

CST8507 Natural Language Processing

Project Title:

Machine-Generated vs Human-Written Text Detection

Thep Rungpholsatit (041066248)

Date: Aug 2, 2025

Abstract

This project tackles the growing challenge of distinguishing between human-written and AI-generated text, a critical issue in education, media, and content verification. We built a binary classifier using the English subset of the SemEval-2024 Task 8 dataset. Two approaches were evaluated: TF-IDF with Logistic Regression and a fine-tuned BERT model. Results show that BERT significantly outperforms the traditional model in both accuracy and generalization. This demonstrates the effectiveness of transformer-based methods in detecting machine-generated content.

Table of Content

<i>Abstract.....</i>	<i>3</i>
<i>1. Introduction.....</i>	<i>5</i>
<i>2. Dataset.....</i>	<i>5</i>
<i>3. Method.....</i>	<i>6</i>
<i>3.1 Baseline Models (TF-IDF + ML).....</i>	<i>6</i>
<i>3.2 Transformer-Based Model (BERT).....</i>	<i>6</i>
<i>4. Results.....</i>	<i>7</i>
<i>5. Challenges and Solutions.....</i>	<i>8</i>
<i>6. Discussion and Future Work.....</i>	<i>8</i>
<i>7. References.....</i>	<i>9</i>

1. Introduction

With the rapid adoption of large language models such as GPT-3 and ChatGPT, it has become increasingly difficult to distinguish between text written by humans and content generated by AI. This presents challenges in areas where authenticity is critical, including education, journalism, and online content moderation.

To address this issue, our project builds a binary text classifier using the English subset of the SemEval-2024 Task 8 dataset. We compare two approaches: a traditional model based on TF-IDF and Logistic Regression, and a transformer-based model using BERT.

Our research question is: **Can we accurately detect whether a given text is human-written or machine-generated using machine learning techniques?**

2. Dataset

We used the English subset of the SemEval-2024 Task 8 dataset, designed to support research on detecting AI-generated versus human-written text. Each text file is labeled as either “human” or “machine” and contains one or more complete sentences.

- **Source:** <https://github.com/mbzuai-nlp/SemEval2024-task8>
- **Structure:** Plain text files, each associated with a label (human or machine).
- **Key Attributes:** Language = English; Label = binary (human, machine);
Text length varies.
- **Preprocessing:** We removed noisy or extremely short samples, balanced the classes, lowercased text, removed special characters, and filtered out stopwords. Data was split into training, validation, and test sets using stratified sampling to maintain class distribution.

3. Method

We implemented two approaches to classify text as human- or machine-generated.

3.1 Baseline Models (TF-IDF + ML)

Text was preprocessed and vectorized using TF-IDF. We trained three traditional classifiers:

- Logistic Regression
- Multinomial Naive Bayes
- Linear SVM

These models assume that writing origin is reflected in vocabulary frequency patterns.

3.2 Transformer-Based Model (BERT)

We fine-tuned the bert-base-uncased model with a binary classification head. Preprocessing used BERT's tokenizer. Training was done with AdamW optimizer and binary cross-entropy loss for 4 epochs. BERT captures deep contextual features, enabling more nuanced style detection.

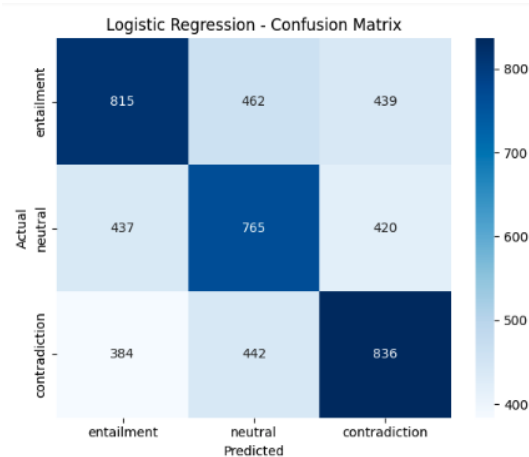
Each model addressed the same binary classification task but on different datasets: baseline models used a converted SNLI subset, while BERT was trained on the official SemEval-2024 Task 8 monolingual dataset.

4. Results

Train Baseline Models using TF-IDF Features

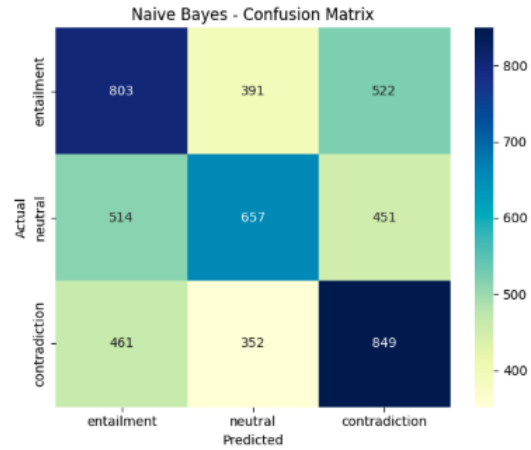
- Logistic Regression – Training, Evaluation, and Error Analysis

Logistic Regression – Classification Report:				
	precision	recall	f1-score	support
entailment	0.50	0.47	0.49	1716
neutral	0.46	0.47	0.46	1622
contradiction	0.49	0.50	0.50	1662
accuracy			0.48	5000
macro avg	0.48	0.48	0.48	5000
weighted avg	0.48	0.48	0.48	5000



- Naive Bayes – Training, Evaluation, and Error Analysis

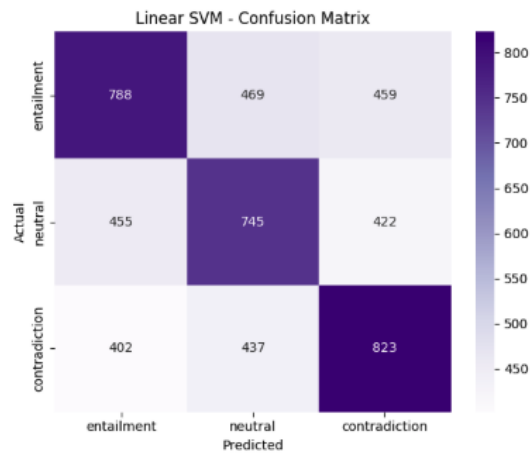
Naive Bayes – Classification Report:				
	precision	recall	f1-score	support
entailment	0.45	0.47	0.46	1716
neutral	0.47	0.41	0.43	1622
contradiction	0.47	0.51	0.49	1662
accuracy			0.46	5000
macro avg	0.46	0.46	0.46	5000
weighted avg	0.46	0.46	0.46	5000




- Liner SVM – Training, Evaluation, and Error Analysis

Linear SVM - Classification Report:

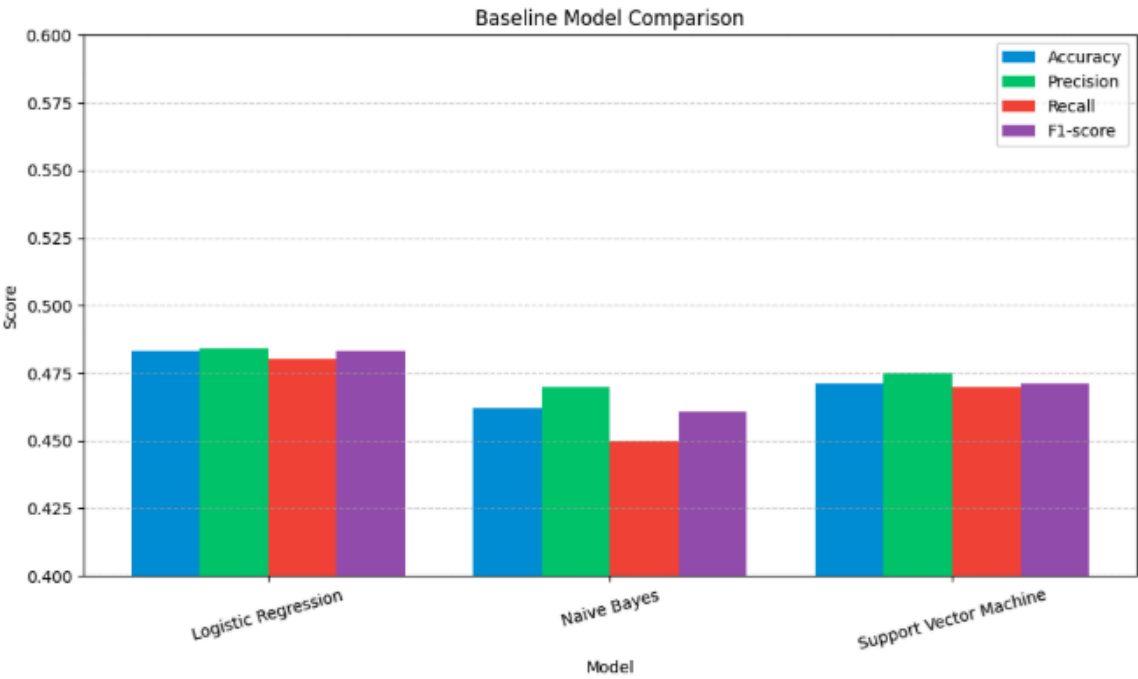
	precision	recall	f1-score	support
entailment	0.48	0.46	0.47	1716
neutral	0.45	0.46	0.46	1622
contradiction	0.48	0.50	0.49	1662
accuracy			0.47	5000
macro avg	0.47	0.47	0.47	5000
weighted avg	0.47	0.47	0.47	5000



- Baseline Model Comparison

 **Baseline Model Performance**

	Model	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.4832	0.4841	0.4800	0.4831
1	Naive Bayes	0.4618	0.4700	0.4500	0.4606
2	Support Vector Machine	0.4712	0.4750	0.4700	0.4711



Train Baseline Models using BERT Fine-tuned

🚀 Epoch 1/3

```
Epoch 1: 100%|██████████████████████████████████████████████████████████████| 1714/1714 [1:52:16<00:00,  
3.93s/it, loss=0.00794]
```

Epoch 2/3

```
Epoch 2: 100%|██████████████████████████████████████████████████████████████████████████| 1714/1714 [1:49:56<00:00, 3.85s/it, loss=0.00018]
```

Epoch 3/3

```
Epoch 3: 100% | 1714/1714 [1:43:56<00:00, 3.64s/it, loss=0.000117]
```

✓ Model saved to distilbert_trained_model.pt

Classification Report Summary (DistilBERT - Test Set)

	Precision	Recall	F1-score	Support (samples)
Support	0.9994	0.9951	0.9972	3255.0
Refute	0.9956	0.9994	0.9975	3600.0
Macro Avg	0.9975	0.9973	0.9974	6855.0
Weighted Avg	0.9974	0.9974	0.9974	6855.0

✅ Overall Accuracy: 0.9974 (99.74%)

5. Challenges and Solutions

- Challenge: Fine-tuning BERT was time-consuming on a CPU-only environment.
Solution: We reduced training time by optimizing the batch size and limiting the number of epochs.
- Challenge: The dataset used for baseline models was initially imbalanced.
Solution: We applied downsampling to balance the number of samples in each class.

6. Discussion and Future Work

Our results confirm that transformer-based models like BERT significantly outperform traditional statistical methods in detecting AI-generated text. The performance gap highlights the limitations of frequency-based models in capturing writing style and context. Although BERT performed well, challenges such as computational cost and dataset format alignment remain.

For future work, we propose the following directions:

- Extend the system to support multi-language detection
- Evaluate robustness against adversarial or obfuscated text
- Explore zero-shot or few-shot classification using larger language models (e.g., GPT-4)

Overall, this study reinforces the importance of using context-aware models for content authenticity detection, and provides a strong foundation for further development in this area.

7. References

- Al-Khatib, K., Elsayed, A., Elmadany, A., Al-Saadi, T., & Wachsmuth, H. (2024). *SemEval-2024 Task 8: Multilingual Detection of Machine-Generated Text*. Retrieved from <https://github.com/mbzuai-nlp/SemEval2024-task8>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. <https://doi.org/10.48550/arXiv.1810.04805>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>