Dissecting Authorship:
A Comparative Analysis of Human vs. Language Model Generated Texts

35644 | 26990 | 67892
The London School of Economics and Political Science
London, United Kingdom

Abstract

In the digital age, artificial intelligence has significantly impacted written communication, with advanced language models like OpenAI's GPT series increasingly challenging the distinction between human and machine-generated texts. This study examines the performance of various machine learning models in distinguishing between human and AI-generated content, focusing on three key datasets: DAIGT V2, WebText GPT-2, and Amazon Reviews. The models, including logistic regression, decision trees, and artificial neural networks, were evaluated based on their accuracy and ability to handle varying text complexities.

The findings reveal that logistic regression and artificial neural networks performed exceptionally well, achieving accuracy scores up to 99% and 98%, respectively, in distinguishing between human and AI-generated texts in the DAIGT dataset. Decision trees, while effective, showed variable results, particularly with increased text complexity, achieving a maximum accuracy of 78% in the Amazon Reviews dataset. The study also highlighted key differences between human and AI-generated texts, with human texts generally being longer and more variable.

1 Introduction

In the digital age, artificial intelligence has revolutionized many facets of life, with language models standing at the forefront of this transformation. These models generate text that rivals the nuance and complexity of human writing. From composing emails to generating news articles and crafting creative fiction, language models such as GPT have begun to permeate every corner of written communication, challenging our perceptions of authorship.

Language models have evolved significantly from the early days of simple rule-based algorithms to today's sophisticated neural networks. With the advent of advanced language models, especially the latest iterations like GPT-4, the line between text generated by machines and humans has become increasingly blurred. These models are now capable of producing writings that mimic human style and complexity with remarkable accuracy. This technological milestone poses significant challenges, particularly in fields that rely heavily on the authenticity of written communication, such as journalism, academia, and legal documentation. The ability to generate convincing texts can lead to issues of trust and credibility, where distinguishing between human and machine becomes not just a technical challenge but a societal concern. Given these challenges, this study seeks to investigate the use of machine learning models to distinguish human-authored and AI-generated texts. Specifically, this study seeks to answer the following questions:

(1) How do different machine learning models perform in distinguishing language AI-generated and human-authored texts?
(2) How does text complexity influence model performance?

This research trains a variety of ML models on several datasets to investigate the models' applicability for the task and evaluate whether the dataset used influences the relative performances. The datasets used are the "DAIGT V2 Train Dataset" [1], Open AI's "WebText GPT-2" [2], and the "Amazon Reviews 2023" by McAuely Lab [3]. As the "Amazon Reviews 2023" dataset does not contain LM-generated texts, novel language model-generated texts will be sourced from outputs of models like GPT-3.5 and GPT-4.

Overall, this study found that machine learning models can effectively distinguish between human and AI-generated texts, although their performance varies depending on the complexity of the texts and the models used. Logistic regression and artificial neural networks consistently performed well across multiple datasets, achieving accuracy scores as high as 0.99 and 0.98, respectively, for distinguishing AI-generated texts from human-authored texts in the DAIGT dataset. Decision trees showed varied results, particularly with deeper complexities, achieving a maximum accuracy of 0.78 in the Amazon Reviews dataset. The analysis also highlighted significant differences between human and AI-generated texts, with human texts generally being longer and more variable. The study concludes that machine learning has strong potential in distinguishing between human and AI-generated texts, and it sets a foundation for future research into varying datasets, model families, and architectures.

2 Related Work

The paper, titled "Feature-based detection of automated language models: Tackling GPT-2, GPT-3, and Grover," by Leon Fröhling and Arkaitz Zubiaga, explores a new method for distinguishing between human-written text and text generated by language models.[4] The key contribution of this research is the development of a simple, feature-based classifier that models intrinsic differences between human and machine-generated text, offering a cost-effective alternative to more computationally expensive methods.

The paper describes a feature-based approach for detecting texts generated by automated language models. The authors create features based on observed differences between human-generated and machine-generated text. These features aim to capture syntactic and lexical diversity, repetitiveness, coherence, and purpose of the text. The features are developed by analyzing typical flaws and limitations in the text generated by language models.

The authors consider various machine learning models including Logistic Regression, Support Vector Machines, Neural Networks, and Random Forests. The classifier is tested with samples from different language models (GPT-2, GPT-3, Grover) using a variety of datasets to evaluate performance across different generation techniques and model complexities. The model's performance is assessed using accuracy and the area under the curve (AUC) of the receiver operating characteristic (ROC), which provides a comprehensive metric that balances sensitivity and specificity.

For GPT-2 (small and large), the accuracy was higher for texts generated with likelihood-maximizing sampling methods, reaching up to 92.3% accuracy and 0.972 AUC for the most controlled sampling conditions. For GPT-3 and Grover models, which are more complex, the classifier faced more challenges but still managed accuracies around 78% and 86% AUC for GPT-3, indicating a capability to adapt to advanced model architectures.

The paper "Testing of Detection Tools for AI-Generated Text" provides a detailed evaluation of the effectiveness of various detection tools in identifying AI-generated text within an academic context.[5] It scrutinizes both the general functionality of these tools and their performance under different text manipulation techniques like machine translation and paraphrasing.

The researchers created English-language documents in six categories to cover a broad range of potential academic integrity challenges. The study tested 14 different detection tools which were evaluated on their ability to classify texts accurately. Each document type was tested for detection accuracy using a structured testing approach where each text was evaluated by the tools to determine if they were identified as human or AI-generated.

The tools varied significantly in their ability to correctly identify human-written and AI-generated texts. The best-performing tool, Turnitin, showed an accuracy of up to 76% under optimal conditions. However, even the best tools struggled with texts that had been altered through paraphrasing or manual editing, with accuracy rates dropping substantially in these scenarios. Overall, the results show that detection tools' reliability is compromised under conditions where text has been altered or manipulated.

The paper "ChatGPT or academic scientist? Distinguishing authorship with over 99% accuracy using off-the-shelf machine learning tools" focuses on differentiating texts written by academic scientists from those generated by ChatGPT.[6] This study addresses the challenge posed by the sophisticated text generation capabilities of AI models like ChatGPT, especially in the context of academic writing where the authenticity of authorship is critical.

The team identified key linguistic features that differ between academic scientists' writing and ChatGPT's outputs. They focused on aspects such as paragraph complexity, use of specific punctuation, and frequency of certain conjunctions and transitional phrases that reflect human academic writing styles. They compiled a dataset comprising both human-written texts (Perspective articles from the journal Science) and corresponding AI-generated texts produced by ChatGPT. Each article was paired with a ChatGPT output that was prompted by the article's title or a human-simulated title capturing the article's essence. The research team manually analyzed these texts to pinpoint 20 discriminative features. These included measures of sentence and paragraph complexity, diversity in sentence length, usage of specific punctuation marks, and the presence of particular "popular words". Using these features, they trained a model using a robust cross-validation technique.

The model achieved a paragraph-level classification accuracy of 94% and 99% on the document-level, indicating its robustness in identifying the nuances of language use at a fine granularity. This metric is particularly relevant in academic settings where entire articles or papers must be authenticated.

The study not only showcases a highly effective method for distinguishing between human and AI-generated academic texts but also underscores the potential of using tailored, feature-based machine learning models for authorship verification in scholarly contexts.

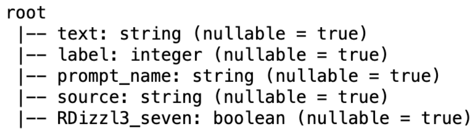
3 Methodology and Numerical Results

We conduct several separate experiments to explore language model capabilities in replicating and generating human text as well as to understand how the text complexity influences the model capabilities. We detail the data sets, models, and procedures employed in the analysis. Numerical outcomes, including the accuracy of classification models in differentiating human from language model-generated text are discussed within the respective model sections.

3.1 Datasets

3.1.1 DAIGT One dataset used for this project is the "DAIGT V2 Train Dataset", which was sourced from Kaggle. The DAIGT (Distinguishing AI-generated Text) dataset is specifically designed for tasks that involve differentiating between text generated by humans and by artificial intelligence. The dataset contains various pieces of text, along with labels. The dataset is valuable for research and development in the field of natural language processing (NLP), particularly in areas such as AI-generated content detection and text classification. The dataset provides a balanced and comprehensive set of examples, enabling robust training and evaluation of machine learning models aimed at distinguishing between human and AI-generated text. It contains 44868 unique data points (≈17000 AI-generated, ≈27000 human). Each includes text, label, source (source dataset), and prompt (original persuade prompt), as shown in Figure 1.

We begin by loading the data into a Google Cloud cluster using the API corresponding to the dataset, before unzipping and saving the dataset in a predefined bucket. We create a spark session and load the data into a spark data frame.



```
root
 |-- text: string (nullable = true)
 |-- label: integer (nullable = true)
 |-- prompt_name: string (nullable = true)
 |-- source: string (nullable = true)
 |-- RDizzl3_seven: boolean (nullable = true)
```

Figure 1: Schema of DAIGT V2 spark data frame

3.1.2 Amazon Reviews Millions of product reviews available on Amazon have been archived in various repositories. These reviews represent a rich corpus of human-generated text that varies in length, complexity, and tonality. We study the capability of GPT-3.5-turbo-0125 and GPT-4-turbo-2024-04-09 to rephrase a selected subset of these reviews in a manner that challenges a basic classification model to distinguish between texts produced by humans and those generated by the language model. Initially, a GPT-2 model was expected to contribute to the dataset, but it became apparent that the model was unable to rewrite the reviews as intended and consistently generated unrelated text instead. We analyze reviews from three distinct product categories: magazine subscriptions, appliances, and video games. The data we use was collected in 2023 by the McAuely Lab at the University of California San Diego [3] and was processed in a Jupyter Notebook. In the notebook, Spark is employed to prepare and transform the data from compressed JSONL files into a structured format suitable for prompting language models. The datasets are read into Spark DataFrames, where initial cleaning and filtering are conducted to remove unnecessary columns and standardize textual fields. We combine all reviews in a single data frame and sample 500 entries with a word count between 150 and 250 for each category. Although the data sets could support larger samples, the processing times and financial costs associated with OpenAI's API necessitate limiting our sample size to a manageable number. Nonetheless, we perform all data processing operations in a distributed manner to ensure scalability. Unfortunately, applying OpenAI API calls across distributed Spark nodes presents significant challenges. The need for each executor node to independently manage HTTP requests results in inefficiencies like higher latency, and increased failure risks. By consolidating the data into a single machine with pandas (for this step only), the application of API calls becomes more controlled and efficient. Of course, this is only feasible because the dataset size is relatively small. An alternative, fully distributed solution leveraging Kafka and a Flask application may provide a scalable distributed processing framework for larger datasets. The notebook's appendix outlines such a solution. Using the pre-processed data, we can then prompt the models using the following:

"Reforumlate the following text in your own words: [review]"

The resulting DataFrame now includes the primary columns: 'text' (human-generated text), 'GPT3.5-rewrite,' and 'GPT4-rewrite.' This pandas DataFrame is then converted back into a Spark DataFrame with an appropriate schema, allowing us to proceed with text cleaning. It's crucial to eliminate any artifacts or non-content symbols in both the human and language model texts to ensure the classification models focus on learning content differences rather than being influenced by unrelated text features, such as HTML artifacts present in some of the original human text.

Before moving on to model training, we reorganize the data to include a 'label' column that distinguishes between human- and LM-generated text, as well as a single 'text' column that houses both original and rewritten reviews. As we want to keep the GPT-3.5 and GPT-4 data separate, this process generates two distinct DataFrames of the aforementioned form. Any further pre-processing and the full data analysis are outlined in the section *Model Pipeline*. The primary goal of the analysis with this dataset is to investigate how effectively the employed language models can replicate concise, informal human text regarding various topics and inconsistent writing styles. Additionally, we aim to determine if classification performance significantly varies between the two language models and whether it depends on factors like text length or review category.

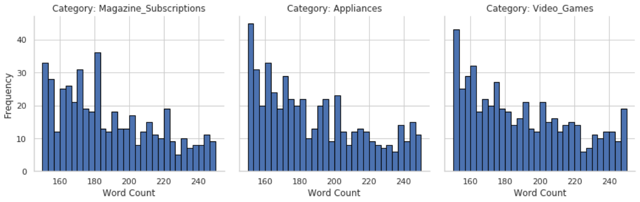


Figure 2: Word Count histogram of Amazon Review datasets: Magazine Subscriptions, Appliances, Video Games

3.1.3 WebText GPT-2 In the last data set we will be working with, the AI data comes from GPT-2 generated text. In particular, we have a data set for 4 versions of the GPT-2 model (where the different models vary in terms of the number of parameters used). Each of the datasets is comprised of a training set with 250,000 data points and a test set of 5,000 data points, wherein the data set corresponding to 1 version of GPT-2 will consist of text that was generated solely from that version of GPT-2. For a given version of GPT-2, the text in its data set was generated by random sampling: that is, the next token in the sequence was selected by sampling from the probability distribution of the autoregressive language model conditional on the prior sequence of tokens. We also investigate 1 other dataset (whose source is the same as that of the initial 4 datasets) where the text was generated using top-k 40 sampling, which truncates the distribution that is sampled from to the top 40 tokens in terms of probability of being the next token. The human data comes from text scraped from the internet (similarly to the AI dataset, the training and testing sets have 250,000 and 5,000 data points, respectively). Because of this, the text is highly varied in terms of structure and style, and so we would expect the classifier to struggle at its task. Another thing we would expect is that the longer an AI text is in terms of word count, the more likely it is to be being correctly classified (as the increase in words increases the exposure of the text). The data was processed on a Jupyter notebook running on a GCP cluster. The datasets are initially stored as .jsonl files, and read into the notebook as spark data frames. After extracting the datasets, the text is cleaned using a user-defined function to remove unwanted artifacts (like emails and HTML tags).

3.2 Model Pipeline

The model pipline was created using the spark.ml library. In the pipeline, the text data undergoes a series of transformations to prepare it for machine learning models. First, the data is tokenized using Spark's Tokenizer class, which splits the text into individual words. This step is crucial for breaking down the textual data into manageable units for further analysis.

Next, we apply HashingTF, a feature transformer that converts the tokenized words into numerical vectors based on their term frequencies. The parameter numFeatures, which defines the dimensions of the embedding space, is set depending on the size of the dataset to balance computational cost and accuracy.

Lastly, we utilize the IDF class, which stands for Inverse Document Frequency, to weight the raw term frequencies and transform them into TF-IDF features. This process adjusted features reflecting the importance of each word in the context of the entire corpus.

Overall, these pre-processing steps ensures that the text data was effectively represented in a numerical format, enabling efficient and meaningful input into subsequent machine learning algorithms.

3.3 Logistic Regression

The Logistic Regression model was initialized and included as the final stage in the pipeline discussed previously. The dataset was split into training and testing sets in an 80:20 ratio using a random seed for reproducibility (42). The model was trained using the training data, and predictions were made on the testing set.

3.3.1 DAIGT LR Model The model's performance was evaluated using the area under the ROC curve (AUC), resulting in an AUC of 0.99 using an embedding space of dimension 500, which indicates a very strong performance for this binary classification task. Additionally, the accuracy was assessed by comparing the predicted labels against the true labels. The model achieved an accuracy of approximately 0.97, with 8680 matches and 296 mismatches between the predicted and true labels.

These evaluation metrics suggest that the Logistic Regression model effectively distinguishes between AI and human-written text. The model's robust performance, as evidenced by both AUC and accuracy metrics, underscores the effectiveness of Logistic Regression in handling this binary classification task.

3.3.2 Amazon Reviews LR Model

Model	Logisitic Reg	Desiction Tree	ANN
GPT-3.5-Based	0.89	0.78	0.93
GPT-4-Based	0.88	0.75	0.95

Table 1: Accuracy Scores for Amazon Reviews Classification Models

All accuracy results for this dataset are presented in Table 1 more information regarding the decision tree and ANN results for this dataset can be located in their corresponding sections). The classification models generally perform well, consistently distinguishing between the original and rewritten reviews. The logistic model achieves an accuracy of 0.89 when differentiating between GPT-3.5-rewritten and original human text, while the score decreases slightly to 0.88 for reviews rewritten by GPT-4. These findings suggest that both GPT-3.5 and GPT-4 are comparably ineffective at imitating human text. Despite being a simple model, logistic regression can accurately distinguish whether a review was authored by a human or generated by a language model. We employ logistic regression for its simplicity and reliable accuracy to conduct further classification tasks on selected subsets of the Amazon review dataset we've compiled. Our objective is to investigate whether slight variations in features such as text length and content, or adjustments in the training methodology, have a notable impact on performance. Initially, we divide the dataset into shorter and longer halves based on the length of the original human reviews and then train the logistic model on each subset. The results show that this segmentation does not yield a noticeable difference in classification accuracy. Subsequently, we partition the dataset into three distinct sets based on review categories (magazines, appliances, video games). Once again, there's no notable difference in model performance between these categories. However, it's worth noting that the accuracy decreases slightly for each subset compared to the aggregate dataset, likely due to the reduced dataset size in each partition. Lastly, we alter the training setup so that the model only encounters either the original or the rewritten version of a review, rather than both pairs. This adjustment aims to assess whether the model benefits significantly from observing two versions of the same text during training, rather than just a mix of original human content and LM-generated content. Interestingly, employing this modified training setup does not impact the model's performance either.

3.3.3 WebText GPT-2 LR Model

The results in Table 2 describe the following: each cell represents the AUC of the LR model trained on the dataset of the GPT-2 model corresponding to the row of the cell against the test set of the GPT-2 model corresponding to the column of the cell. A striking feature of the results is that the performance of the model against all of the test sets is independent of which GPT-2 version we use as the source of our AI text for the training set of the model. In terms of the performance of the LR model, regardless of which data set it was trained on, we observe that there is deterioration as we test against data sourced from more complex GPT-2 models. The explanation behind this phenomenon is that text generated from more complex models, i.e. models with a larger number of parameters, is more human-like, on an account of being sourced from a more powerful model, and therefore harder to detect as being AI in origin.

The other set of results we have is the of the model trained on GPT-2 medium against the test set which was also generated from GPT-2 medium, but using the top-k 40 text generation strategy. The test accuracy we got was 0.40, which shows that our LR models are non-transferable against data generated under a different mechanism to that of the data which the model was trained on. Finally, we were able to make a plot of the size of the AI text (in terms of word count) against the proportion of text with that size which was classified correctly. We expected that this plot would be downward trending, however, this did not happen and we instead observed a noisy graph.

	Small	Medium	Large	Extra Large
Small	0.79	0.84	0.61	0.60
Medium	0.77	0.85	0.58	0.59
Large	0.77	0.85	0.58	0.59
Extra Large	0.77	0.85	0.58	0.59

Table 2: Performance of LR models, trained on each of the 4 GPT-2 datasets, against the 4 test sets

3.4 Decision Tree

For the Decision Tree model, a range of depths was explored to evaluate its performance in distinguishing between AI and human-written text. The pre-processing steps for the input data remained as outlined.

A loop was set up to train and evaluate Decision Tree models with several depths. Each model was included as the final stage in a pipeline, which also contained the tokenizer, hashing, and IDF transformers. The dataset was split into training and testing sets in an 80:20 ratio using a random seed (42) for reproducibility. For each depth, the Area Under the ROC Curve (AUC) and accuracy were computed to assess model performance.

3.4.1 DAIGT DT Model Decision trees trained on the DAIGT model used two sets of embedding dimensions, 100 and 500, with the latter providing higher scoring results. Figure 3 shows the AUCs and accuracies across depths one to five where the embedding space dimension was set to 500 (for dimension 100 see Appendix B, Figure 5). The results indicate that the accuracy generally improved with increasing depth, but the AUC did not follow the same trend, peaking at depth 1 and declining thereafter.

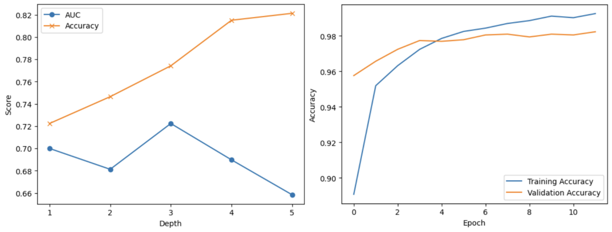


Figure 3: LEFT: DAIGT trained Decision Tree AUC and accuracy across depths (embedding space of dimension 500), RIGHT: ANN accuracy trained on DAIGT dataset (embedding space of dimension 250, 50 epochs, early stopping)

The decrease in AUC suggests that the model's ability to distinguish between AI-generated and human-written text is diminishing. With a decrease in AUC, the true positive rate (sensitivity) might be decreasing relative to the false positive rate (1-specificity). This means that the model's ability to correctly identify AI-generated text instances (true positives) among human-written texts is reducing, leading to a poorer discrimination ability.

The increase in accuracy indicates that the overall proportion of correctly classified instances (both AI-generated and human-written) is improving. However, due to the class imbalance, this improvement in accuracy could be primarily driven by the model's ability to correctly classify the majority class (human-written text), as there are more instances of it.

In this scenario, the model may prioritize maximizing accuracy by focusing on correctly classifying the abundant class (human-written text), even at the expense of its ability to distinguish the minority class (AI-generated text). As a result, while the model achieves a higher overall accuracy, it becomes less effective at correctly identifying AI-generated text instances, leading to a decrease in AUC.

3.4.2 Amazon Reviews DT Model Using a decision tree for author classification in the Amazon review dataset results in a slightly lower model score compared to logistic regression. Accuracy is significantly influenced by model depth. The highest accuracy score is attained at a depth of four with 0.78 (though a depth of six provides comparable accuracy results with a higher AUC). For the GPT-4 generated reviews, accuracy is again slightly lower with 0.75 at a model depth of six.

3.4.3 WebText GPT-2 DT Model We trained a decision tree model on the GPT-2 small data set, wherein the depth of the model was set to 5. We observed that the AUC fell from 0.79 to 0.60, when compared to the logistic model trained on the same dataset.

3.5 Artificial Neural Network

The final family of models trained were Artificial neural networks (ANNs). The architecture was a simple sequential model with the input layer defining the input shape as the number of dimensions of the embedding space used, with the default having been 100. This layer is followed by three pairs of dense followed by dropout layers. The dense layers have neurons 128 → 64 → 32, whilst the dropout layers are set to 0.3 → 0.2 → 0.1. The output layer uses the *sigmoid* activation function due to the binary nature of the classification task.

All ANNs were trained for 50 epochs using the *binary cross entropy* loss function, *Adam* optimizer with the learning rate set to 0.0005 and the metric set to *accuracy*. Furthermore, a callback function in the form of an early stop was used. It monitored the validation loss with that *patience* set to 5.

This approach allows the neural network model to continue training even if the monitored metric doesn't improve for a few epochs. This prevents premature stopping due to temporary fluctuations in performance and gives the model more opportunity to reach its optimal state in a time-efficient manner. Additionally, setting *restore_best_weights=True* ensures that the model's best performance during training is retained, even if it experiences temporary performance dips. This helps prevent overfitting and enhances the model's generalization ability, as it stops training at the point of best validation performance. Together, these strategies offer a robust and flexible approach to early stopping, balancing the need to explore potential improvements while minimizing the risk of overfitting.

3.5.1 DAIGT ANN Model Two different ANNs were trained using the DAIGT dataset. One with embedding space dimensions of 100 and one with 250. Again, the larger embedding space led to better performance. Upon examining the model's accuracy of the better performing embedding space of 250 in Figure **??** (for dimension 100 see Appendix B, Figure **??**), the training and validation accuracy improved steadily in the initial epochs, with both metrics stabilizing around an accuracy of approximately 0.98. The use of early stopping ensured that the model stopped training before overfitting occurred, resulting in an optimal model that generalizes well to unseen data. The validation accuracy closely followed the training accuracy, indicating that the model was not memorizing the training data but rather learning meaningful patterns, which is crucial for robust performance on new inputs.

The ANN's performance demonstrates that it effectively learned to distinguish between AI and human-written text, achieving high accuracy on the test set. The steady increase in accuracy, coupled with the early stopping callback, illustrates a successful training process. A high test accuracy of 0.98 confirms the model's effectiveness, validating the choice of architecture and hyperparameters for this binary classification task. The small gap between training and validation accuracy suggests that the model's generalization capabilities are strong, indicating that the network is well-suited for this type of text classification problem.

3.5.2 Amazon Reviews ANN Model An artificial neural network was also trained on the two available review datasets (GPT-4 and GPT-3.5). Given the limited dataset size, we expanded the embedding space to 1000 dimensions. Training proceeded swiftly, with validation accuracy rising rapidly in the early epochs before leveling off after around 10 epochs, prompting an early stop. Accuracy for the GPT-3.5 data is 0.93. Interestingly, evaluation accuracy for the GPT-4 data is slightly higher, at 0.95. The reason for this discrepancy could be that the neural network more easily identifies patterns specific to GPT-4's text generation style. GPT-4, being more advanced, may have distinctive characteristics that the model can distinguish, despite its improved ability to imitate human text. This finding contrasts with other models, which generally show higher accuracy for GPT-3.5 due to its simpler architecture. Intuitively, this makes sense, as GPT-3.5's simpler design may struggle to convincingly replicate human text, making it easier to identify. Overall, the differences in classification accuracy between the two datasets do not seem to be highly significant, so these figures should not be overinterpreted.

3.6 Using GPT-4 for Classification

In an additional experiment, we use OpenAI's API once more to differentiate between human-authored and LM-generated text. From the Amazon review data generated by GPT-3.5, we select a sample of 100 and prompt GPT-4 to classify each text as either original human content or GPT-3.5-rewritten text. The specific prompt used is:

"You will see a brief product review. The review is either an original human-written review or a human review that has been rewritten by GPT-3.5. Decide whether the review has been rewritten by the language model or not. Output '1' for rewritten and '0' for original. This is the review: [review]."

Classification performance is surprisingly low, with only 55.23% of texts being correctly classified. In this binary problem, this result indicates that the model cannot reliably make accurate predictions. To investigate further, GPT-4 was prompted to explain its reasoning. It appears that GPT-4 holds the mistaken assumption that human-authored texts are of a higher quality than LM-generated content. As a result, it incorrectly interprets the raw, unstructured nature of many human-written reviews as characteristics of a language model that struggles to produce clear, structured text. Visual inspection suggests the opposite might be true: while reviews rewritten by GPT-3.5 are of high textual quality, coherent, and well-structured, the original human-authored content often lacks structure, contains errors, and has limited flow and coherence. This aligns with expectations, as human reviews are typically written quickly, with little planning or consideration for textual quality. Presumably, this distinct difference in textual patterns enables the other models to achieve such high prediction accuracy. There could be several reasons why GPT-4 struggles with this classification task. First, the prompt provided may not sufficiently guide GPT-4 toward recognizing the specific linguistic nuances that differentiate rewritten from original reviews. The underlying training data used to develop GPT-4 likely contains diverse writing styles and patterns, leading the model to generalize its assumptions about human and machine-authored content, which may not align with this particular dataset's characteristics. Second, GPT-4 might inherently possess biases favoring certain linguistic traits that are more common in its training corpus. These biases could influence its belief that human-written text is more polished and organized, even though the opposite is often the case for quick, informal Amazon reviews. Third, the model might be overly confident in its internal heuristic about writing quality, causing it to dismiss errors and inconsistencies that typically distinguish human-authored content. This overconfidence may prevent it from accurately distinguishing between reviews rewritten by GPT-3.5 and original reviews.

4 Text Analysis

We created a custom function to perform a comprehensive textual analysis on a given PySpark DataFrame containing textual data. It processes the texts and calculates various statistical metrics based on the specified text and label columns. The purpose of the function is to analyze and compare the textual content based on differently labelled texts.

The function first replaces any non-breaking spaces (\xa0) in the text column with regular spaces to clean up the data. The cleaned text is then split into words using spaces as the delimiter. After splitting the text into words, the function filters out any empty strings from the resulting list. The function then splits the text into sentences using periods (.), exclamation marks (!), and question marks (?) as delimiters, using the improved regular expression [.!?]+ which accounts for one or more of these characters. Additionally, the function explicitly removes empty strings after splitting the text into words to ensure cleaner tokenization.

For each label, the function calculates the average and standard deviation of the number of words, the number of sentences, and the words per sentence, while also counting how many texts correspond to each label. The function then identifies the top five most common words for each label, excluding common stop words such as "and" and "the" and common punctuation marks. To enhance the relevance of the output, the function uses a set of common punctuation marks to filter out unwanted characters. It compiles the statistics and common words into a dictionary for each label and returns the overall result.

4.1 DAIGT Texts

Table 3 presents a comprehensive textual analysis of both human and AI-generated texts found in the DAIGT dataset. The table highlights differences between the two types of texts in terms of their structural and linguistic characteristics.

Statistic	**AI**	**Human**
Average Number of Words	324.89	413.89
STD Words	91.97	188.51
Average Number of Sentences	19.06	22.31
STD Sentences	6.31	10.15
Average Words per Sentence	17.64	19.96
STD Words per Sentence	3.4	11.82
Label Count	17497	27371
Most Common Words	**AI Freq.**	**Human Freq.**
help	1.23	
also	1.33	
This	1.34	
students	1.82	2.87
I	2.45	2.26
would		2.75
people		2.20
could		1.53

Table 3: Statistics and 5 most common words with their average frequency per text for both human and AI texts in the DAIGT dataset

The table shows that human-generated texts tend to be longer than AI-generated texts, with an average word count of 413.89 compared to 324.89. This aligns with the common understanding that human writers often provide more detailed and nuanced information. The standard deviation of the word count for human-written texts (188.51) further supports this, suggesting a wider variance in human responses. Although still high, AI-generated texts display more consistency, with a lower standard deviation of 91.97.

The average number of sentences is also higher for human-generated texts (22.31) compared to AI-generated texts (19.06). Again, the standard deviation for human-generated texts (10.15) is larger than that for AI-generated texts (6.31), indicating greater variability in the length of human-written content.

The smallest difference can be seen in the average number of words per sentence with 19.96 for human-written texts and 17.64 for AI-generated ones. However, the standard deviation is substantially higher for human texts (11.82) than for AI texts (3.4). This suggests that while AI-generated sentences tend to maintain a more uniform length, human-written sentences exhibit greater variability, potentially reflecting more complex sentence structures and narrative styles.

The most common words in the dataset offer additional insights into the nature of the texts. For AI-generated texts, the top five words align with common narrative constructs and indicate an attempt to engage with educational or informative topics.

For human-written texts, the most common words are generally similar but also include more conversational or suggestive terms such as "would" and "could".

The disparities in word and sentence statistics suggest that models leveraging features such as average word count, sentence count, and words per sentence may be effective in differentiating AI-generated texts from human-written ones. Additionally, focusing on the variability of these features, as indicated by their standard deviations, could further improve classification accuracy. On the other hand, the large variability between individual shorter texts might hinder models from using these textual features to classify texts of similar lengths. Future work should explore interpretability techniques to understand if and how the models studied in this paper utilize the insights gained from the text analysis. This would allow for the development of pre-processing methods and model architectures that incorporate more sophisticated feature extraction techniques to improve upon the initial results.

4.2 Amazon Review Texts

Table 4 presents a detailed comparison of the human and AI-generated texts in the Amazon Review dataset. It shows that the average number of words for human-written reviews is higher than that for both GPT-3.5 and GPT-4 generated texts, with humans averaging 189.19 words, while GPT-3.5 and GPT-4 average 127.96 and 166.98 words respectively. This suggests that human reviewers tend to provide more detailed feedback. Similarly, the average number of sentences is also higher for human reviews at 13.39 sentences, compared to 7.95 for GPT-3.5 and 10.11 for GPT-4. This implies that humans tend to write longer and more elaborate sentences, which could be indicative of their natural writing style.

The average number of words per sentence across all text types shows a notable consistency, with the figures closely clustered around the mid-teens. Specifically, human reviews average 15.91 words per sentence, while GPT-3.5 and GPT-4 models generate slightly more at 16.29 and 16.86 words per sentence, respectively. This close range suggests that all text types maintain a similar sentence complexity and length, illustrating that the AI models have been effectively trained to mimic human-like sentence structuring. This similarity in sentence length among the different sources of text could pose a challenge for machine learning models trying to distinguish between them based purely on this metric.

The standard deviations for all metrics are generally higher for human-written reviews, suggesting greater variability in human writing. For example, the standard deviation for the number of words per sentence, where the standard deviation is 9.3 for human reviews, while GPT-3.5 and GPT-4 have standard deviations of 2.61 and 2.7, respectively. The exception to this pattern is in the number of words, where GPT-3.5 exhibits a slightly higher standard deviation (31.67) compared to humans (28.77).

This indicates that while AI-generated texts tend to maintain a more uniform structure, human-generated texts display greater diversity, potentially due to the natural variation in human expression and writing styles. This variability could be a distinguishing factor for machine learning models designed to classify whether a text is generated by an AI or a human, although the increased variability between individual shorter texts might pose a challenge in using these textual features for classification when the texts have similar lengths.

The most common words across all reviews are generally similar, with “I” and “The” being frequently used. However, the frequency of these words differs, with human reviews using “I” more frequently (5.0) compared to GPT-3.5 (3.5) and GPT-4 (3.53). This could be indicative of more personal or subjective reviews written by humans. Similarly, the word "like" appears more frequently in human reviews (0.64) compared to GPT-3.5 (0.45) and GPT-4 (0.49), possibly reflecting a more conversational style.

4.3 WebText GPT-2 Texts

Table 5 offers a description of the text data found in GPT-2 Medium, Medium top-k 40 and WebText datasets. Immediately we can see that there is, again, higher variability present in the human text versus the AI in terms of the average number of words and

Statistic	**Human**	**GPT 3.5**	**GPT 4**
Avg no. of Words	189.19	127.96	166.98
STD Words	28.77	31.67	26.49
Avg no. of Sentences	13.39	7.95	10.11
STD Sentences	5.03	2.0	2.04
Avg Words / Sentence	15.91	16.29	16.86
STD Words / Sentence	9.3	2.61	2.7
Label Count	1505	1505	1505
Most Common Words	**Freq.**	**Freq.**	**Freq.**
I	5.0	3.5	3.53
The	1.19	1.28	1.21
magazine	0.56	0.58	0.58
like	0.64	0.45	0.49
one	0.59		
It		0.47	
This			0.46

Table 4: Statistics and 5 most common words with their average frequency per text for both human and AI texts in the Amazon Review dataset

Statistic	Medium	Medium top-k 40	Human
Avg no. of Words	474.74	520.22	424.42
STD Words	268.68	280.38	266.84
Avg no. of Sentences	22.87	26.44	25.43
STD Sentences	13.79	15.30	17.79
Avg Words / Sentence	21.62	20.23	18.19
STD Words / Sentence	12.24	11.19	16.22
Label Count	250,000	250,000	250,000
Most Common Words	**Freq.**	**Freq.**	**Freq.**
I	1.44	3.08	1.80
The	1.35	2.14	1.64
one	0.83	1.26	0.84
would	0.63	1.37	0.78
like	0.81		
also		1.28	
said			0.78

Table 5: Statistics and the 5 most common words with their average frequency per text for both human and AI texts in the WebText GPT-2 dataset

average number of sentences (for example, the STD for the number of sentences is 18 for human text, whereas its 14 and 15 for Medium and Medium top-k 40 text, respectively). When comparing between the random sampling and top-k 40 datasets, we can see that the latter registers a higher average number of words and sentences (520 vs 475 and 26 vs 23, respectively). An explanation for this could be that the mechanism by which text is generated in the GPT-2 top-k 40 model limits the word choice of the text. Thus, words that could capture a deep meaning are excluded from the text, which induces the model to generate multiple words to capture the same meaning (this, of course, will drive up word and sentence count).

In terms of the most common words from each text, we can see that there is strong agreement across the data sets in terms of which words are the most popular. What's interesting to note, however, is that Medium top-k 40 exhibits a higher magnitude for its frequency of the shared popular words in comparison to the other 2 data sets. This fits neatly with what we said above, as this could be due to the Medium top-k 40 model requiring the usage of the same words multiple times to capture the same meaning of a single word which it is not allowed to use. The final figure we would like to mention is that of the size of the vocabulary of each of the corpora. Their sizes are 8363290, 2761432 and 4536841 for Medium, Medium top-k 40 and Human, respectively. This, once again, is explained by the differences in the mechanism for text generation found the 2 GPT-2 models, in that the limits placed on word choice for the Medium top-k 40 leads to a far smaller vocabulary size. It's interesting to note that the human corpus' vocabulary size is in between the 2 models', which could indicate that humans generate text according to their limited vocabulary (which is equivalent to the Medium top-k 40 placing restrictions on its word choice) but that this limit is less so than that placed on the Medium top-k 40 model. Thus in the future it might be prudent, in the aim of making text more human like, to increase the sampling size from 40 to a value which best agrees with that of the average human.

5 Conclusion and Future Work

This study evaluated the performance of various machine learning models in distinguishing between human-authored and AI-generated texts across multiple datasets, including DAIGT, Amazon Reviews, and WebText GPT-2. We aimed to assess how different models coped with this challenging task and explored the textual features that could help differentiate between human and machine-generated content.

Our findings indicate that logistic regression and artificial neural networks consistently showed strong performance, achieving high accuracy and AUC scores. Decision trees, while effective, demonstrated variability in results, particularly at increased depths. The analysis also revealed significant differences between human and AI-generated texts: human texts were typically longer and displayed more variability, whereas AI-generated texts showed more uniform structures.

Key insights include that logistic regression and ANNs excelled across most datasets, while decision trees struggled with deeper complexities and longer texts. Human texts showed greater complexity in sentence structures and vocabulary, challenging the models differently than the more consistent AI-generated texts.

For future research, several areas appear promising. Investigating multiple languages and cultural contexts could provide broader insights into the effectiveness of text classification models. Techniques like resampling could be employed to address class imbalance and improve model performance. Developing interpretability methods would help in understanding how models use textual features, guiding improvements in model architecture and pre-processing.

In conclusion, the study underscores the potential of machine learning in distinguishing between human and AI-generated texts and acts as a foundation for further research into varying datasets, model families, and architectures.

References

[1] Darek KŁECZEK. Daigt v2 train dataset, Nov 2023.

[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[3] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[4] Leon Fröhling and Arkaitz Zubiaga. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443, 04 2021.

[5] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1), December 2023.

[6] Heather Desaire, Aleesa E. Chua, Madeline Isom, Romana Jarosova, and David Hua. Chatgpt or academic scientist? distinguishing authorship with over 99

A Individual Contributions

- 35644: Model Pipeline & Text Cleaning, Text Analysis function, DAIGT dataset analysis
- 26990: Automated GPT Prompting, Amazon Review dataset analysis
- 67892: WebText GPT-2 dataset analysis

B Additional Figures

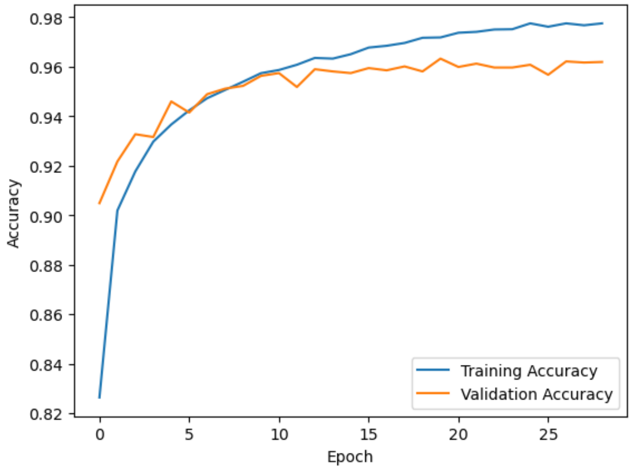


Figure 4: ANN accuracy trained on DAIGT dataset with an embedding space of dimension 100 over 50 epoch training with early stop

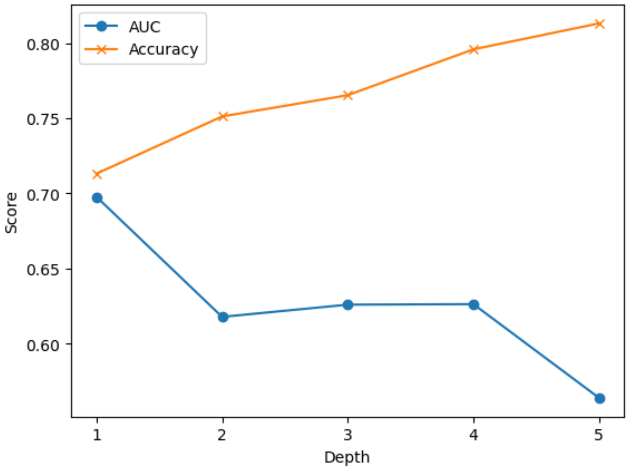


Figure 5: DAIGT trained Decision Tree AUC and accuracy over different depths with an embedding space of dimension 100