Evaluating AI Detectors

Group 3

Alphan Yang, Anshul Arvind, and Chris Mongelli

Algorithms in Society Section 2

Introduction

Essay writing, a core and ancient focus of the education system, has significantly decreased in difficulty through the past few centuries. The introduction of the typewriter, and a few decades later, the keyboard and mouse, as well as internet search engines, have left a lasting impact on writers' efficiency. In the widespread digital landscape of the 21st century, Artificial Intelligence (AI) chatbots have emerged as the latest and greatest essay-writing assistant. Whether it be to correct grammar, enhance style, or just flat-out write the whole essay, many students have fallen to these chatbots as a crutch to lean on. The sheer versatility of the highly-adaptive tool almost makes it difficult not to use it throughout the writing process. However, with this shift in approach towards AI reliance, many issues have arisen regarding the ethics of AI use, as well as the difficulties behind pinpointing when a text is written or assisted by AI.

Though it is a very useful tool for students and many others alike, AI-generated work goes against the values of academic integrity, since at its core, it is a souped-up predictive text algorithm that takes its ideas from the vast library of resources at its disposal. In turn, no outputs provided by the chatbot contain original thought, and responses are typically lifted from their knowledge base. This creates a pressing need in academic settings for reliable AI detectors since it can be difficult for the naked eye to distinguish between what is original and what is computer-generated. These detectors are designed with the intention to accurately discern the origin of the writing. The emergence of these AI detectors, however, introduces a new issue: **How reliable are these monitors of authenticity?** Through this project, we will dive into a few research questions regarding the intricacies of these chatbot detectors.

Can we Train a Prediction Model to Detect AI-written Text with our Dataset?

In the digital age, distinguishing between human and AI-generated text has become a difficult challenge. In response, we decided to have a go at training a predictive model capable of detecting AI-authored essays. Utilizing linguistic features such as readability, diversity, entropy, etc., we plan to detect patterns unique to human and AI writing. This section will explore our model's construction.

Defining Textual Metrics:
To begin, here are the linguistic features our model will be calculating and evaluating:

- **Readability**
  - This calculation is based on the Flesch Reading Ease Score. It indicates how easy a text is to understand. It takes into account total words, sentences, and syllables to calculate its score. The higher the score, the easier the text is to read.
- **Percent SAT Words**
  - The percentage of words in the essay that are considered SAT-level vocabulary (based on a predefined list of SAT words). A higher percentage indicates a more advanced vocabulary.
- **Simplicity**
  - The percentage of common words found in the essay (based on a predefined list of common words). Unlike the Percent SAT score, the higher the percentage, the easier the text is to read and understand.
- **Lexical Diversity**
  - A measure of the percentage of unique words used in the text – used as a way to calculate the text's complexity.
- **Burstiness**
  - Measures the variability in sentence length within the text, which impacts readability. Higher burstiness suggests that a text has diverse sentence lengths throughout. Burstiness is calculated using the standard deviation of the lengths of sentences in the essay.
- **Average Sentence Length**
  - Simply a calculation of the mean number of words per sentence. This metric also affects readability, where a longer average length indicates a more complex piece of writing.
- **Text Entropy**
  - Entropy is a concept that measures unpredictability or randomness of information. In this case, a higher entropy indicates the distribution of words is more uniform, which suggests more complex language use. The calculation for this metric assesses the probability distribution of words in the given text.
- **N-Grams**
  - N-grams are continuous sequences of 'n' items from a sample of text. This model uses n-grams as features that can be indicative of writing patterns typical of human or AI writers, as opposed to calculating a score directly from the n-grams.

These metrics are derived using established linguistic methodologies. These include the CMU Pronouncing Dictionary (for syllable counts), and some natural language processing techniques for the text analysis. For SAT words, we used NLTK Corpus as an approximation for this list. For common words, we used the first 100 words from NLTK.

Applying Metrics to Essay Data and Preparation for Machine Learning:

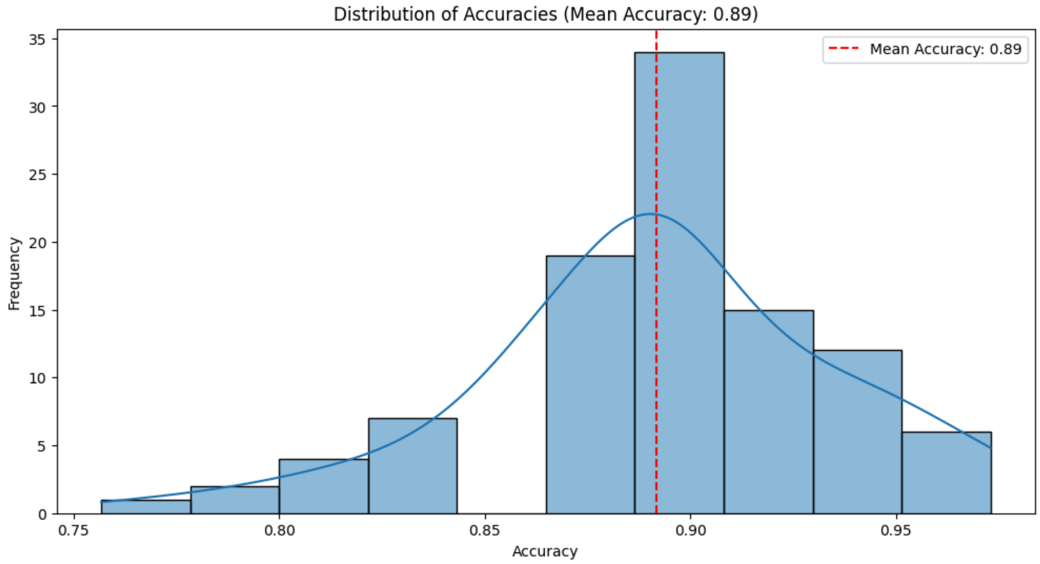
- After defining our calculations, multiple new columns applying each of these calculations to the set of essays and returning their respective scores were added to the dataset. Another column was added distinguishing how the text was written (AI or human) as a binary value, 0 if human-written, 1 if AI-written. This value was obtained through the "Written By" column already in the dataset, which distinguishes if the text is entirely written by a human (0), written by a human but the grammar was corrected by AI (0), partly written by a human, partly written by AI (1), or completely AI-generated (1).
- In some cases throughout the dataset, data cleansing was in order, since there were missing values for some detector scores. To fill these missing scores in, we simply used the median values from the rest of the dataset for the respective detector.
- We also created a new column called "N_Gram_Text", which contains a string of the entire sequence of N-grams for each given essay, which we will use for the text-vectorization process.
- After copying the data frame, we then extracted all the feature columns (All metric columns), and labeled them "x_all", and we extracted the AI-written binary column and labeled it "y_all"
- We then split the dataset into training and testing sets, with 20% of the data reserved for the testing model.
- A 'ColumnTransformer' was set up with 'TfidVectorizer' to convert the "N_Grams_Text" columns into TF-IDF features, which reflect the importance of words in a collection of documents. All other columns are left unchanged.

Machine Learning:

- A “Pipeline” was created to combine the TF-IDF transformation and a “RandomForestClassifier” as the classification model.
- The pipeline was used to train the RandomForestClassifier on the training set (“X_train”, “y_train”)
- The trained model was used to predict the labels on the testing set “X_test”, and “classification_report” was printed, which provides a detailed performance analysis of the

	precision	recall	f1-score	support
0	0.88	0.88	0.88	16
1	0.90	0.90	0.90	21
accuracy			0.89	37
macro avg	0.89	0.89	0.89	37
weighted avg	0.89	0.89	0.89	37

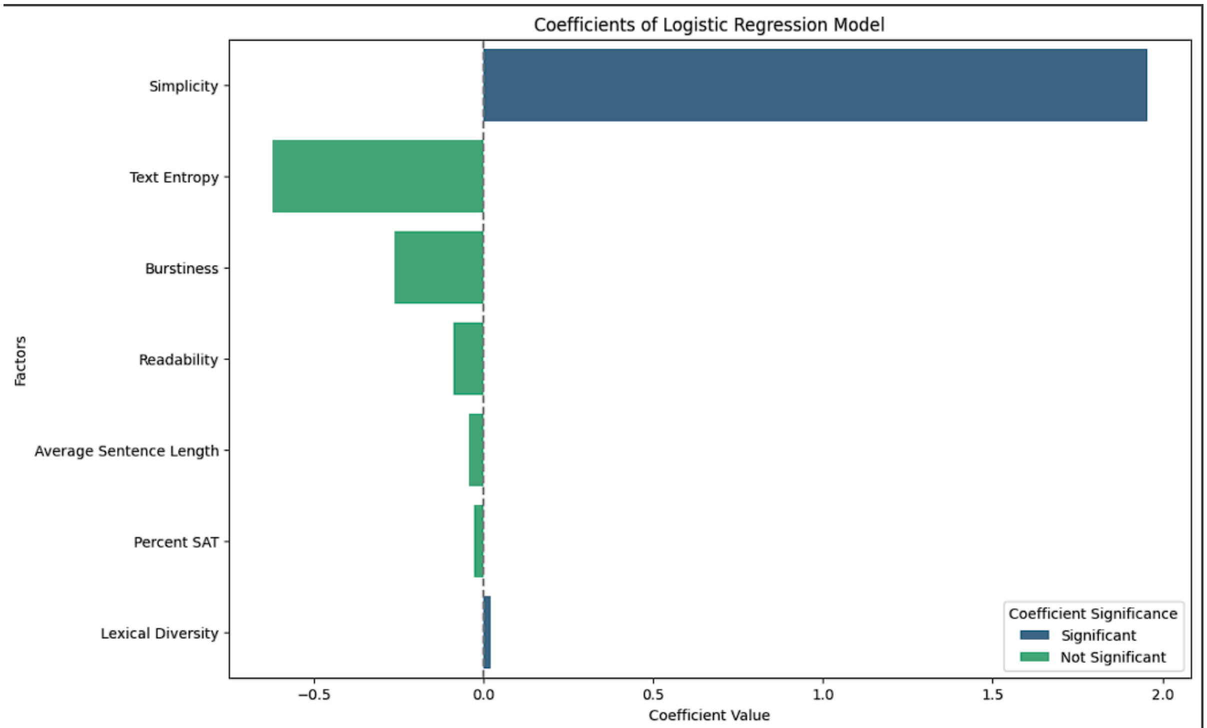
model, including precision, recall, f1-score for each class, and the overall accuracy of the model.
    -
    - Precision/Recall average (F1-score): 0.88 for 0, 0.90 for 1
        - 88% of the instances predicted as Human-written were correctly predicted, and 90% of the instances predicted as AI-written were correctly predicted.
    - Accuracy: 0.89
        - The proportion of total correct predictions over all predictions made. Correctly predicted 89% of total instances
    - This model exhibits a relatively high accuracy, precision, and recall level for both classes, however, it performs slightly better for identifying human-written text as opposed to AI-written.
- Next, we created a function to evaluate the stability and performance of our pipeline across multiple iterations of training and testing (called run_iterations_and_plot_accuracies_with_pipeline).
    - We ran this function 100 times, where each iteration conducts a train-test split on the data, trains the pipeline on the training set, and then makes predictions on the test set.
    - Once all iterations have been completed, we plotted a histogram to show the distribution of accuracy scores across the iterations. We also plotted a KDE to show the probability density of the accuracies, as well as the mean accuracy.



-
  - The most prevalent part of this graph is the large cluster of data near the mean. This highlights that our model has relatively consistent performance across different splits. The distribution is, however, slightly skewed to the left, showing that more iterations have accuracies that score less than the mean as opposed to greater. Otherwise, the model performs decently well, but variation is certainly present.
- After that, we used logistic regression to discern essay authors
  - Like before, we did a train test split on the data, with 20% of the data reserved for testing.
  - We then used a logistic regression classifier with a max number of iterations set to 1000.

	precision	recall	f1-score	support
0	0.88	0.83	0.86	18
1	0.85	0.89	0.87	19
accuracy			0.86	37
macro avg	0.87	0.86	0.86	37
weighted avg	0.87	0.86	0.86	37

  The classifier was then trained on the training data.
    - These metrics suggest that this model could be less accurate than our previous model, however, this is only a prediction
  - After the training, the coefficients and intercept are extracted and placed into a new data frame, which will include the linguistic metrics, as well as their respective coefficients, and a column called “Significance” which evaluates whether the metric is significant based on if their coefficient is greater than or equal to zero.
  - We then plotted a bar plot to visualize the logistic regression model and see which metrics end up being significant.

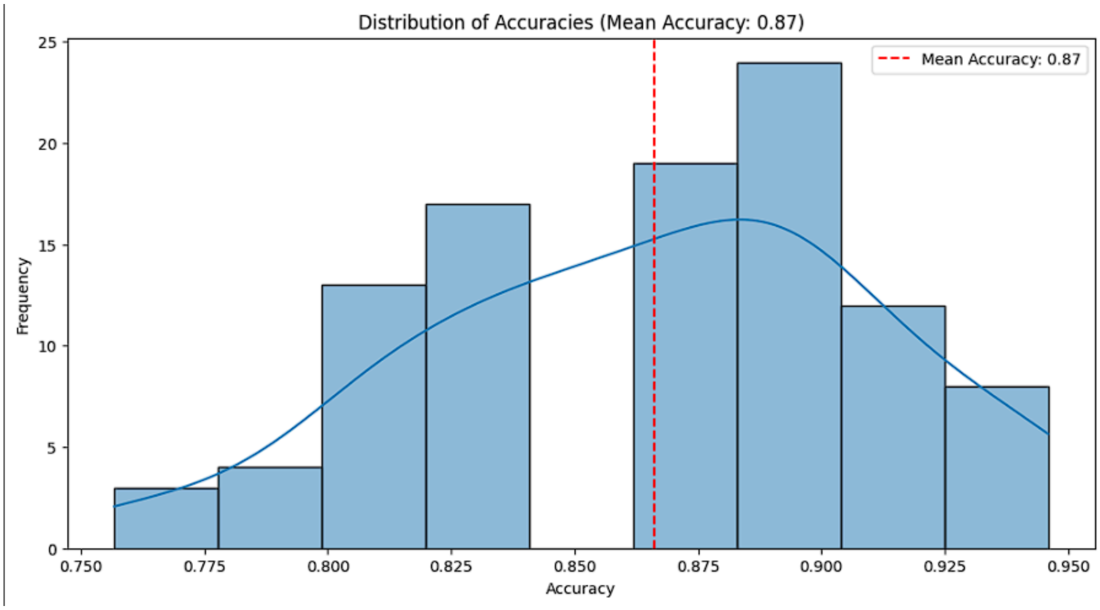


	Factors	Coefficients	Significance
2	Simplicity	1.955072	Significant
6	Text Entropy	-0.620899	Not Significant
4	Burstiness	-0.260885	Not Significant
0	Readability	-0.089580	Not Significant
5	Average Sentence Length	-0.043609	Not Significant
1	Percent SAT	-0.028109	Not Significant
3	Lexical Diversity	0.021498	Significant

-
    - This table suggests that, according to our model, both simplicity and lexical diversity are characteristics of AI-written material, while text entropy, burstiness, readability, average sentence length, and percent SAT are indicative of human-written material. Furthermore, simplicity seems to be very strongly correlated with AI-written material.
- We then filtered the training and testing data to only include the significant metrics and ran our initial model.

	precision	recall	f1-score	support
0	0.84	0.89	0.86	18
1	0.89	0.84	0.86	19
accuracy			0.86	37
macro avg	0.87	0.87	0.86	37
weighted avg	0.87	0.86	0.86	37

    - **Here are the predictions it calculated**:

- This suggests that it will be less accurate than the first model, and although the specific precisions and recalls differ, the averages stay relatively the same as the logistic regression model with all metrics included.
- Here is the accuracy plot this model outputted:
- 

    - This model achieves a decent level of accuracy on average, and the distribution shows that the lowest accuracy achieved was above 0.75, which suggests that our model, for the most part, tends to be accurate. However, such variation in performance makes it so the model is not completely usable if you want accuracy on every test, which someone using an AI detector would likely want.

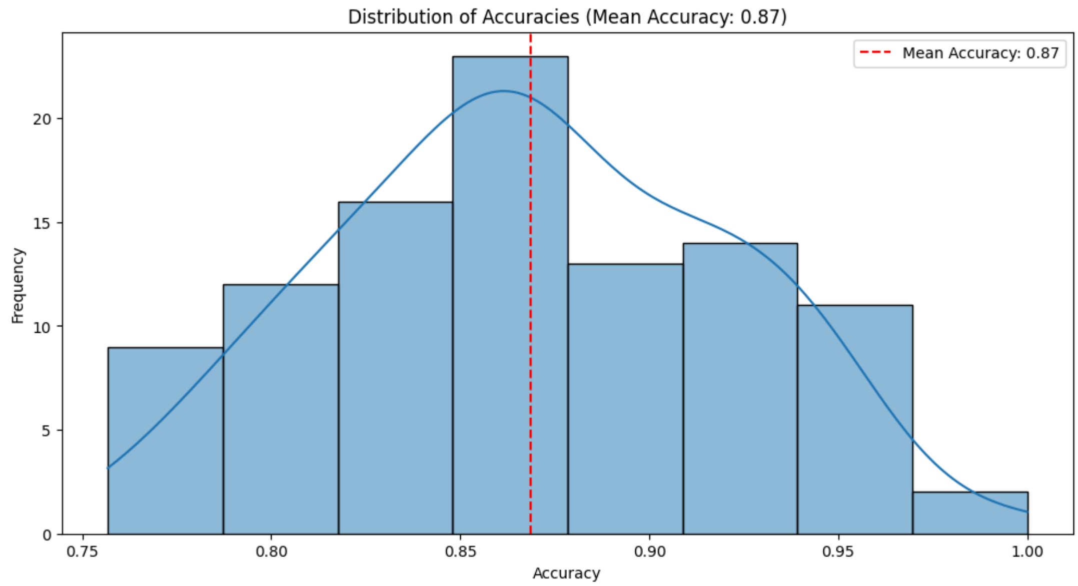
- We then performed Chi-squared and P-tests to extract whether a metric was statistically significant when determining if a text was AI-written.
    -

	Metric	Chi-Square	P-Value	Significant
0	Readability_Cat	58.274964	2.216932e-13	True
1	Percent SAT_Cat	5.420927	6.650598e-02	False
2	Simplicity_Cat	61.160296	5.238542e-14	True
3	Lexical Diversity_Cat	32.044673	1.100494e-07	True
4	Burstiness_Cat	32.044673	1.100494e-07	True
5	Average Sentence Length_Cat	6.076684	4.791426e-02	True
6	Text Entropy_Cat	64.570234	9.522518e-15	True

- These chi-squared statistics and significance levels provide evidence that all metrics except for # of SAT words could aid in predicting the authorship of a piece of written material.

- We then ran the original model again, now with our new list of significant columns, and here are the predictions and graph:

	precision	recall	f1-score	support
0	1.00	0.76	0.86	21
1	0.76	1.00	0.86	16
accuracy			0.86	37
macro avg	0.88	0.88	0.86	37
weighted avg	0.90	0.86	0.86	37



- This data suggests similar attributes regarding our model, where our average accuracy is relatively high, and our distribution spans only higher-accuracy percentages, but now with some even reaching 100% accuracy.

Conclusion:

In our investigation into the feasibility of training a model to detect AI-written text, we have applied various linguistic metrics to a diverse, but limited dataset of essays. Through rigorous statistical tests and machine learning models, including logistic regression and RandomForestClassifier, we have identified key features that differ between human and AI writing. Our models have demonstrated commendable accuracy, with the ability to predict AI authorship in essays semi-reliably. However, variability in accuracy indicates a need for further refinement. One who intends to use a tool like this expects the output to be very highly accurate since serious punishment could be on the line if a text were to come back as AI-generated. With this much variability, it would not be fair to use our tool as a definitive judge. With that being said, the success of our models lays a promising foundation for developing robust AI-written text detection tools, signaling a significant step forward in distinguishing AI-generated content in academic and professional domains.

Testing AI Detector Metrics, What are they testing for?

We used OLS statistical significance test from the library, statsmodel.api, in order to test significant features that were taken into account for each AI detector. There were several targets we wanted to test for. First, we cleaned up the data to only include numeric metrics to prepare for our statistical analysis. we also removed all of the results from AI detectors because that is technically our target. Then we ran the hypothesis test like this for each of the AI detectors.

GPTZero:

OLS Regression Results

Dep. Variable:	GPTZero Score	R-squared:	0.576
Model:	OLS	Adj. R-squared:	0.549
Method:	Least Squares	F-statistic:	21.12
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	1.30e-26
Time:	02:47:41	Log-Likelihood:	-844.19
No. Observations:	183	AIC:	1712.
Df Residuals:	171	BIC:	1751.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	257.9935	60.109	4.292	0.000	139.343	376.644
Readability	-0.9733	0.153	-6.380	0.000	-1.274	-0.672
Percent SAT	-1.1144	0.460	-2.421	0.017	-2.023	-0.206
Simplicity	82.1343	17.553	4.679	0.000	47.486	116.783
Lexical Diversity	-0.0684	0.324	-0.211	0.833	-0.708	0.571
Burstiness	-2.9231	0.859	-3.403	0.001	-4.619	-1.227
Average Sentence Length	-1.1578	0.580	-1.997	0.047	-2.302	-0.014
Text Entropy	-7.0978	4.832	-1.469	0.144	-16.635	2.440
English	11.6857	18.017	0.649	0.517	-23.879	47.250
History	5.2232	17.953	0.291	0.771	-30.215	40.662
Other	3.7350	17.977	0.208	0.836	-31.751	39.221
Technology	14.4208	18.753	0.769	0.443	-22.596	51.438

Omnibus:	13.911	Durbin-Watson:	1.897
Prob(Omnibus):	0.001	Jarque-Bera (JB):	5.353
Skew:	-0.081	Prob(JB):	0.0688
Kurtosis:	2.178	Cond. No.	3.25e+03

The OLS regression results for the GPTZero Score as the dependent variable reveal the following:

- We had an R-squared value of 0.576, which signifies that 57.6% of the variability in GPTZero's scoring system is explained by the variables used in the model.
- Of the several metrics being tested, five of them significantly affected the overall score outputted by GPTZero:
    - **Readability**: Shows a significant negative impact, indicating that higher readability scores are associated with lower GPTZero Scores (Human-written).
    - **Percent SAT**: Demonstrates a significant negative impact, suggesting that essays with a higher proportion of SAT-level words tend to have lower GPTZero Scores (Human-written).

- ○ **Simplicity**: Exhibits a significant positive impact, suggesting that simpler texts are more likely to have higher GPTZero Scores (AI-written).
- ○ **Burstiness**: Has a significant negative impact, indicating that essays with varied sentence lengths tend to have lower GPTZero Scores (Human-written).
- ○ **Average Sentence Length**: Also shows a significant negative impact indicating that essays with lower average sentence lengths tend to have lower GPTZero Scores (Human-written).

- **<u>Non-Significant Variables</u>**:
  - ○ Lexical Diversity,Text Entropy, and Subject Categories (English, History, Other, Technology) did not show a significant impact on the GPTZero Score.

Takeaways:

The analysis suggests that the GPTZero Score is influenced by several linguistic features of the essays. Higher readability, more complex vocabulary (Percent SAT), and varied sentence structures (Burstiness, Average Sentence Length) tend to be associated with lower scores, while simplicity in text correlates with higher scores. This could reflect characteristics more commonly found in human-written essays or biases in the GPTZero detector.

ContentDetectorAI:

OLS Regression Results

Dep. Variable:	ContentDetectorAI Score	**R-squared:**	0.276
Model:	OLS	**Adj. R-squared:**	0.230
Method:	Least Squares	**F-statistic:**	5.935
Date:	Fri, 08 Dec 2023	**Prob (F-statistic):**	3.78e-08
Time:	22:59:39	**Log-Likelihood:**	-797.41
No. Observations:	183	**AIC:**	1619.
Df Residuals:	171	**BIC:**	1657.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	127.3683	46.551	2.736	0.007	35.480	219.256
Readability	-0.4409	0.118	-3.731	0.000	-0.674	-0.208
Percent SAT	0.6871	0.356	1.927	0.056	-0.017	1.391
Simplicity	37.8232	13.594	2.782	0.006	10.990	64.657
Lexical Diversity	-0.6729	0.251	-2.681	0.008	-1.168	-0.177
Burstiness	-1.0922	0.665	-1.642	0.102	-2.405	0.221
Average Sentence Length	-0.8760	0.449	-1.951	0.053	-1.762	0.010
Text Entropy	-5.0784	3.742	-1.357	0.177	-12.465	2.308
English	-26.1078	13.953	-1.871	0.063	-53.651	1.435
History	-25.1774	13.904	-1.811	0.072	-52.622	2.268
Other	-20.2471	13.922	-1.454	0.148	-47.729	7.235
Technology	-24.8271	14.523	-1.709	0.089	-53.495	3.840

Omnibus:	4.786	**Durbin-Watson:**	1.696
Prob(Omnibus):	0.091	**Jarque-Bera (JB):**	4.431
Skew:	0.317	**Prob(JB):**	0.109
Kurtosis:	2.576	**Cond. No.**	3.25e+03

The OLS regression results for the ContentDetectorAI Score as the dependent variable reveal the following:

- We had an R-squared value of 0.276, which signifies that 27.6% of the variability in ContentDetectorAI's scoring system is explained by the variables used in the model.
- Of the several metrics being tested, three of them significantly affected the overall score outputted by ContentDetectorAI:
    - **Readability**: Demonstrates a significant negative impact, suggesting that higher readability scores are associated with lower ContentDetectorAI Scores (Human-written).
    - **Simplicity**: Shows a significant positive impact, indicating that simpler texts are more likely to have higher ContentDetectorAI Scores (AI-written).
    - **Lexical Diversity**: Exhibits a significant negative impact, suggesting that a greater variety of vocabulary is associated with lower ContentDetectorAI Scores (Human-written).
- Non-Significant Variables:
    - Variables like Percent SAT, Burstiness, Average Sentence Length, Text Entropy, and Subject Categories (English, History, Other, Technology) did not show a significant impact on the ContentDetectorAI Score.

Takeaways:

The results suggest that the ContentDetectorAI Score is also influenced by several linguistic features of the essays. Higher complexity in vocabulary (Lexical Diversity) and higher readability tend to be associated with lower scores, while simplicity in text correlates with higher scores. The relationship with Percent SAT and sentence structure (Burstiness, Average Sentence Length) is less clear.

GPT2:

OLS Regression Results

Dep. Variable:	GPT2 Score	**R-squared:**	0.204
Model:	OLS	**Adj. R-squared:**	0.153
Method:	Least Squares	**F-statistic:**	3.990
Date:	Fri, 08 Dec 2023	**Prob (F-statistic):**	3.48e-05
Time:	22:59:39	**Log-Likelihood:**	-867.88
No. Observations:	183	**AIC:**	1760.
Df Residuals:	171	**BIC:**	1798.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	137.7868	68.417	2.014	0.046	2.736	272.838
Readability	0.2837	0.174	1.634	0.104	-0.059	0.627
Percent SAT	-1.0724	0.524	-2.047	0.042	-2.107	-0.038
Simplicity	107.1825	19.979	5.365	0.000	67.744	146.621
Lexical Diversity	-0.6795	0.369	-1.842	0.067	-1.408	0.049
Burstiness	1.6044	0.978	1.641	0.103	-0.326	3.534
Average Sentence Length	-0.7787	0.660	-1.180	0.240	-2.081	0.524
Text Entropy	-6.7418	5.500	-1.226	0.222	-17.598	4.114
English	-3.7599	20.508	-0.183	0.855	-44.240	36.721
History	-19.5649	20.435	-0.957	0.340	-59.902	20.772
Other	-19.1040	20.462	-0.934	0.352	-59.495	21.287
Technology	-17.2244	21.345	-0.807	0.421	-59.358	24.909

Omnibus:	65.213	**Durbin-Watson:**	1.548
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	140.302
Skew:	1.668	**Prob(JB):**	3.42e-31
Kurtosis:	5.696	**Cond. No.**	3.25e+03

The OLS regression results for the GPT2 Score as the dependent variable show the following:

- We had an R-squared value of 0.204, which signifies that 20.4% of the variability in GPT2's scoring system is explained by the variables used in the model.

- Of the several metrics being tested, three of them significantly affected the overall score outputted by GPT2:
    - **Percent SAT**: Shows a significant negative impact, suggesting that essays with a higher proportion of SAT-level words tend to have lower GPT2 Scores (Human-written).
    - **Simplicity:** Demonstrates a significant positive impact, indicating that simpler texts are more likely to have higher GPT2 Scores (AI-written).
    - **Lexical Diversity**: Exhibits a near-significant negative impact, suggesting that a greater variety of vocabulary is associated with lower GPT2 Scores (Human-written).
- Non-Significant Variables:
    - Variables like Readability, Burstiness, Average Sentence Length, Text Entropy, and Subject Categories (English, History, Other, Technology) did not show a significant impact on the GPT2 Score.

Takeaways:
The results suggest that the GPT2 Score is also influenced by several inguistic features of the essays. Higher complexity in vocabulary (Percent SAT) tends to be associated with lower scores, while simplicity in text correlates with higher scores. Lexical diversity also appears to have an impact, although not as strongly as simplicity.

AcademicHelp:

OLS Regression Results

Dep. Variable:	AcademicHelp Score	**R-squared:**	0.337
Model:	OLS	**Adj. R-squared:**	0.294
Method:	Least Squares	**F-statistic:**	7.887
Date:	Fri, 08 Dec 2023	**Prob (F-statistic):**	5.21e-11
Time:	22:59:39	**Log-Likelihood:**	-863.91
No. Observations:	183	**AIC:**	1752.
Df Residuals:	171	**BIC:**	1790.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	200.5161	66.950	2.995	0.003	68.361	332.672
Readability	-0.6400	0.170	-3.766	0.000	-0.975	-0.305
Percent SAT	0.5496	0.513	1.072	0.285	-0.462	1.562
Simplicity	65.1848	19.551	3.334	0.001	26.592	103.777
Lexical Diversity	-1.3651	0.361	-3.782	0.000	-2.078	-0.653
Burstiness	-2.1718	0.957	-2.270	0.024	-4.061	-0.283
Average Sentence Length	-1.2280	0.646	-1.902	0.059	-2.503	0.047
Text Entropy	-14.9241	5.382	-2.773	0.006	-25.547	-4.301
English	14.5610	20.068	0.726	0.469	-25.052	54.174
History	18.7199	19.997	0.936	0.351	-20.752	58.192
Other	19.2580	20.024	0.962	0.338	-20.267	58.783
Technology	15.0095	20.887	0.719	0.473	-26.221	56.240

Omnibus:	5.929	**Durbin-Watson:**	1.714
Prob(Omnibus):	0.052	**Jarque-Bera (JB):**	5.710
Skew:	0.383	**Prob(JB):**	0.0575
Kurtosis:	2.597	**Cond. No.**	3.25e+03

The OLS regression results for the AcademicHelp Score as the dependent variable show the following:

- We had an R-squared value of 0.337, which signifies that 33.7% of the variability in AcademicHelp's scoring system is explained by the variables used in the model.
- Of the several metrics being tested, five of them significantly affected the overall score outputted by AcademicHelp:
    - **Readability:** Shows a significant negative impact, suggesting that higher readability scores are associated with lower AcademicHelp Scores (Human-written).
    - **Simplicity:** Demonstrates a significant positive impact, indicating that simpler texts are more likely to have higher AcademicHelp Scores (AI-Written).
    - **Lexical Diversity:** Exhibits a significant negative impact, suggesting that a greater variety of vocabulary is associated with lower AcademicHelp Scores (Human-written).
    - **Burstiness:** Shows a significant negative impact, indicating that essays with varied sentence lengths tend to have lower AcademicHelp Scores (Human-written).
    - **Text Entropy:** Also shows a significant negative impact indicating that essays with less unpredictability tend to have lower GPTZero Scores (Human-written).
- Non-Significant Variables:
    - Percent SAT, Average Sentence Length, and Subject Categories (English, History, Other, Technology) did not show a significant impact on the AcademicHelp Score.

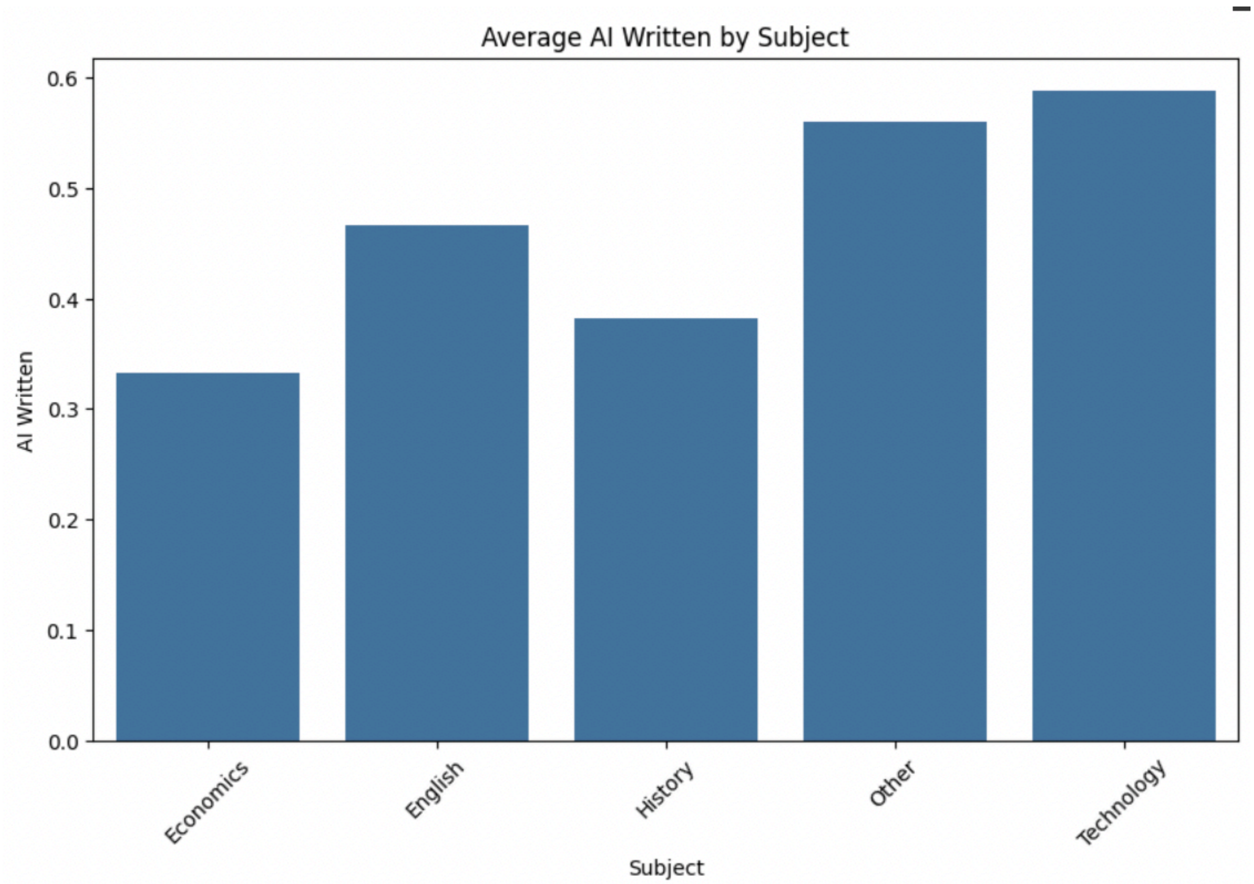
Takeaways:
The results suggest that several linguistic features of the essays influence the AcademicHelp Score. Higher readability, greater lexical diversity, and varied sentence lengths (Burstiness) tend to be associated with lower scores, while simplicity in text correlates with higher scores. This could reflect preferences or standards used by AcademicHelp in evaluating essays.

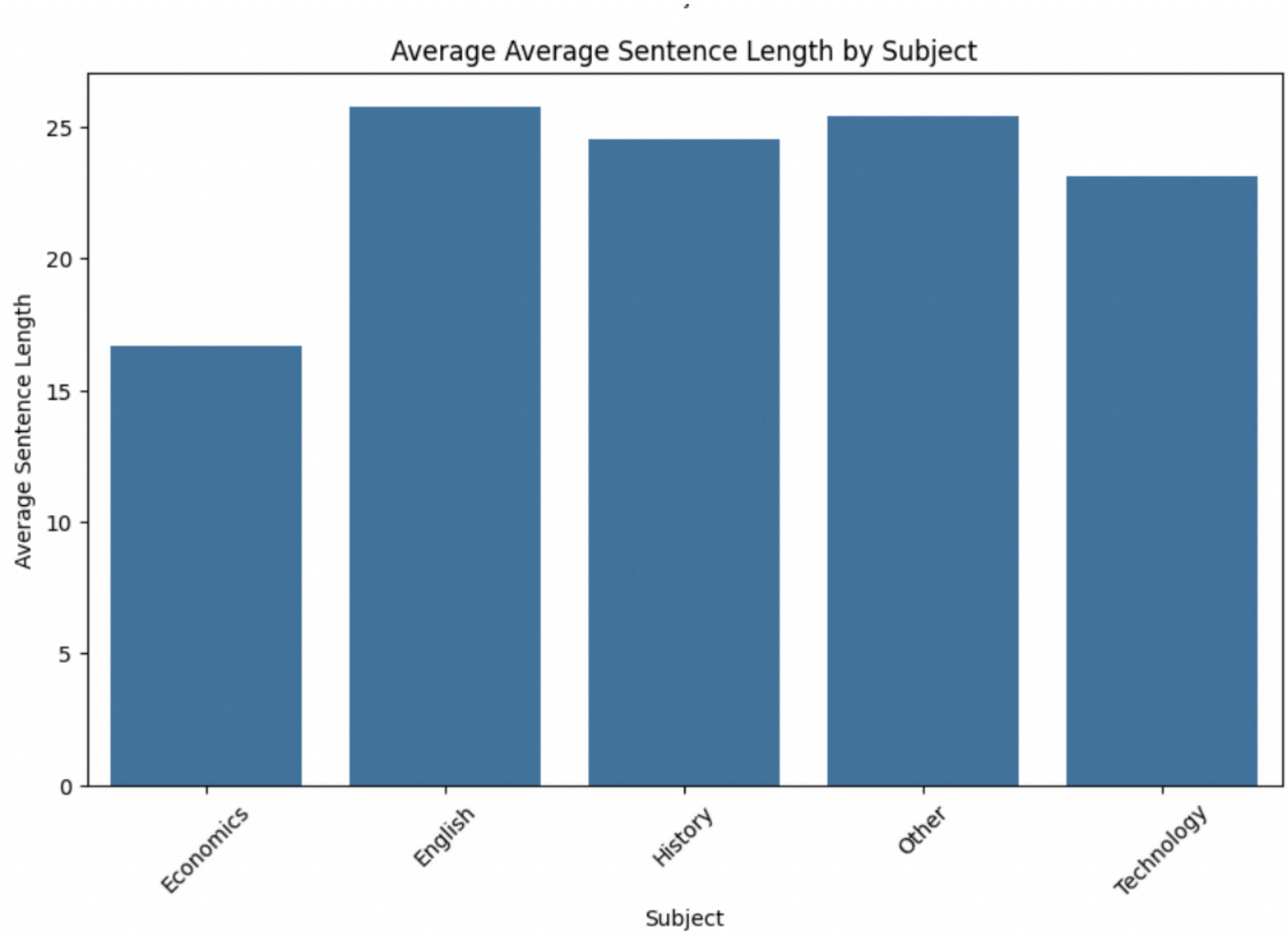
Conclusion:
After rigorous analysis across four different AI detectors, we have come to the conclusion that certain linguistic features, such as simplicity and lexical diversity consistently influence scores outputted by these detectors. This signifies that these features are pivotal for these detectors to distinguish AI-written text from human-written text. Readability and text structure were also seen to play roles within these calculations, however, their significance varied between detectors. We also found that across the board, the percentage of SAT words, as well as average sentence length, were not significant in identifying AI authorship. For two of the models (ContentDetectorAI and GPT2), text entropy and burstiness were also insignificant. This signifies that there features tend to be not very predictive of AI-written text.

Different Subjects and Their Impact on AI Detectors

Essays revolving around technology display remarkable characteristics that notably impact their detection rates by AI systems. A key feature is the deliberate use of technical terminology intricately woven into a structured, formal writing style. This distinct approach is marked by the employment of specialized terms and concepts inherent to the realm of technology. The deliberate structuring and adherence to formal language norms create a predictable pattern recognizable by AI detectors. This consistency not only aids in the identification of familiar structural elements but also streamlines the process of categorization within the technology domain. Moreover, the uniform and standardized utilization of vocabulary within this field significantly assists AI models in pinpointing and precisely categorizing these essays. The reliance on established terminologies and consistent vocabulary choices allows for more accurate recognition of the subject matter, contributing to the heightened likelihood of detection. On the other hand, English and History essays tend to have more diverse writing styles ranging from descriptive to analytical. There are a lot of diversities in the way ideas are expressed, making it difficult for AI detectors to really detect if an English or History essay has been actually written by AI or not. Moreover, English and History essays have wide ranges of vocabulary and use words that are more complex and unique, as opposed to words that are used more for casual conversation. Another major thing about English and History essays is that the sentences tend to be longer as most students or anyone else writing on those topics tend to be comparing many things or analyzing stuff. The average sentence length for English essays is about 25 while it's less for other subjects. This tends to happen due to the topics being something that you have to defend such as defending a thesis, and tend to try to be more connected. They want to make the sentences feel more connected, which allows the paper to be more fluid in terms of its writing style, and humans tend to want to connect the sentences when they write while AI will keep its sentences more precise while using higher level words. This makes it very difficult for AI detectors to really identify an AI-made essay or not. As seen in the chart below, one can clearly see that the percentage of Technology-related essays that get flagged is higher than any of the other subjects.

The dichotomy between technology-related essays, with their structured language and consistent technical terminology facilitating detection, and the fluid and diverse nature of writing styles and vocabulary in English and History essays, highlights the nuanced intricacies that govern AI detection across different subject matters. This divergence underscores the need for AI models to adapt to the diverse characteristics inherent in various subjects, aiming for a more comprehensive understanding and recognition of distinct writing styles and content nuances across disciplines. Efforts toward enhancing AI detection accuracy should prioritize accommodating these divergent attributes to create more robust and adaptable detection systems.




Average Average Sentence Length by Subject
Average Sentence Length
0
5
10
15
20
25
Economics
English
History
Other
Technology
Subject