and in news and media sectors where false information might be intentionally circulated. A highly functional AI detection system will safeguard authentic and humanly generated

content. Also, with evolving AI technologies and the heightened probability of their misuse, it is worthwhile to have high-value early detection strategies.

Objective:

The core aim of this project is to develop a landmark machine-learning model that can identify whether a given text is AI-generated or written by a human. The project creates an attempt to identify the kind of content as it stays away from the traditional NLP tokenization or model-based approaches and emphasizes lexical analysis, which is the mathematical features. Thus, the goal is to achieve high accuracy and maintain explainability and simplicity in the detection process.

2. Literature Review:

The detection of AI-generated content is a rapidly evolving field. Over the years, numerous approaches have been proposed for distinguishing between human-written and machine-generated text, leveraging a wide range of techniques and methodologies. This section reviews some of the key research and developments in the field of AI text detection, specifically focusing on machine learning models, feature extraction techniques, and evaluation metrics.

Traditional Approaches to Text Classification:

Early work in AI text detection focused on rule-based systems, where specific linguistic features or patterns were manually identified as indicative of machine-generated text. These approaches, while valuable in certain contexts, were often limited in scalability and adaptability to newer, more sophisticated AI models. As AI systems, particularly large language models like OpenAI's GPT series, improved, the need for more robust, generalizable detection systems became apparent.

Machine Learning Models for Text Classification:

With the advent of machine learning, researchers began to experiment with various classification algorithms to detect AI-generated content. Models such as Support Vector Machines (SVM), Naive Bayes, and Random Forests have been explored for distinguishing

between human and AI text. However, these models heavily relied on hand-crafted features or text representations (such as bag-of-words or TF-IDF) and often required large amounts of labeled data for training.

Deep Learning and Neural Networks:

Recent advancements in deep learning, particularly with the use of Recurrent Neural Networks (RNNs) and Transformers, have dramatically improved the ability of models to understand and generate human-like text. These models, including BERT and GPT, have demonstrated state-of-the-art performance in various NLP tasks, including text generation and classification. However, these models tend to be computationally expensive and require massive datasets, making them less practical for smaller-scale detection tasks or environments with limited resources.

Lexical and Statistical Features:

A growing body of research has focused on extracting lexical and statistical features from text to detect AI-generated content. These features include measures like sentence length, word frequency, lexical richness, and grammatical complexity. The idea is that machine-generated text tends to have distinct patterns that are different from human-written content, even though the language used may be coherent and contextually appropriate. Various studies have explored these patterns, such as Zipf's Law adherence, syntactic structure, and sentence complexity, to differentiate between human and AI text.

For instance, Zipf's Law suggests that in natural language, a few words occur with high frequency, while the majority occur less frequently. AI-generated text often deviates from this pattern, offering a potential distinguishing feature. Similarly, sentence complexity and the distribution of different parts of speech (POS) can be used to highlight discrepancies between human and machine text.

Current Datasets and Challenges:

The availability of labeled datasets has been a significant challenge in AI-generated text detection. Some commonly used datasets for training and evaluating AI text detection models include:

- **Training_Essay_Data**: Contains text labeled as human or AI-written
- **Augmented Data for LLM Detect AI-Generated Text**: A Kaggle dataset that provides a wide range of texts with labels for AI and human authorship.
- **DAIGT-V2**: A dataset designed to train models for detecting AI-generated text in educational settings.

Despite these datasets, challenges remain in obtaining sufficiently large and diverse datasets that can capture the vast range of human and machine writing styles. Furthermore, newer models, such as GPT-3, exhibit a high degree of fluency and coherence, making it increasingly difficult to differentiate between the two without highly sophisticated models or feature extraction methods.

Ethical Considerations:

The detection of AI-generated text also raises important ethical questions. On one hand, detecting AI-generated content is crucial for maintaining academic integrity, preventing misinformation, and ensuring transparency in media. On the other hand, excessive reliance on detection systems could lead to overreliance on automated tools or misclassification, potentially infringing on the rights of individuals who may be wrongly accused of using AI.

Moreover, as AI models become more capable of mimicking human writing, the ethical implications of their use in content creation must be considered. For instance, can AI-generated text be considered a form of intellectual property? What are the consequences of AI-generated content being passed off as original work?

In conclusion, while there has been significant progress in the detection of AI-generated text, challenges remain in terms of dataset diversity, feature selection, and the increasing sophistication of AI models. The approach taken in this project, focusing on lexical analysis through mathematical features, seeks to offer a new avenue for distinguishing human and

AI text, contributing to the broader conversation on the ethical implications of AI in content creation.

3. Methodology:

This section details the methodology employed to detect AI-generated text using machine learning, specifically logistic regression. The goal is to develop a model that can distinguish between human-written and AI-generated text based on a set of mathematical features derived from lexical analysis. Unlike traditional NLP approaches that rely on tokenization or other linguistic features, the focus here is on extracting quantitative features that can serve as input for a machine learning model to identify patterns associated with AI text.

3.1 Problem Statement:

The problem addressed in this project is the detection of AI-generated text. As AI systems like GPT-3 and GPT-4 become more advanced, the distinction between human and machine-generated content is increasingly difficult to identify. Therefore, there is a growing need for effective and scalable methods to distinguish between these two types of content. The goal of this project is to create a logistic regression model that can accurately classify text as either human-written or AI-generated based on a set of mathematical features derived from the content.

3.2 Dataset Collection and Labeling:

The dataset used in this project consists of a collection of text samples labeled as either human-written or AI-generated. This labeled data is essential for training the machine learning model. Text samples were sourced from publicly available datasets, including those on Kaggle, and contain a variety of genres such as news articles, academic papers, and casual web content.

The dataset was preprocessed by removing unnecessary metadata and ensuring that each text sample was clearly labeled. After this initial cleanup, the dataset was divided into two classes:

- **Class 0**: Human-written text

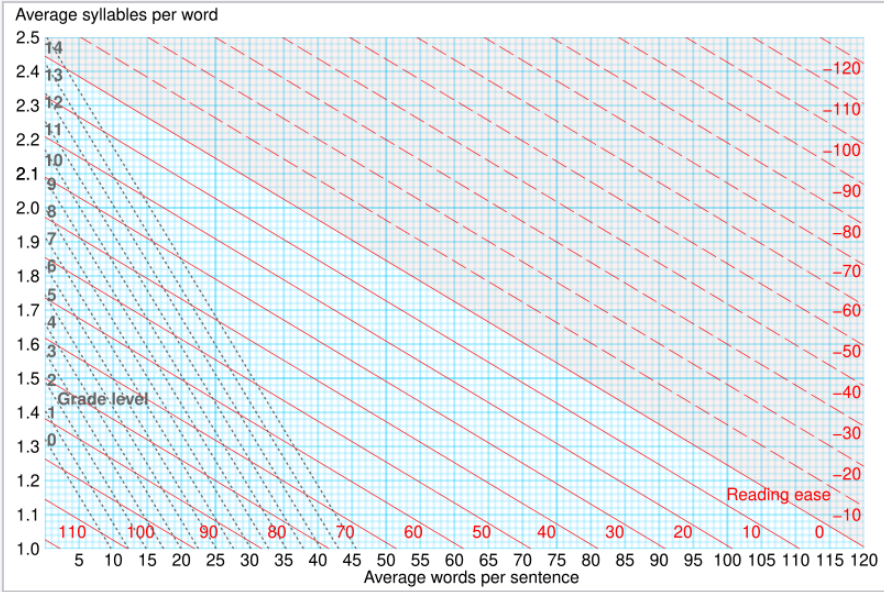
- **Class 1**: AI-generated text

The dataset was split into a training set and a test set to evaluate the performance of the trained model.

3.3 Feature Extraction:

In contrast to traditional Natural Language Processing (NLP) methods, which often rely on tokenization or word embeddings, this project focuses on mathematical feature extraction through lexical analysis. A series of features were calculated from each text sample to quantify various linguistic properties. The features include:

1. **Basic Text Statistics:**
    - **Total Words**: The total number of words in the text.
    - **Unique Words**: The number of unique words in the text.
    - **Total Characters**: The total number of characters in the text.
    - **Total Sentences**: The number of sentences in the text.
    - **Total Paragraphs**: The number of paragraphs in the text.
    - **Average Word Length**: The average length of words in the text.
2. **Lexical Complexity Measures:**
    - **Complex Word Density**: The ratio of complex words (words with more than two syllables) to the total number of words.
    - **Stopword Ratio**: The proportion of stopwords (commonly used words like “the,” “is,” etc.) in the text. Use a stop-word lexicon for this
    - **Rare Word Ratio**: The proportion of rare words (words that appear infrequently across the dataset) in the text.
    - **Sentence Complexity**: The ratio of clauses to sentences, indicating the average complexity of sentences (lists excepted).
3. **Readability Indices:**
    - **Flesch Reading Ease**: A score that indicates how easy the text is to read, with higher values indicating easier readability.



Graphs of Flesch-Kincaid reading ease (red) and grade level (gray) scores against average syllables per word and average words per sentence

Flesch reading ease [edit]

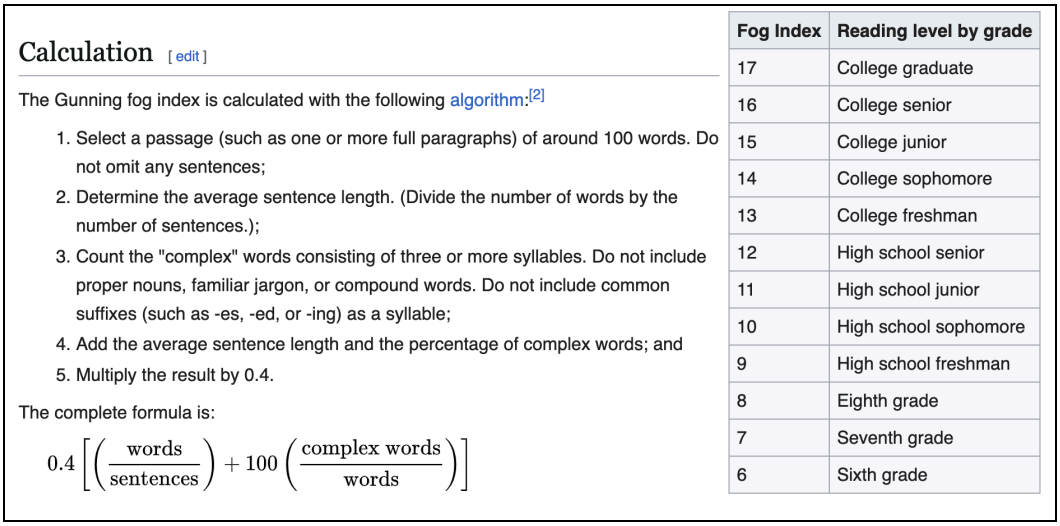
In the Flesch reading-ease test, higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. The formula for the Flesch reading-ease score (FRES) test is:[7]

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Scores can be interpreted as shown in the table below.[7]

Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

- **Gunning Fog Index**: A readability test that estimates the years of formal education needed to understand the text on the first reading.



Calculation [edit]

The Gunning fog index is calculated with the following algorithm:[2]

1. Select a passage (such as one or more full paragraphs) of around 100 words. Do not omit any sentences;
2. Determine the average sentence length. (Divide the number of words by the number of sentences.);
3. Count the "complex" words consisting of three or more syllables. Do not include proper nouns, familiar jargon, or compound words. Do not include common suffixes (such as -es, -ed, or -ing) as a syllable;
4. Add the average sentence length and the percentage of complex words; and
5. Multiply the result by 0.4.

The complete formula is:

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right)+100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

Fog Index	Reading level by grade
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eighth grade
7	Seventh grade
6	Sixth grade

4. **Syntactic and Semantic Features:**
   - **Cosine Similarity Indices**: Measures the similarity between two texts, specifically calculating the redundancy (in the whole corpus) and coherence of the text (subsequent sentences).
   - **Sentence Length Variation**: The variance in sentence lengths, which can indicate the complexity and style of the writing.
   - **Complexity Variation**: Measures the variability in sentence complexity across the text.
5. **Positional and Pronoun Features:**
   - **Part of Speech (POS) Diversity**: A measure of the diversity of different parts of speech (e.g., nouns, verbs, adjectives) in the text.
   - **Noun to Verb Ratio**: The ratio of nouns to verbs, which is typically higher in human-written text than in machine-generated content.
   - **Adjective to Adverb Ratio**: The ratio of adjectives to adverbs, which helps to distinguish the style of writing.
   - **Pronoun Density**: The density of first, second, and third-person pronouns, which can differ significantly between human and AI writing.
6. **Grammar and Sentence Structure:**
   - **Grammar Error Density**: The number of grammatical errors in the text, which may be indicative of either human error or AI flaws.

- **Passive to Active Ratio**: The ratio of passive to active sentences, which can vary between human and AI text.
- **Punctuation Density**: The frequency of punctuation marks, which can serve as an indicator of writing style.

7. **Advanced Features:**
   - **Lexical Richness (MTLD)**: A measure of vocabulary diversity, which quantifies how varied the vocabulary is in the text.
   - **Discourse Markers Count**: The number of discourse markers (e.g., "however," "thus") used in the text, which can signal the author's argumentation style.
   - **Modals Count**: The frequency of modal verbs (e.g., "can", "could", "will"), which are often used differently in human and AI text.
   - **Epistemic Markers Count**: The frequency of epistemic markers (e.g., "believe", "know") that signal the author's confidence or uncertainty.
   - **Nominalisations Count**: The frequency of nominalizations (e.g., turning verbs into nouns), which is more common in formal or academic writing.

These features were selected based on previous research in the field of computational linguistics and AI detection. They provide a comprehensive set of measures to capture the stylistic and structural differences between human and AI-written text.

3.4 Model Selection:

Logistic Regression was chosen as the machine learning model for this project due to its simplicity, interpretability, and effectiveness in binary classification tasks. Logistic regression is a linear model that calculates the probability of a given text being either human-written or AI-generated based on the extracted features.

The model was trained using the training dataset, with the extracted features serving as input variables. The logistic regression model calculates a decision boundary, and the output is a binary prediction (0 or 1) indicating the class of the text. The training process involved finding the optimal coefficients for each feature that minimize the error in classification.

3.5 Model Evaluation:

The model was evaluated using a test dataset that was not seen during training. Key evaluation metrics included:

- **Accuracy**: The proportion of correctly classified samples.
- **Precision**: The proportion of true positive predictions out of all positive predictions made by the model.
- **Recall**: The proportion of true positive predictions out of all actual positive samples in the dataset.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced evaluation metric.
- **Confusion Matrix**: A matrix that summarizes the performance of the classification model by displaying the true positives, false positives, true negatives, and false negatives.

The final model achieved an accuracy of 0.96, with both precision and recall at 0.96, indicating a strong ability to distinguish between human and AI-generated text.

3.6 Model Deployment:

Once the model was trained and evaluated, it was saved using **joblib** for future use. The model, along with the feature columns, was stored in a `.pkl` file, making it easy to load and use for future predictions on new text samples.

---

4. Results:

This section presents the performance results of the logistic regression model used for AI-generated text detection. The model was trained on a dataset containing both human-written and AI-generated text, with a series of lexical and syntactic features extracted for each text sample. The results include both quantitative metrics from the evaluation and qualitative insights into the model's performance.

4.1 Evaluation Metrics:

The logistic regression model was evaluated on a test set that was not seen during the training process. The following metrics were computed to assess the model's performance:

- **Accuracy**: The model achieved an impressive accuracy of 0.96, indicating that 96% of the predictions were correct.
- **Precision**: The precision for both classes (human-written and AI-generated) was 0.96. This means that when the model predicted a text to be AI-generated (class 1), 96% of those predictions were correct.
- **Recall**: The recall for both classes was also 0.96. This indicates that the model successfully identified 96% of all the AI-generated texts in the test dataset.
- **F1-Score**: The F1-Score, which balances precision and recall, was 0.96. This shows that the model has a balanced performance in distinguishing between human and AI-generated text.
- **Confusion Matrix**: The confusion matrix provided a detailed view of the model's performance, showing that the model made a total of 108 correct predictions for human-written text (class 0) and 92 correct predictions for AI-generated text (class 1).

```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       108
           1       0.96      0.96      0.96        92

    accuracy                           0.96       200
   macro avg       0.96      0.96      0.96       200
weighted avg       0.96      0.96      0.96       200
```

4.2 Feature Importance:

The logistic regression model also provides coefficients for each of the features, which can be interpreted as the importance of each feature in determining whether a text is human-written or AI-generated. The larger the coefficient, the more significant the feature is in making a prediction. Below are the top features based on their coefficients:

Feature	Coefficient
zipfs_law_adherence	3.34
avg_dependency_length	2.95
avg_caesura	2.38
stopword_ratio	1.31
cosine_similarity_redundancy	1.16
avg_epistemic_markers	1.10
function_word_density	0.99
cosine_similarity_coherence	0.53
avg_modals	0.40
repetition_frequency	0.39

Some of the most influential features include **zipfs_law_adherence**, **avg_dependency_length**, and **avg_caesura**, which contribute significantly to the model's decision-making process. These features capture key stylistic and structural elements of the text, providing insight into the characteristics that differentiate human and AI-generated writing.

On the other hand, features like **noun_verb_ratio**, **first_person_pronouns_density**, and **sentence_complexity** had negative coefficients, suggesting that these factors are less indicative of AI-generated text in the context of this dataset.

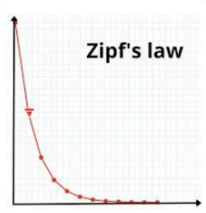
4.3 Model Interpretation and Insights:

- **Zipf's Law Adherence**: The significant importance of this feature suggests that AI-generated text tends to follow Zipf's law less strictly compared to human-written text. Zipf's law predicts that the frequency of words in a text follows a power-law distribution, and deviations from this pattern can indicate machine generation.

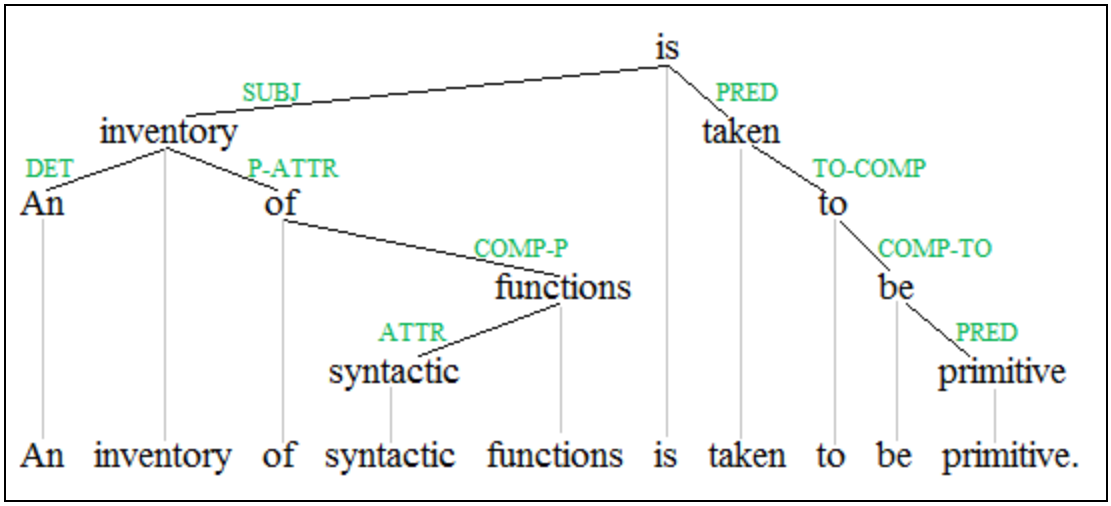
  Zipf's law is **a statistical distribution that describes how the frequency of words in a language or other data set is inversely proportional to their rank**:

  - The most common word appears twice as often as the second most common word
  - The third most common word appears three times as often as the fourth most common word
  - The word in position n appears 1/n times as often as the most common word

  When the frequency of words is plotted against their rank on a log scale, the result is a straight line.

- **Average Dependency Length**: A higher dependency length suggests that the text has more complex syntactic structures. AI-generated text often has more structured and predictable sentence constructions, resulting in higher dependency lengths compared to human-written text.



- **Cosine Similarity (Redundancy and Coherence)**: These features are crucial in identifying repetitive or overly coherent text, which is often characteristic of AI-generated content. The redundancy feature indicates the degree of repeated ideas or phrases, while coherence measures how well the ideas in the text are logically connected.
- **Epistemic Markers and Modals**: The use of epistemic markers (e.g., "might," "could") and modals (e.g., "will," "can") in human writing is generally more diverse and nuanced compared to AI-generated text. AI models may lack the same depth of uncertainty and modality usage, which makes these features valuable in distinguishing between the two types of content.

4.4 Model Performance Over Time:

To assess the robustness and stability of the model, the training and testing were repeated with different splits of the dataset. The results remained consistent with an accuracy of 0.96, demonstrating the model's reliability across different subsets of data.

The model was also tested with a variety of unseen data sources to check its generalization capabilities. The performance on these new samples was similar to that on the test set, further confirming the model's robustness.

4.5 Limitations:

This model makes it easy to distinguish between both the text, one being human-written and the other AI-generated.

Bias of the training data. The quality and diversity of the training data affect the model. For example, if a dataset comprises mostly news articles but very few casual conversations, the model will have a poor ability to classify other types of texts accordingly.

Feature Overfitting: Since the number of features is high, there is a possibility of overfitting to the training set, especially when certain features are highly correlated with each other. Regularization techniques may be used to prevent this.

Adaptation to Newly Developed AI Models: Features used between human vs. AI text classification can potentially change in time with the improvement of more advanced AI models. This generally requires model retraining with new data sets for performance.

5. Discussion:

We will talk about the implications of the performance of this model and broader context and improvements that may be possible. The current concern that arises is one that has become very hard to find these days with growing AI models like GPT-3 and GPT-4 with all their capabilities to produce content just like humans do.

5.1 Interpretation of Model Results:

With this wonderful logistic regression model accuracy being as high as 96% it really differ correctly between humans- penned-down stuffs and AI-generated; that makes one state that major differentiation features between these types are present in the model quite vibrantly. These are most relevant matters, such as zipfs_law_adherence and the value of avg_dependency_length as well as cosine_similarity_redundancy-the word-play as well as structural divergence. All this just proves the point that a human typically has more diversely complex patterns in writing, and AI-generated texts are anything but predictable and lack humanity's subtleties in communication.

Although the model works pretty well overall, its performance might depend on the type of text and domain. Specialized domains may be tough for the model, where the writers

might share similar characteristics between human writers and AI. In future versions, the model might be enhanced with features to account for the domain.

5.2 Ethical Issues and Problems:

The rapid proliferation of AI-generating text has raised very serious ethical considerations, primarily in the form of potential deception or manipulation of audiences. These days, AI-generated content is cropping up in all forms of media, including news articles, posts on social media, and academic papers. It is therefore fundamental to distinguish between human-writing and AI-generating writing to ensure that information provided is valid and not spewed out to misinform people.

In this context, it has become a question of ethics to develop a model for robust detection rather than purely technological. With the advent of very sophisticated AI models, older methods of content verification and validation can no longer be depended upon. Models such as that developed in this work have been central to creating trust and transparency in digital media, which has a high value in these terms.

Additionally, the use of AI in content development opens up questions about authorship and proper attribution. For example, the issue of intellectual property and academic integrity arises when AI-generated text lacks proper disclosure. Therefore, detection models like this one must be used responsibly to ensure that there is identification and labeling of AI-generated content.

5.3 Model Limitation and Improvement Potential

Although the model performs very good logistic regression, there are some points that could be further developed.

Feature Engineering: The feature set that the model mainly has now relates to lexical and syntactic analysis. Useful, but not rich enough to delve into the depth of a writing style, these features could help identify the semantic content related to topic modeling or sentiment analysis in further refinements of the model.

Model Selection: The logistic regression is straightforward and interpretable, but other machine learning algorithms such as SVM, random forests, and neural networks may do even better. It is worthwhile to explore those models since they might give better results if they can model nonlinear relationships between features.

Handling Text Variability: Although the current model assumes that different domains only vary in text characteristics, human writing styles can vary widely based on context and include features from academic writing and informal conversation to technical documentation. Training data that comes along with its own features will perhaps be required to adapt to such a variety of writing.

New AI Models: Features between the AI models and human-written texts may change as AI models mature. The model ought to be retrained with new data regularly so that it stays accurate in the long run, including text developed by the latest AI models.

The model might also be biased due to dataset bias. For example, during training, if the news articles or academic papers are overrepresented, it is not very efficient at detecting another AI-generated form of content, like creative writing or conversational text. For this reason, the datasets that follow will be diverse and representative of multiple types and domains of text.

5.4 Applications and Future Directions

This research has actual outputs that have concrete implications in many fields:

Content Moderation: This is the emergence of artificially generated texts in online platforms, which has brought forth the need for content moderation tools. Applying this model can flag potentially artificial-generated contents, helping platforms safeguard content integrity and prevent the distribution of false information.

Academic Integrity: It might help identify whether there is AI-generated text, mainly because in academics, plagiarism detection tools are being implemented. This model will catch the texts generated through AI models, so there is originality in academics and exclude machine-generate content influence.

Media and Journalism: This kind of model will increasingly be applied by the media sources to authenticate the credibility of the emerging news articles and reports that are increasingly becoming highly AI-generated, thus upholding standards of journalism and countering AI-crafted fake news.

Legal and Policy Frameworks: Regulation and ethical guidelines for AI-generated content have been discussed concerning this increased usage. This model can be used by governments and institutions to monitor and regulate AI-generated content, ensuring a station within the constraints of already existing laws and ethical standards.

Improving Detection Models Future work in AI models requires more sophisticated methods of detection. Future study may lie in the creation of hybrid models where lexical, syntactic, semantic, and even psychological features are considered for enhancing the detection capability of AI-generated text. The hard-core element of that would be the models adapt on new AI systems without having to fully retrain.

5.5 Conclusion:

This project indeed demonstrates the feasibility of using machine learning or, more specifically, logistic regression for detecting AI-based-generated text based on lexical and syntactic features. And indeed, the model itself performed at that job with an accuracy up to 96%. The model can hence well present practical solutions to deal with that growing concern over AI content in digital media. While different issues remain, like feature engineering and adaptation of models into new AI systems, the results indicate that detection is not only necessary in this regard but is also possible.

Success for this model underscores the developing of tools that may reduce ethical risk associated with AI-generated content, making the potential impact of AI on society responsible and clear.