machine_learning_project

Academic Honesty in the Age of Generative AI

Introduction

Education plays an essential role in quality of life and social mobility. In the U.S., tens of millions of students go to school everyday to learn new skills to prepare for their future career. While academic honesty lies at the heart of an effective education, the rise of generative AI tools such as ChatGPT allows students to produce an essay in seconds with the click of a few buttons.

Literature Review

- Researchers have created various data sets and used many traditional machine learning techniques to detect the use of AI in essays [1].
- Many previous studies have focused on detecting stylistic differences between student and AI written essays, but this has become less feasible as more and more models are released to the public [3].
- Researchers have thoroughly investigated many different machine learning techniques such SVMs, Naive Bayes, and K nearest neighbors. However, a BERT transformer has been found to be extremely effective [2].
- Other research papers have found success using a N-gram bag-of-words discrepancy language model [4].

Dataset

The dataset linked below contains 40,000 essay samples either written by humans or generated using AI. They are labeled with either AI or Human in binary encoding (1, 0). The data contains a few other features including the prompt used to generate the essay, but we chose to only use features extracted from each essay in our preprocessing.

https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data

Problem

Artificial Intelligence is being used more and more frequently for text generation, especially in academic settings.

Motivation

If teachers are able to leverage accurate models to predict whether an essay was written by a student or AI, teachers can hold students accountable for their learning and improve the education system as a whole.

Methods

Data Preprocessing

In our preprocessing, we chose a few features that made sense to help distinguish between AI written text and Human Written Text. We first extracted the necessary portions of the data (the actual text and the label), as the dataset came with a few extra features that were not relevant to our project. The features we chose to extract for our first few tests of the model are listed below:

- **Number of Apostrophes:** The number of apostrophes in a chunk of text is a relatively good estimate for the number of contractions. Contractions tend to be used in more casual writing, and based on observation, AI generated text takes a more formal tone.
- **Variance in Length of Paragraphs:** Based on observation, AI generated text has a very consistent paragraph length, while human generated text is more likely to have higher variance.
- **Shannon Entropy:** Since AI generated text is generated according to a probability distribution, it makes sense that there might be a pattern in the entropy of a given text. This pattern could be picked up on by a machine learning model.
- **Sentiment Analysis:** This is a measure of how monotone a piece of text is. We will create this feature using several NLP libraries from NLTK. We expect AI generated texts to have a much larger monotone values.
- **Lexical Diversity:** This is a measure of the lexical diversity or a rating of how many different words are found in the text. We will create this feature using several NLP libraries from NLTK. We expect AI generated text to have lower lexical diversity than humans.
- **PoS:** PoS stands for Parts of Speech. This simply looks at the percent of nouns per number of words in the text. We read that AI generated texts often use lots of nouns, so we are hoping this will be a solid feature for detecting AI text.
- **Readability Metrics:** The Gunning Fog Index and the Flesch Reading Ease Score are features used to measure the readability of a given text. The Gunning Fog Index determines the years of education needed to understand a piece of writing based on average sentence length and the use of complex words. In contrast, the Flesch Reading Ease Score calculates readability

using total words, syllables, and sentences, yielding a score between 0 and 100, where higher scores translate to easier comprehension. These metrics can help distinguish between human and AI-generated speech since AI tends to produce more uniform text with predictable scores, while human writing tends to be more varied and nuanced.

ML Algorithms/Models:

- **GNB**: We started with a Gaussian Naive Bayes Model. This allowed us to classify our data quickly, meaning we had more time to try tweaking some of the features that we applied in preprocessing. Since it's a supervised method, we are also able to train the model on some of our data and test it on the remaining points. We calculated the F1 and accuracy scores shortly after running the model.
- **SVM**: We selected the Support Vector Machine algorithm for linearly separable data. Also a supervised method, we chose this because it's less prone to overfitting and also more interpretable than other methods because its vectors can shed light on which data points are most significant.
- **NN**: We included a supervised Neural Network because it excels with finding complex relationships in large datasets because it is trained by tuning numerous parameters.
- Code for this portion of the project can be found in the main branch of our github repository.

Project Goals:

Our goal is to help teachers detect the use of AI assistants in written essays. We will exercise extreme caution in order to avoid false-positives that could devastate a student's career. In terms of sustainability, training machine learning models such as neural networks can be extremely energy intensive and leave a substantial environmental footprint. We aim to meet the precision goals specified below.

Expected Results:

We expect the model would meet the following criteria:

- F1 Score: 70%+
- Fowlkes-Mallows Index: 0.7+
- Precision: 70%+
- Recall: 75%+

Actual Results

Gaussian Naive Bayes:

- Accuracy: 0.7064
- F1 Score: 0.7026
- Precision: 0.5445
- Recall: 0.9453
- Fowlkes-Mallow Index: 0.6116

Support Vector Machine:

- Accuracy: 0.8206
- F1 Score: 0.8249
- Precision: 0.7888
- Recall: 0.6602
- Fowlkes-Mallow Index: 0.7409

Neural Network:

- Accuracy: 0.8713
- F1 Score: 0.8718
- Precision: 0.8277
- Recall: 0.8030
- Fowlkes-Mallow Index: 0.7956

Analysis of Results

GNB:

- Our recall score is very high, meaning that we are very likely to classify an AI written text as AI written. However, the precision is pretty low, indicating that only about 55.5 percent of text classified as AI was actually AI. There was a very high false positive rate.
- Accuracy was about 71%, meaning that the model correctly predicted the classification of an essay 71% of the time. This was higher than our original goal, so now we hope to improve this score in future iterations and models.
- F1 score and Fowlkes Mallow Index both measure the balance of precision and recall. With F1 measuring the harmonic mean, it shows that the precision and recall are fairly balanced. However, Fowlkes Mallow is more influenced by very high or low values, so it is a lower score (most likely due to the low precision score).

SVM:

- Precision and recall indicated that about 78 percent of text classified as AI is actually AI, and that about 66 percent of AI generated text was actually classified as AI. The false positive rate was within the goal range, but the model missed a lot of AI generated text and classified it as human instead.
- The accuracy score indicates that the SVM Model correctly classified about 82 percent of datapoints.
- The high F1 score shows that precision and recall are well balanced, but Folwkes Mallow indicated that the balance could still improve. Since it is highly sensitive to low precision or recall values, this score shows that we should focus on improving the recall in the next model.

NN:

- The accuracy score means that 87 percent of data was correctly classified. This is well above the goal for accuracy performance.
- Precision and Recall indicate that 82 percent of AI classfied texts were actually AI generated, and that the model correctly classified around 80 percent of AI texts.
- F1 and Fowlkes Mallow show that the precision and recall balance is high.
- To validate the performance of the Neural Network, we used k-folds validation and took the averages of each metric. (All visualizations show the run of all 5 folds concatenated together)

Comparison of Models:

Criteria: Due to our model potentially being used to catch cheaters, we have given serious consideration to whether we prefer False Positives or False Negatives. We intend for this to simply be a tool that educators use to reduce the sample size of essays they should thoroughly inspect. However, AI and machine learning tools are often treated like blackboxes and their results are often given far more weight then they should have. As a result, we would prefer more False Negatives than False Positives in order to avoid falsely accusing and punishing students who did not cheap.

- The Neural Network Model performed higher than both other models in all metrics except GNB's recall. Recall performance from the GNB can be attributed the the fact that the model was much more likely to classify ANY text as AI generated, so it was more likely to catch the ones that were actually AI. This led to much worse precision. This can be seen both from the metrics and from the GNB Precision-Recall Curve that is not very close to (1, 1).
- Based on the accuracy metrics, it is clear that the Neural Network has the lowest amount of incorrectly classified datapoints. This is important in practical use because any amount of inaccuracy can have large consequences. However we also wanted to analyze the split between False Positives and False negatives. GNB has a lot of False Positives that when

coupled with its low accuracy mean this model is not very viable. SVMs accuracy is significantly lower than Neural Network, but it does lean towards more False Negatives. Compared to the Neural Network which has a fairly even split between False Negatives and False Positives, it might at first seem like the SVM might fit our use case slightly better. However, the SVM has nearly twice the amount of False Positives as the average neural network run. As a result, we prefer the Neural network when analyzing accuracy and incorrect classifications.

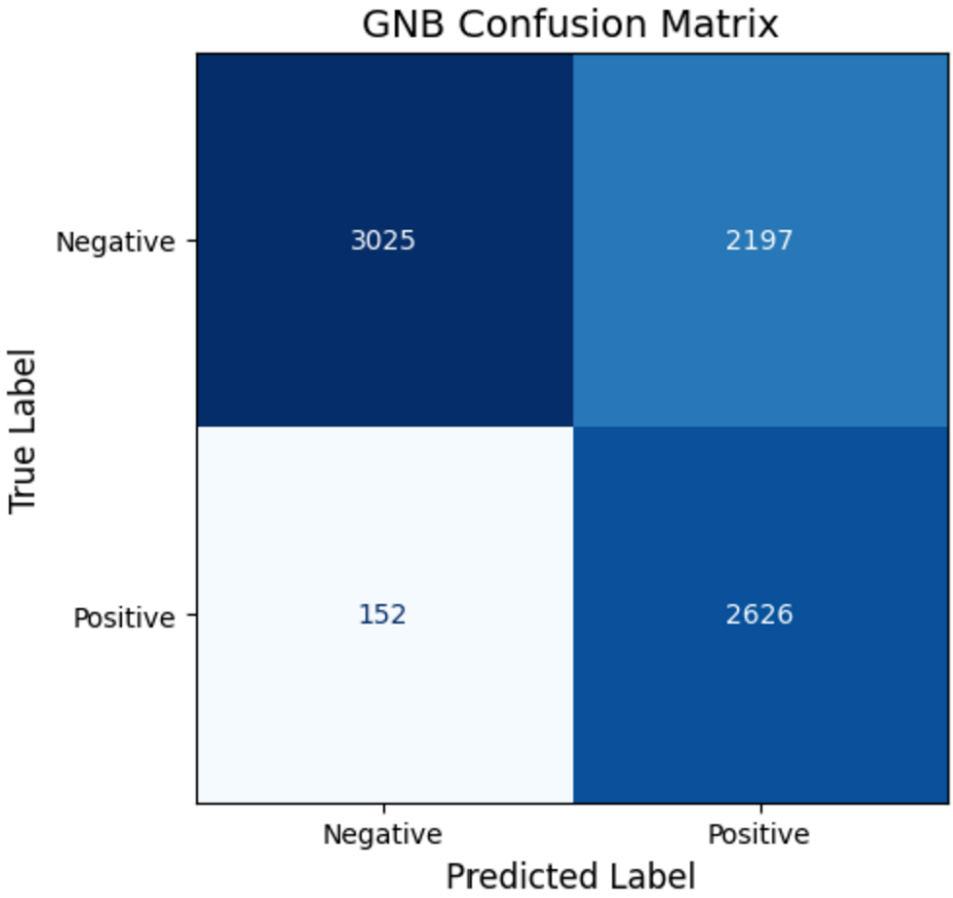
- Even though our dataset is slightly unbalanced, we wanted to analyze an ROC curve to see how well our model was able to classify data points. When analyzing a ROC curve, we specifically look at the area under the curve. The area under the curve represents the chance given a positive and negative datapoint that the positive is ranked above the negative point. The Neural Network was still found to the best model with an AUC of 0.94. GNB and SVM both had AUCs that were only slightly smaller, but because of their lower accuracies we continue to prefer the neural network as our final model.
- Since the slight imbalance in our dataset could result in the AUCs being overly optimistic, we decided to include a Precision-Recall Curve. Looking at the Precision-Recall Curves, we see that the Neural Network curve is the closest to the ideal shape. On top of this, its precision and recall values are by far the most closest of any of the models. This means our model is actually distinguishing between the AI and Human text instead of just guessing a certain label in order to maintain the highest accuracy. This balance is further shown by the Neural Network having the highest F1 and Fowlkes Mallow scores.
- Its important to note that each model had it's own feature importance graph. Each model was tuned to only use the features that provided the best performance. GNB performed best with the fewest features, while SVM and the neural network saw significant performance hits when we removed any more features even if they were deemed not very important. We believe we tuned each model to the best of its abilities with the available features.

As a result, we present our Neural Network as our final model.

Visualizations of Results:

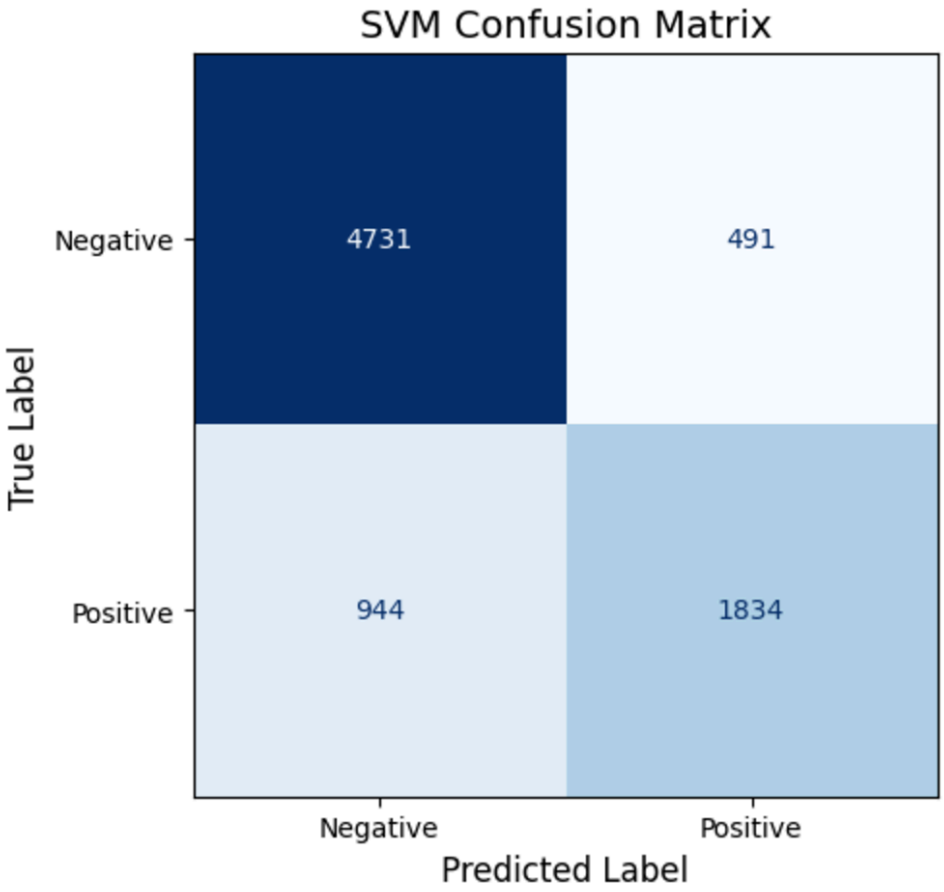
Confusion Matrices

GNB



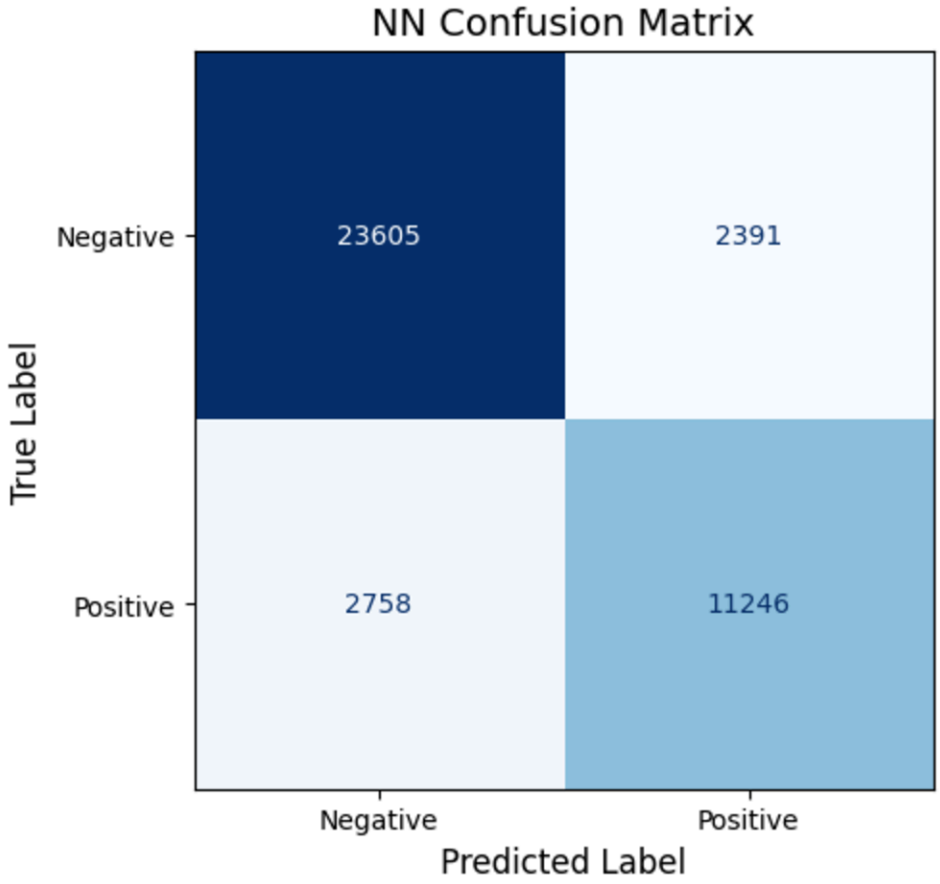
Base don the confusion matrix, GNB tends to guess AI more than it should.

SVM



Based on the confusion matrix, SVM tends to guess human more than it should.

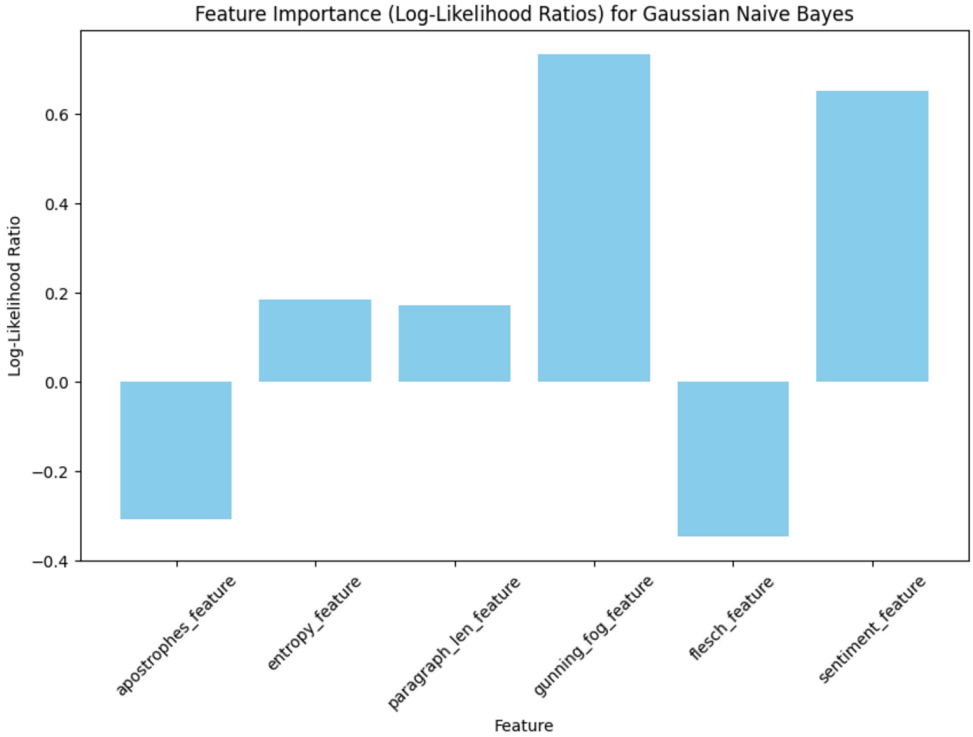
NN



This shows the result for the neural network running over 5 folds. This means it has 5 times the amount of data than the other confusion matrices.

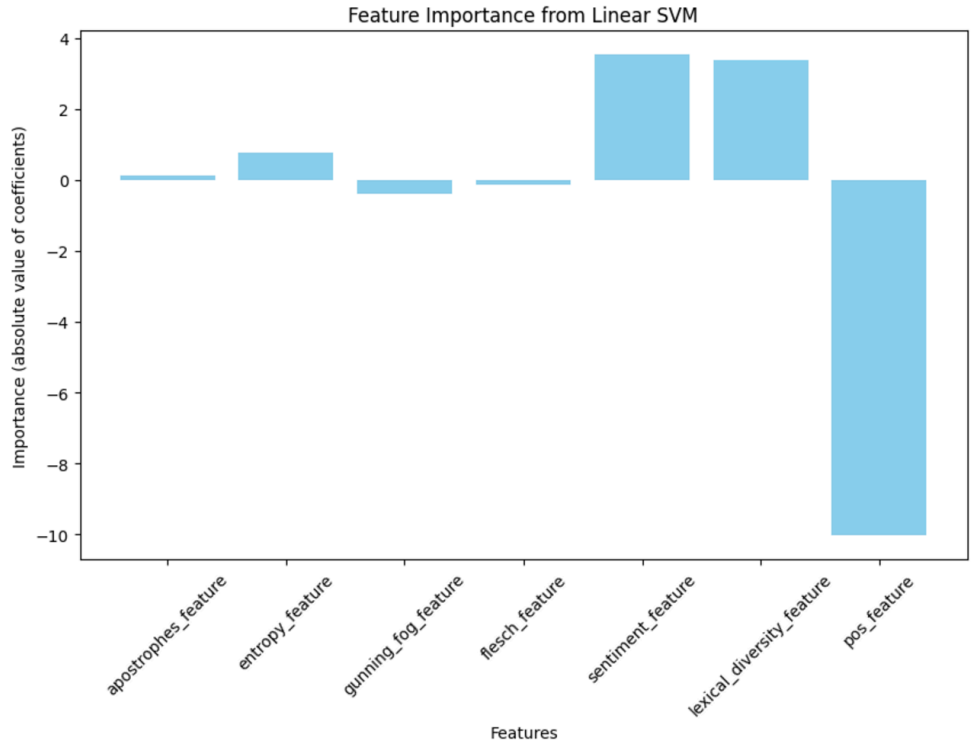
Feature Importance

GNB



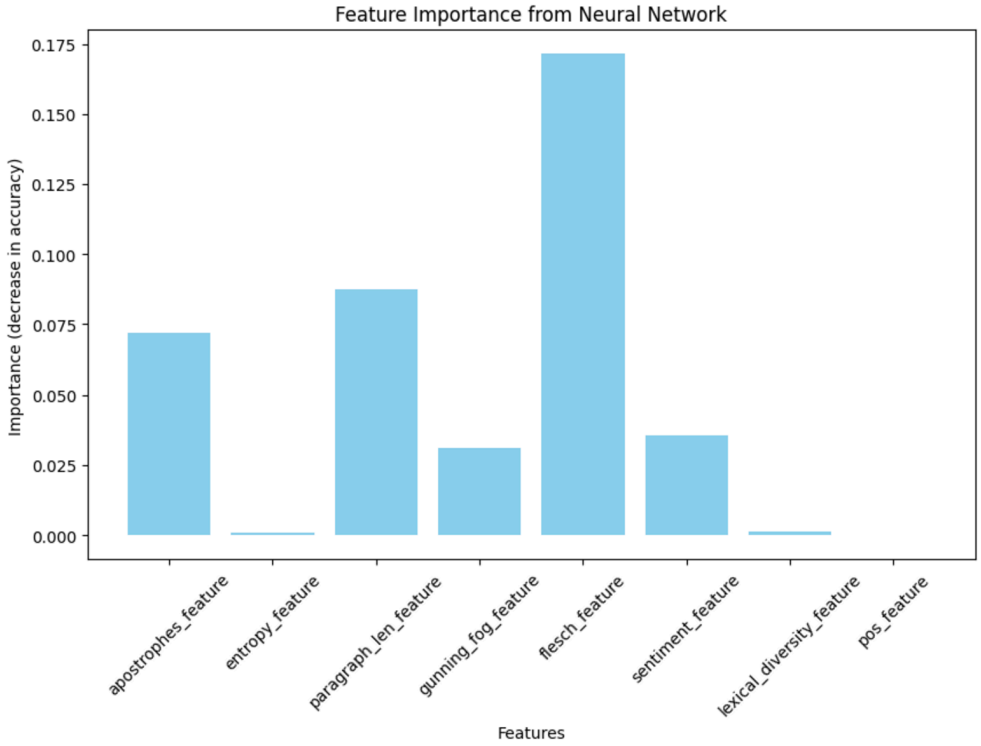
After removing unimportant features, GNB performed signifcantly better.

SVM



Removing more features did not significantly increase performance and in most cases led to some performance degradation.

NN

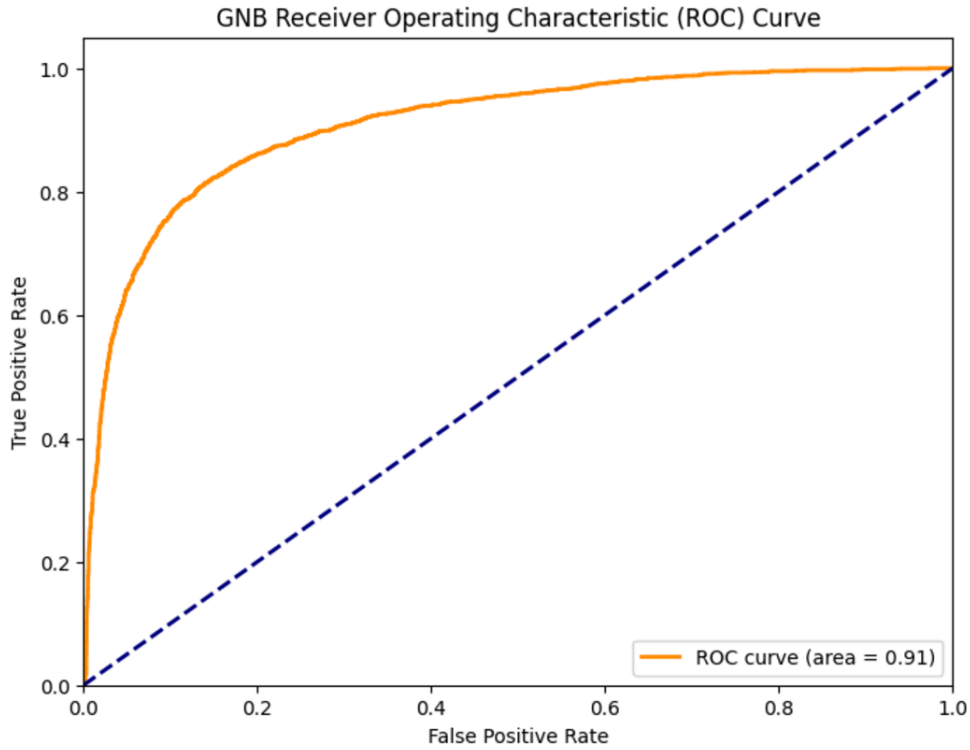


Even though this graph several unimportant features, we found that deleting them very slightly reduced accuracy.

ROC Curves

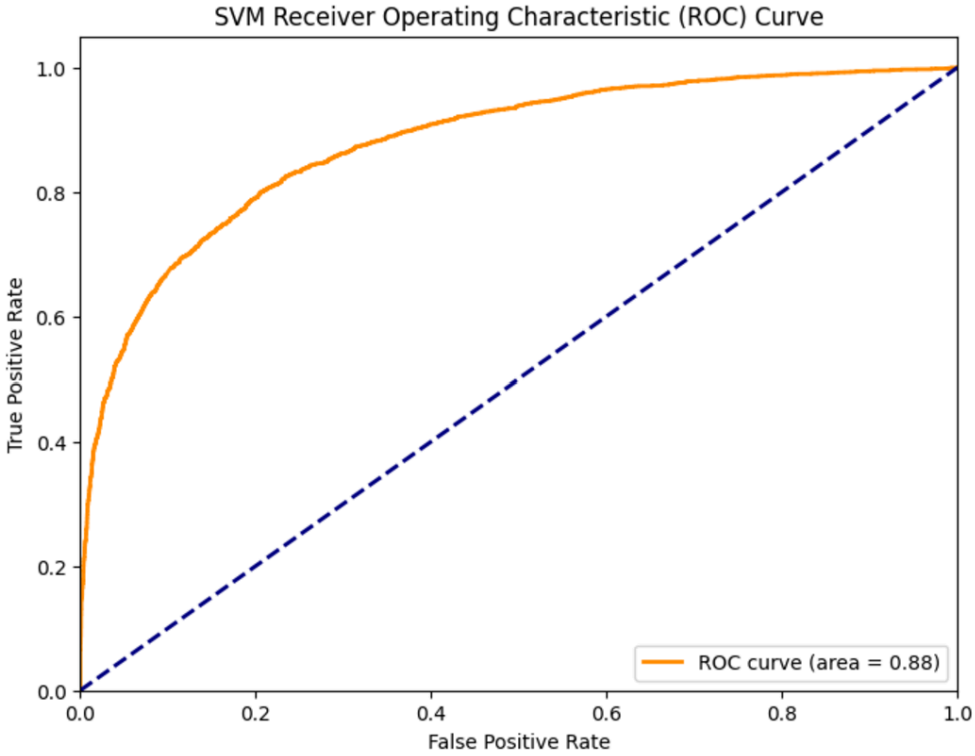
This curve represents the trade off between the True Positive Rate and the False Positive Rate. An ideal curve would have the AUC (Area Under Curve) be 1 and an AUC of 0.5 would be a completely random model. Even though this curve can be misleading with imbalance classes and our data set does have around a 65% human and 35% AI, it provides useful insights into the different models' ability to discriminate between classes.

GNB



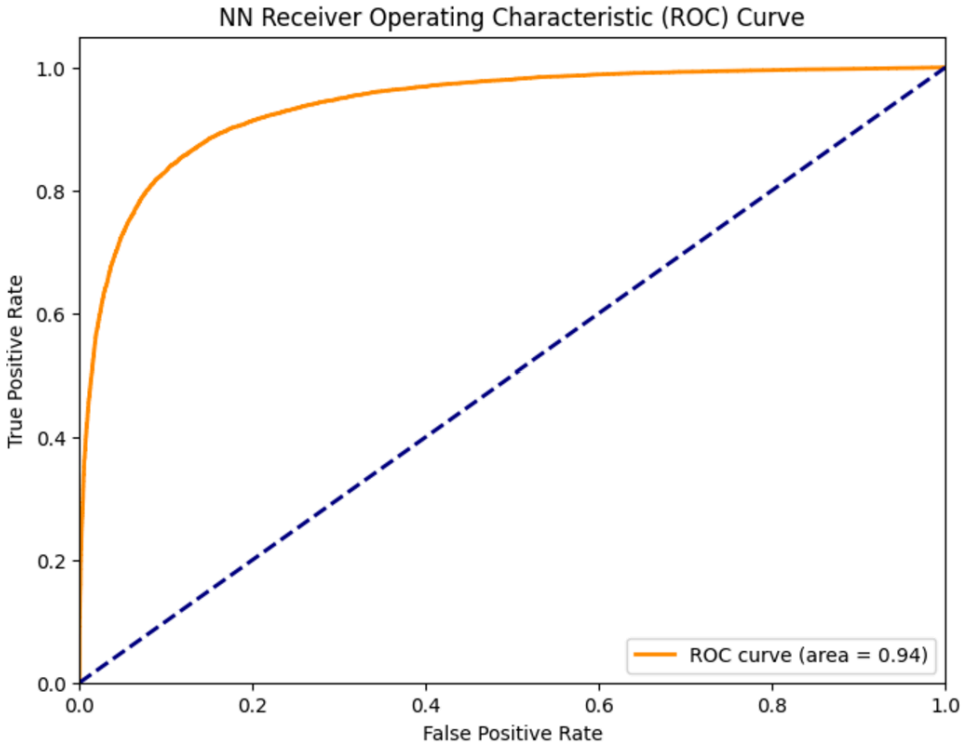
This model has an AUC of 0.91 which means the model is a great classifier.

SVM



This model has an AUC of 0.88 which means the model is a good classifier.

NN



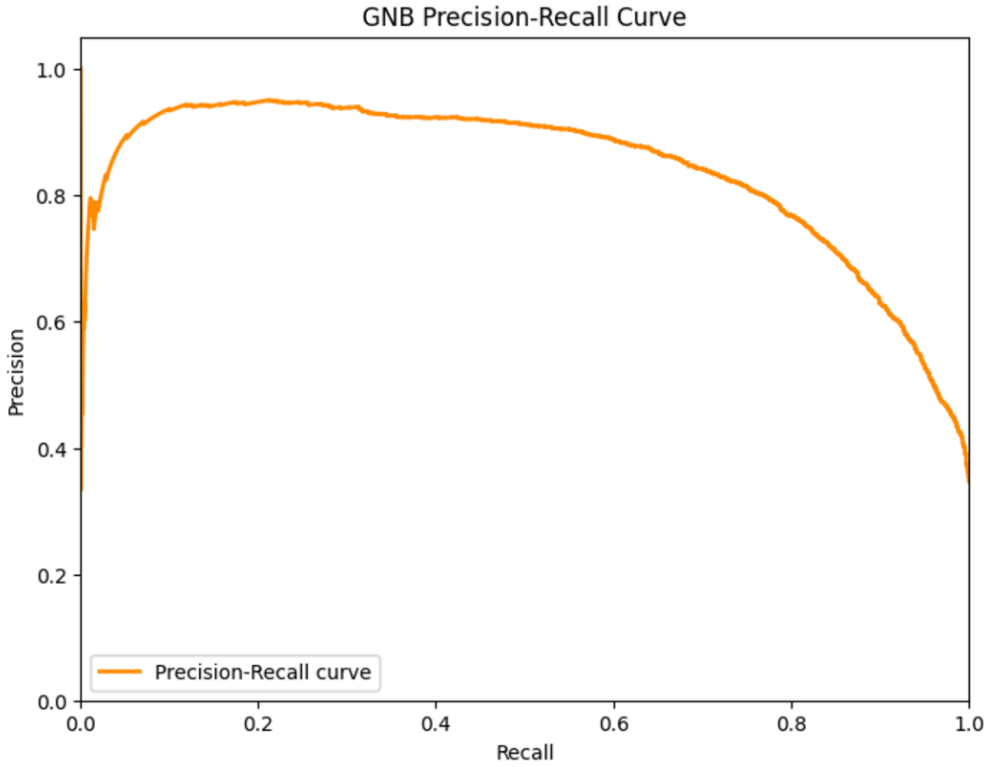
This model has the best ROC Curve with an AUC of 0.94 An AUC of 0.94 is generally considered to mean that the model is a great classifier. This means that it will correctly classify a piece of text as AI written or Human written 94% of the time when distinguishing between the two classes.

Precision-Recall Curves

This curve represents the trade off between precision and recall. We include this graph to show that we can correctly distinguish between classes while minimizing false positives. This curve is also important because the ROC curve could be misleading since our data set has some imbalance.

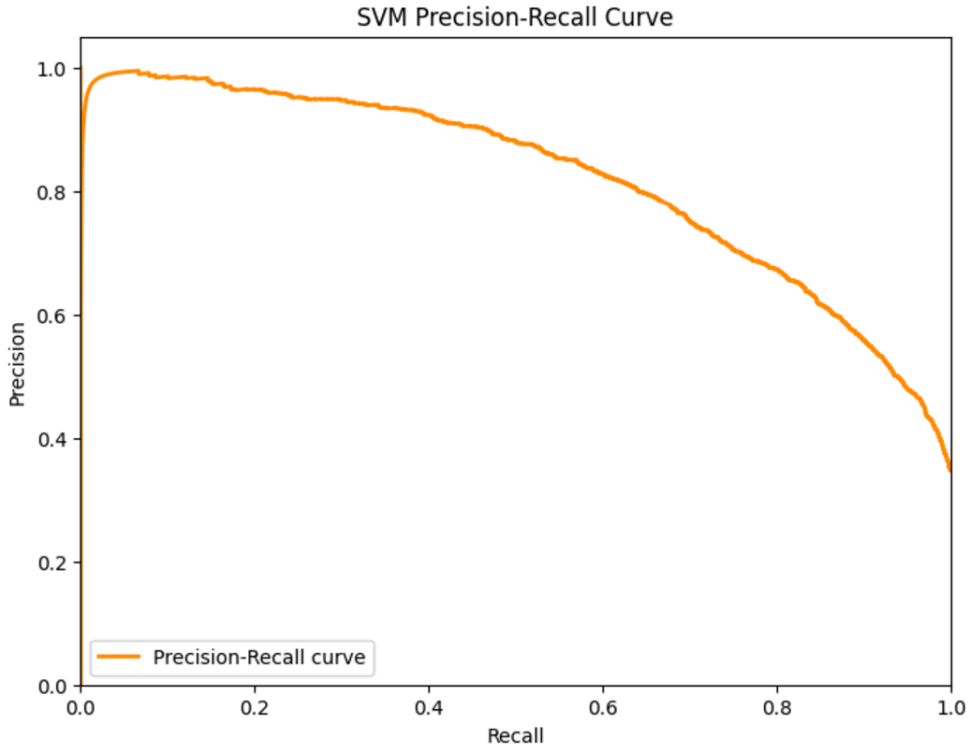
An ideal Precision-Recall Curve would be a horizontal line to y = 1 and then becoming a vertical line at x = 1.

GNB



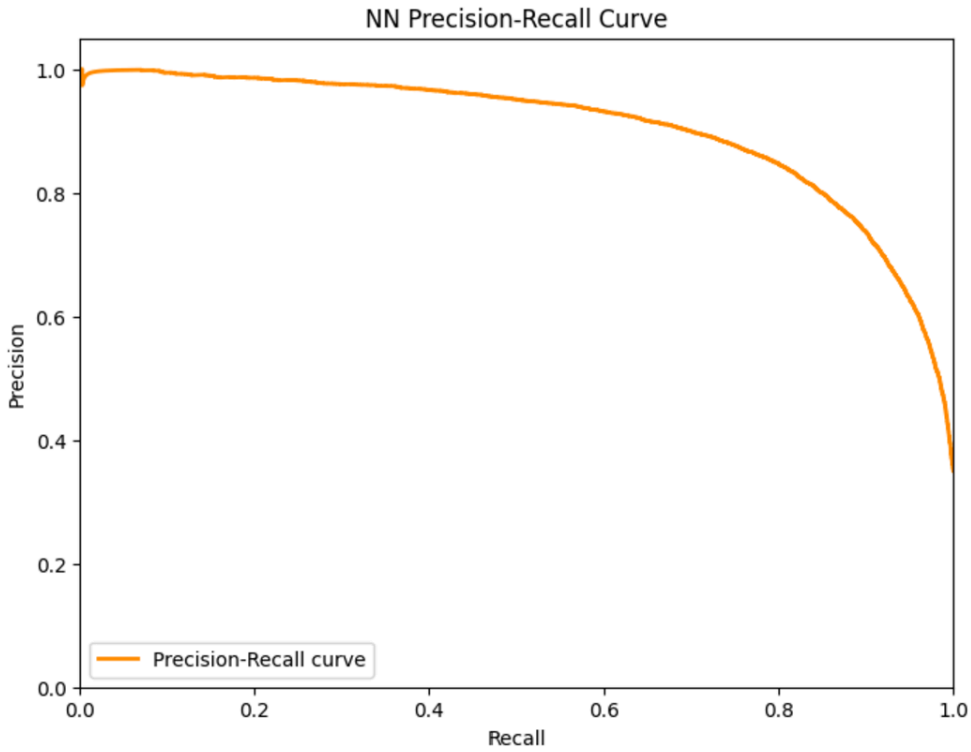
This Precision-Recall Curve shows the large trade off between precision and recall for this model.

SVM



This Precision-Recall Curve shows the large trade off between precision and recall for this model.

NN



This Precision-Recall curve is quite good, The two two values are quite balanced. We do wish the curve was slightly closer to (1,1), but it strikes a good balance overall.

Gantt Table

High-Level Gantt Table

GANTT CHART

PROJECT TITLE	AI Essay Detection

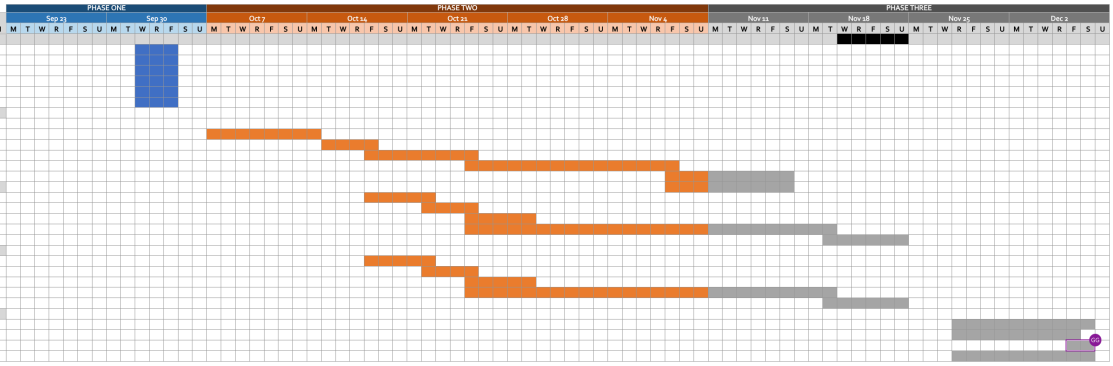
				PHASE ONE		PHASE TWO					PHASE THREE			
TASK TITLE	START DATE	DUE DATE	DURATION	Sep 27	Oct 4	Oct 11	Oct 18	Oct 25	Nov 1	Nov 8	Nov 15	Nov 22	Nov 29	Dec 6
Project Proposal	9/27/2021	10/7/2021	10											
Model 1	10/7/2021	11/16/2021	39											
Model 2	10/18/2021	11/24/2021	36											
Model 3	10/18/2021	11/24/2021	36											
Evaluation	11/24/2021	12/7/2021	13											

Extended Gantt Table

GANTT CHART

PROJECT TITLE: AI-Generated Text Detector

TASK TITLE	TASK OWNER	START DATE	DUE DATE	DURATION
Project Proposal				
Introduction & Background	Mingkuan	10/2/2024	10/4/2024	3
Problem Definition	Annelise	10/2/2024	10/4/2024	3
Methods	Greg + Ronojoy	10/2/2024	10/4/2024	3
Potential Results & Discussion	Mingkuan	10/2/2024	10/4/2024	3
Video Recording	Greg + Annelise	10/2/2024	10/4/2024	3
GitHub Page	Andrew	10/2/2024	10/4/2024	3
Model 1				
Data Sourcing and Cleaning	Andrew	10/7/2021	10/14/2021	7
Model Selection	Annelise	10/15/2021	10/18/2021	4
Data Pre-Processing	Greg	10/18/2021	10/25/2021	8
Model Coding	Mingkuan	10/25/2021	11/8/2021	15
Results Evaluation and Analysis	Ronojoy	11/8/2021	11/16/2021	9
Midterm Report	All	11/8/2021	11/16/2021	9
Model 2				
Data Sourcing and Cleaning	Andrew	10/18/2021	10/22/2021	5
Model Selection	Ronojoy	10/22/2021	10/25/2021	4
Data Pre-Processing	Greg	10/25/2021	10/29/2021	5
Model Coding	Mingkuan	10/25/2021	11/19/2021	25
Results Evaluation and Analysis	Annelise	11/19/2021	11/24/2021	6
Model 3				
Data Sourcing and Cleaning	Andrew	10/18/2021	10/22/2021	5
Model Selection	Annelise	10/22/2021	10/25/2021	4
Data Pre-Processing	Mingkuan	10/25/2021	10/29/2021	5
Model Coding	Greg	10/25/2021	11/19/2021	25
Results Evaluation and Analysis	Ronojoy	11/19/2021	11/24/2021	6
Evaluation				
Model Comparison	All	11/29/2021	12/7/2021	8
Presentation	All	11/29/2021	12/6/2021	7
Recording	All	12/6/2021	12/7/2021	1
Final Report	All	11/29/2021	12/7/2021	8



References

[1] X. Peng, Y. Zhou, B. He, L. Sun, and Y. Sun, “Hiding the Ghostwriters: An Adversarial Evaluation of AI-Generated Student Essay Detection,” Feb. 2024. Available: https://arxiv.org/pdf/2402.00412

[2] M. Maktab Dar Oghaz, K. Dhame, G. Singaram, and L. Babu Saheer, “Detection and Classification of ChatGPT Generated Contents Using Deep Transformer Models,” www.techrxiv.org, Aug. 09, 2023. https://www.techrxiv.org/articles/preprint/Detection_and_Classification_of_ChatGPT_Generated_Contents_Using_Deep_Transformer_Models/23895951/1 (accessed Sep. 06, 2023).

[3] M.-G. Kim and H. Desaire, “Detecting the Use of ChatGPT in University Newspapers by Analyzing Stylistic Differences with Machine Learning,” Information, vol. 15, no. 6, p. 307, Jun. 2024, doi: https://doi.org/10.3390/info15060307.

[4]I. Cingillioglu, “Detecting AI-generated essays: the ChatGPT challenge,” The International Journal of Information and Learning Technology, vol. 40, no. 3, May 2023, doi: https://doi.org/10.1108/ijilt-03-2023-0043.

Contribution Table

Contribution Table:

Name	**Midterm Contributions**
Andrew	- Experimented with ensemble approaches - GMM implementation - Presentation metrics
Annelise	- SVM and NN implementation - Feature analysis - Compiled Final Results and Visuals
Greg	- SVM and NN implementation - Feature analysis - Compiled Final Results and Visuals
Mingkuan	- Tested GMM, did not yield expected results, so pivoted to other ML methods - Model descriptions + reasoning on the website - Presentation visuals
Ronojoy	- Researched new potential features to include within training and experimented with overall accuracy - Minor presentation edits