Team 9 – Human vs AI Final Report

Logan Caraway
University of Tennessee
Knoxville, Tennessee
lcaraway@vols.utk.edu

Bella Matasic
University of Tennessee
Knoxville, Tennessee
bmatasic@vols.utk.edu

Alyssa Yzabelle Quijano
University of Tennessee
Knoxville, Tennessee
aquijano@vols.utk.edu

Silvia Shenouda
University of Tennessee
Knoxville, Tennessee
selkomos@vols.utk.edu

Ashwin Vinod
University of Tennessee
Knoxville, Tennessee
avinod@vols.utk.edu

Abstract

This work goes on to investigate the challenges of distinguishing human-written text from AI-generated text using machine learning models. With AI models rapidly evolving, identifying who or what wrote a text has become more difficult. Our project evaluates both LinearSVC with TF-IDF and DistilBERT with a Kaggle dataset that has 1,000+ labeled samples. Both approaches achieved accuracies between 51% and 59%. These results indicate that AI models can only be as good as the data it is trained on and highlights the importance of high-quality large datasets to reliably classify human-written and AI-generated texts.

Keywords—*LinearSVC, DistilBERT, machine learning, transformers, text classification, artificial intelligence*

1. Introduction

With Artificial Intelligence (AI) models rapidly evolving, it's getting harder to identify human-written and AI-generated works apart. Newer AI models are capable of mimicking how humans write text which makes it seem fluent. The Penn State Information Knowledge and wEb (PIKE) Lab conducted an experiment that shows that humans distinguished AI-generated and human written texts apart only 53% of the time [3]. This creates challenges, especially in areas where the author of the materials matter. For example, in academic settings, instructors need to make sure that students are submitting their own work and not just using AI to generate texts. In journalism, readers want to verify the content they are reading to make sure that it is credible and accurate. These are just some of the fields where this issue of differentiating AI and human texts persists.

The existing AI-detection tools have shown a lot of variability when it comes to distinguishing AI vs Human-written texts. GPTZero claims that it can accurately detect AI-usage with a 99% accuracy, and when tested, it does perform within the 95%-100% accuracy range as of July 2024, according to the paper "Detecting AI-Generated Writing Using GPTZero" [4]. The PIKE Lab also built an AI solution that analyzes texts which provides an answer with 85% to 95% accuracy [3].

Our project aims to address this problem by seeing if we can develop a model that accurately and reliably distinguish between human-written text and AI-generated text with the Kaggle dataset provided to us.

2. Technical Approach

Our initial approach to this problem was to use a linear support vector machine to help distinguish the AI versus human text. Our hypothesis was that there should be features that would distinguish the line between AI vs human text and that could be plotted with a hyperplane using an SVM. At first, we used the extensive numeric values present in the dataset (which will be discussed later) but quickly decided to drop these in favor of using only the text content. We used TF-IDF (term frequency versus

inverse document frequency) to vectorize the text into numeric values the classifier could understand. Then we used unigrams and bigrams to create 141,442 features and passed that into a scikit-learn LinearSVC model that used a standard regularization value. This resulted in an accuracy of 59.12%. As seen in the word cloud in figure 1, the top positive features (i.e. the features that push the model the most towards AI) are primarily noise. Therefore, this model was unable to capture a deep relationship between the AI and human text. It is worth mentioning that before we settled on using an SVM, we did experiment with using Logistic Regression with the numeric features. However, the performance was rather poor in comparison to the SVM (with the highest being around 48%). Using the SVM with both the numeric features and the text content also yielded a poor performance. Considering that the final model used an NLP, this was the first and last time these numeric features were considered in this problem.

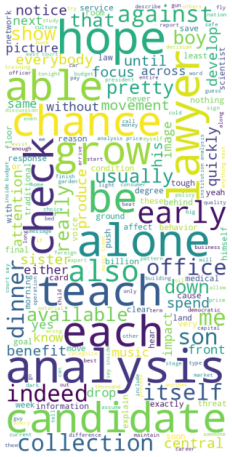

Figure 1: Word cloud depicting top positive coefficients for LinearSVC.

Our next approach to this problem was to move beyond classical machine-learning models and apply a transformer-based natural language processing method using DistilBERT. Our hypothesis was that because DistilBERT is a pretrained language model capable of understanding linguistic context and relationships between words, it would be better able to distinguish between AI-generated and human-written text.

To begin, the dataset was split into training and testing subsets using an 80/20 ratio. We then used a pretrained DistilBERT tokenizer that was applied to convert each text sample into subword token identifiers. Both the cased and uncased variants of DistilBERT were evaluated, but the uncased version produced better results because the removal of capitalization removed unnecessary noise. After tokenization, the outputs were wrapped into a custom PyTorch dataset class to allow the HuggingFace Trainer to

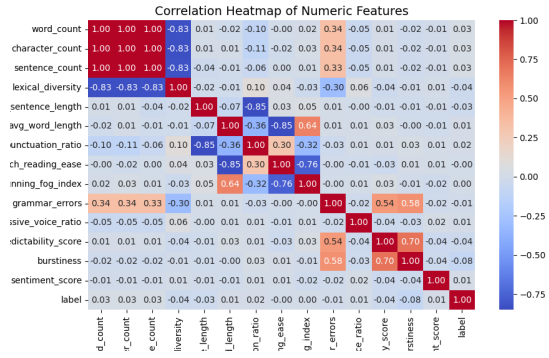

Figure 2: Heatmap visualization of numerical features

feed examples into the model during training.

The pretrained DistilBERT model was then loaded and configured for two output labels corresponding to AI and human text. Training was performed using the HuggingFace Training Arguments interface with a learning rate of 3×10^{-5}, a batch size of 8, a weight decay of 0.01, and a total of five epochs. The HuggingFace Trainer managed the full optimization and evaluation loop during training.

During training, we also experimented with replacing DistilBERT's default loss function (cross-entropy loss) with BCEWithLogitsLoss. To do this, we created a custom subclass of the HuggingFace Trainer and overrode the compute_loss method so the model's raw logits were passed directly into BCEWithLogitsLoss. We also added label smoothing of 0.1 to reduce overconfidence. However, BCEWithLogitsLoss assumes each output logit represents an independent binary decision, which did not match our single-label classification setup, so it performed worse in practice. As a result, we returned to using DistilBERT's standard cross-entropy loss.

This DistilBERT approach achieved an accuracy of approximately 59.48%. However, the model's performance was unstable, which we believe is due to the limited size of the dataset. With only approximately 1,000 samples available, the model did not have sufficient data for DistilBERT to fully learn the deeper patterns required to differentiate between AI-generated and human-written text.

3. Dataset Description

The dataset we used for this project was from Kaggle and was titled "*AI vs Human Content Detection 1000+ record in 2025*". It contained 1,367 samples of both AI-

generated and human-written text. For each sample, the text content was accompanied by numerical values for attributes like Gunning-Fog Index, Flesch reading ease, word/character count, and much more. Figure 2 displays a heatmap of each of these numerical values. From this figure we can see that there is a high inverse relationship between word lengths and Flesch score. There is also a strong negative correlation between Gunning Fog and Flesch scores. Many of the numerical values are also heavily right skewed. One place where this is evident is in the word counts of the various text pieces. The highest word count is 447 words, whereas the lowest is 3. Figure 3 is a histogram showcasing this. Some other features that were right-skewed are sentence count, character count, and grammar errors.

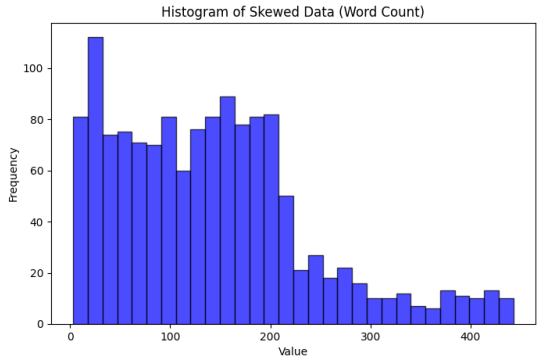

Figure 3: Histogram showing the distribution of word counts across all samples

Besides the numeric features, there is of course the text content and the label. The label was 1 for AI text and 0 for human text. During the exploratory analysis phase, we observed that a substantial amount of text content consisted of syntactically correct but incoherent and semantically incorrect text. One such example is the following: "*Pass enjoy production stand process. Low common wall life matter husband billion."* The documentation for the dataset is sparse and does not explain why the text is laid out like this. We later realized late into the project that all of the samples exhibited the same linguistic style. Although it is not documented in the source and the source does not explain how the text was generated, one possible explanation is that the author asked an AI model to generate deliberately incoherent text and then attempted to replicate this themselves, or vice versa. Of course, this cannot be confirmed from the dataset itself.

As mentioned previously, the dataset consisted of 1,367 samples split evenly between AI and human text. While this amount is sufficient for classical or smaller-scale machine learning techniques, it is significantly smaller than the scale typically expected by large transformer models like DistilBERT. As a result, this imposed restrictions on later model performance.

4. Results and Discussion

4.1. Baseline Linear SVC Performance

Our first set of experiments evaluated a linear Support Vector Machine using TF-IDF representations. The model incorporated unigrams and bigrams, resulting in approximately 141 thousand sparse features. As shown in the baseline results table 1 below, the model achieved an overall accuracy of **59.1 percent** with a macro F1 score near **0.59**. Class 1, which represented AI generated text, showed a notably higher recall than Class 0, indicating that the classifier was more likely to flag text as AI than human.

TABLE1 : BASELINE PERFORMANCE

Metric	***Class 0***	***Class 1***	***Macro Avg***
Precision	0.619	0.574	0.597
Recall	0.474	0.708	0.591
F1-Score	0.537	0.634	0.586

Figure 1 word cloud illustrates the highest weighted features. These features were largely noise instead of meaningful linguistic markers that distinguish human text from AI-generated text. This suggests that the TF-IDF approach captured superficial lexical correlations but was not able to model deeper stylistic or semantic signals. The result aligns with known limitations of linear models when applied to complicated language classification tasks, especially when patterns extend beyond individual words or short n-grams.

We also tested logistic regression with the numeric features provided in the dataset, as well as a combined model using both numeric features and TF-IDF embeddings. These approaches produced inferior performance. The highest accuracy obtained with logistic regression was approximately **48 percent**, which confirmed that the numeric features contributed minimal discriminative value and introduced additional noise. These results motivated the decision to discard the numerical readability and style metrics for the remainder of the modeling pipeline.

Taken together, the SVM experiments established a reasonable baseline but demonstrated clear limitations. The classifier could not capture the richer linguistic structures that modern AI systems generate. This finding provided justification for transitioning to transformer-based methods.

4.2. DistilBERT Fine-Tuning Results

To evaluate whether contextual embeddings and transformer-based language understanding would improve performance, we fine-tuned the DistilBERT model for binary classification. Both cased and uncased variants were tested. The uncased variant achieved higher accuracy because removing capitalization reduced unnecessary variability in the dataset.

The model was trained for five epochs with a learning rate of 3 x 10^-5, a batch size of 8, and weight decay of 0.01. The HuggingFace Trainer framework handled batching, optimization, and evaluation. The training and validation loss curves shown in Figure 4 reveals unstable model behavior. While the training loss decreased steadily, the validation loss oscillated and increased during multiple points in training, indicating overfitting.

Figure 4: Training vs. validation loss curve

The best observed accuracy across runs was approximately **59 percent**, which is consistent with the baseline SVM but not demonstrably higher. Moreover, the performance fluctuated between **54 percent and 59 percent** depending on the random seed and fold. This instability reflects a core limitation of the dataset. With only 1367 samples, DistilBERT lacked sufficient data to reliably learn the nuanced differences between AI and human written text. Large transformer models typically require tens of thousands of examples for stable fine-tuning. In contrast, our dataset size is more aligned with classical machine learning regimes.

Even with this limitation, the model did learn patterns beyond the SVM baseline, as indicated by small improvements in specific runs and more balanced precision-recall tradeoffs. However, the gains were inconsistent and not statistically significant.

4.3. K-Fold Cross Validation

To obtain a more reliable measurement of generalization performance, we implemented five-fold cross validation. The LinearSVC achieved an average accuracy of **52 percent**, while DistilBERT achieved approximately **51 percent**. These values were lower than single-split experiments and confirmed that both models struggle to generalize reliably across different subsets of the dataset.

The reduced performance in cross validation further supports the conclusion that the dataset is too small and too homogeneous in style to allow either classical or transformer-based models to form robust decision boundaries. The dataset also contained syntactically correct but incoherent samples for both classes. This undermines the model's ability to learn meaningful stylistic distinctions because both labels share overlapping linguistic artifacts.

4.4. Dataset Factors Influencing Performance

The dataset characteristics played a dominant role in limiting the performance of all tested models. Several factors contributed to this:

1. *Small sample size.* With 1367 samples split evenly between the two classes, the dataset is significantly smaller than the scale typically required for modern NLP models. Transformer models expect large and diverse corpora.
2. *Right-skewed numerical distributions.* As shown in the histograms and heatmaps included Figure 2, Figure 3, Figure 5, and Figure 6 features like word count, character count, and grammar errors were heavily skewed. These features did not provide meaningful separation between classes.

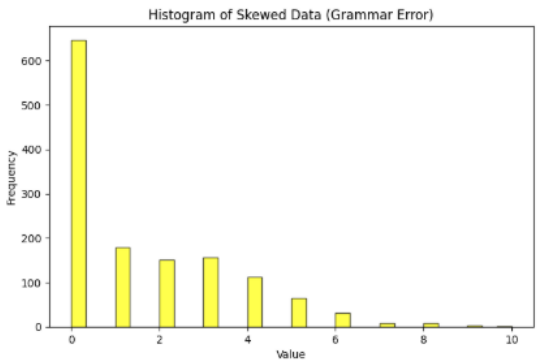

Figure 5: Histogram depicting the skew of data for grammar errors

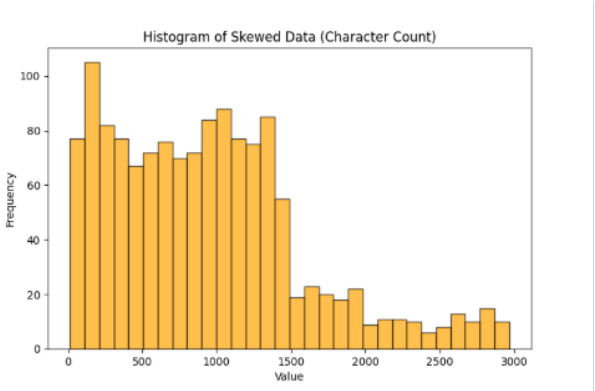

Figure 6: Histogram depicting the skew of data for character growth

3. *Inconsistent text quality.* Many samples contained incoherent sentences such as "Pass enjoy production stand process." These samples were present in both human and AI categories, which suggests that the dataset creator may have produced or modified samples using a mixture of synthetic and manually written text. This inconsistency reduces the reliability of the label assignments.

4. *Lack of stylistic diversity.* Nearly all samples shared an unusually similar linguistic style, which makes it difficult for models to learn real-world distinctions between human-written text and AI-generated content.

These dataset limitations explain why both SVM and DistilBERT converged around similar accuracy levels near random chance for a binary classification task.

4.5. Results

Our experiments demonstrate that model complexity alone does not guarantee improved performance in AI detection tasks. When the dataset lacks linguistic diversity, contains label noise, or is too small for transformer-based learning, even advanced models struggle to outperform classical approaches.

The results also mirror findings from recent literature. Prior studies have shown that as large language models become more fluent, distinguishing them from human text becomes increasingly difficult. The performance ceilings of our models reflects this broader trend.

Although DistilBERT provides richer representations of language, it cannot overcome fundamental dataset constraints. The instability in validation loss and cross-validation scores confirms that the dataset prevented the model from learning generalizable patterns.

However, the project clarifies several directions for improvement. Data augmentation through back translation, generating additional AI samples, or collecting more human-written text could help expand the training set. Improving dataset diversity and reducing incoherent samples may also allow transformer models to leverage their full capacity.

4.6. Summary

Both modeling approaches achieved accuracies between 51 percent and 59 percent, with neither model demonstrating consistent superiority. Classical TF-IDF features with LinearSVC produced a stable but limited baseline. DistilBERT provided more expressive modeling capacity but was constrained by data scarcity, leading to unstable results.

Overall, the experimental findings highlight that the boundary between AI and human text is becoming increasingly subtle and that reliable detection requires large, diverse, and clean datasets. The models developed in this project provide a foundation, but the dataset size and structure ultimately limited their ability to generalize.

5. Conclusion

Our first approach of implementing a solution involved experimenting with various machine learning models, and using a LinearSVC model with TF-IDF features gave the best performance among them, reaching an accuracy of 59.12%. While this approach performed better than logistic regression with numeric features, it was not reliable enough to distinguish between AI-generated and human written text. The top features identified by SVM were mostly noise, which suggested that the model was not learning meaningful patterns and was struggling to find any strong relationships between AI and human text.

We then used DistilBERT, a transformer-based model, to see if it could find any patterns that the SVM could not. We used the same dataset with 1367 samples and achieved an accuracy of 59%, which was similar to the SVM's performance. During our exploratory analysis, we found that much of the text followed an unusual, syntactically correct but semantically incoherent style for AI and human generated text. Due to these similarities that blurred the distinction between texts, the model had very little linguistic cues to learn from. If the dataset had more samples and was more diverse in content, the model could have performed better.

6. Workload Distribution

6.1. Logan Caraway

- Technical Approach (SVM Section)
- Dataset Description
- Initial dataset transformation (numerical values)
- Hyperparameter tuning/exploration for LinearSVC
- Performance curves for DistilBERT, various other visualizations

6.2. Bella Matasic

- Results and Discussion
- Initial Baseline Solution
- Baseline Results
- Project Management
- References

6.3. Alyssa Yzabelle Quijano

- Abstract
- Introduction and Overview
- Confusion Matrix of LinearSVC and DistilBERT
- Helped Test Other Baseline Models
- Compiled Code for Submission (PDF/GitHub)
- Project management
- References

6.4. Silvia Shenouda

- Technical Approach (DistilBERT section)
- Implemented full DistilBERT training pipeline
- Built DistilBERT K-Fold evaluation framework
- Built LinearSVC K-Fold evaluation framework

6.5. Ashwin Vinod

- Conclusion
- Data Preprocessing
- Correlation Heatmap

References

[1] FPaullet, K., Pinchot, J., Kinney, E., & Stewart, T. (2025). Detecting AI-generated writing using gptzero. Information Systems Education Journal, 23(6), 44–52. https://doi.org/10.62273/pzww7741

[2] Elkhatat, A.M., Elsaid, K. & Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 19, 17 (2023). https://doi.org/10.1007/s40979-023-00140-5

[3] "Q&A: The increasing difficulty of detecting AI- versus human-generated text | Penn State University." https://www.psu.edu/news/information-sciences-and-technology/story/qa-increasing-difficulty-detecting-ai-versus-human

[4] S. Dik, O. Erdem, and M. Dik, "Assessing GPTZero's Accuracy in Identifying AI vs. Human-Written Essays," Abstract, [Online]. Available: https://arxiv.org/pdf/2506.23517