# Team ID: 3
# Team Name: Bayes and Bert Text Source Detectors

## Comparison of Multinomial Naïve Bayes and BERT in Detecting Human vs. Machine-Generated Text


## Abstract

This project compares Multinomial Naïve Bayes (MNB) and BERT-based models (including DistilBERT) for detecting human vs. machine-generated text using the SemEval Task 8 and GenAI datasets. The task is complex, requiring models to identify nuanced features between human and machine text. Evaluation focuses on accuracy, precision, recall, and F1-score. Results show that BERT-based models outperform MNB, though MNB still captures some patterns effectively. DistilBERT offers a more efficient alternative to BERT. Hyperparameter tuning plays a critical role in optimizing performance.


## Goal of the Project

- Compare Multinomial Naïve Bayes (MNB) and BERT-based models (including DistilBERT) for detecting human vs. machine-generated text using SemEval Task 8 and GenAI datasets.
- Through comparison, understand the complexity of detecting nuanced features in machine-generated text.
- Understand each model's strengths and weaknesses.
- Assess the impact of hyperparameters on model performance.
- Compare performance across key metrics like accuracy, precision, recall, and F1-score.

## Methodology

1. **Data Preprocessing:**
   - Loaded SemEval and GenAI datasets; labeled Human (0), Machine (1).
   - Stratified GenAI split: 90% train, 10% test; BERT: small subset for validation.
   - BERT/DistilBERT: Preprocessed with BERT tokenizer, max length = 128, padding/truncation applied.
   - MNB: Lowercasing, whitespace handling, punctuation, number, stopword removal, tokenising based on spaces for vocabulary building and likelihood calculations.
2. **Model Implementation:**
   - Built binary classifiers with PyTorch for MNB, BERT and DistilBERT.
   - DistilBERT: Smaller BERT variant with fewer layers, no NSP pre-training, and knowledge distillation.
   - Initial hyperparameters:
     - BERT: Batch size = 16, Learning rate = $2 \times 10^{-5}$, Epochs = 4.
     - DistilBERT: Batch size = 16, Learning rate = $2 \times 10^{-5}$, Epochs = 5.
     - MNB: Smoothing factor = 0.2
3. **Model Training:**
   - Fine-tuned on SemEval and GenAI datasets.
   - Hyperparameter tuning for BERT/DistilBERT:
     - Batch size: 16, 64.
     - Learning rate: $2 \times 10^{-5}$ to $1 \times 10^{-4}$.
     - Epochs: 4, 5.
   - Smoothing factor varied as part of hyperparameter tuning for MNB [0.1, 0.2, 0.5, 1.0, 2.0, 3.0]
   - Best settings: For MNB based on smoothing factors - SemEval: 0.5, GenAI: 3.0; For BERT models: Batch size = 16, Learning rate = $2 \times 10^{-5}$; BERT: 3 epochs, DistilBERT: 4 epochs
4. **Evaluation:**
   - Tested on GenAI and SemEval sets.
   - Metrics: Accuracy, precision, recall, F1-score. Training time included for BERT vs. DistilBERT. MNB has comparatively much lesser training time.

## Results

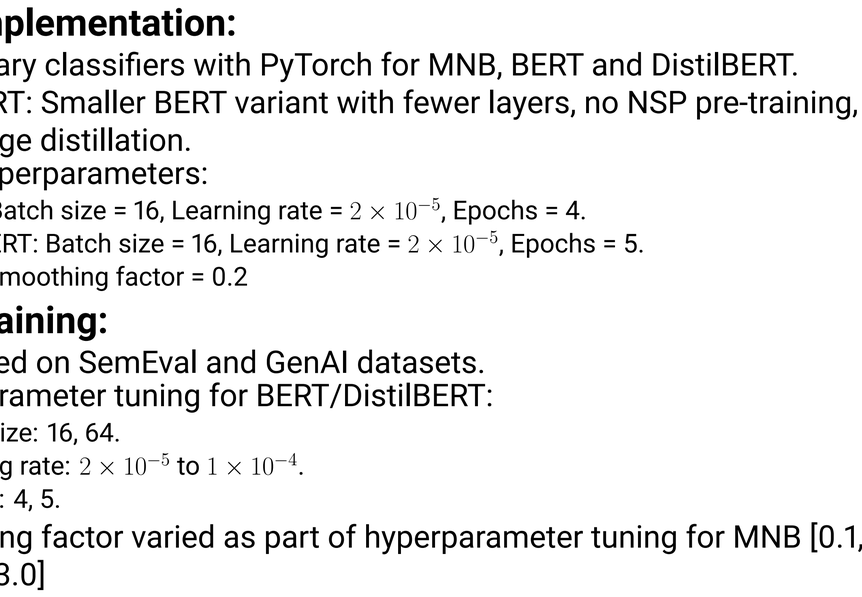


Figure 1. Hyperparameter Tuning and Training Time Comparison Results.

## Results

Table 1. Performance Comparison Across Models and Datasets

| Model | Dataset | Accuracy % | Precision % | Recall % | F1-score % |
|---|---|---|---|---|---|
| BERT | Train (SemEval) | 94.51 | 90.38 | 98.87 | 94.44 |
| | Train (GenAI) | 97.39 | 96.05 | 99.94 | 97.96 |
| | Validation (SemEval) | 94.52 | 90.21 | 99.88 | 94.80 |
| | Validation (GenAI) | 93.56 | 91.25 | 99.20 | 95.06 |
| | Test (SemEval) | 52.77 | 52.65 | 99.99 | 68.98 |
| | Test (GenAI) | 93.56 | 91.27 | 99.19 | 95.06 |
| DistilBERT | Train (SemEval) | 94.22 | 89.65 | 99.19 | 94.18 |
| | Train (GenAI) | 97.81 | 96.66 | 99.95 | 98.28 |
| | Validation (SemEval) | 93.62 | 88.74 | 99.92 | 94.00 |
| | Validation (GenAI) | 93.78 | 91.58 | 99.16 | 95.22 |
| | Test (SemEval) | 52.85 | 52.70 | 99.97 | 69.01 |
| | Test (GenAI) | 93.95 | 91.79 | 99.20 | 95.35 |
| MNB | Train (SemEval) | 72.14 | 72.69 | 72.14 | 72.14 |
| | Train (GenAI) | 74.34 | 76.56 | 74.34 | 74.73 |
| | Validation (SemEval) | 51.44 | 51.44 | 51.44 | 51.40 |
| | Validation (GenAI) | 69.52 | 71.92 | 69.52 | 69.99 |
| | Test (SemEval) | 84.13 | 84.80 | 84.13 | 84.12 |
| | Test (GenAI) | 69.43 | 71.83 | 69.43 | 69.91 |

## Analysis

- **Performance Metrics:**
  - **BERT/DistilBERT:** High accuracy (94%-98%), precision (91%-96%), and recall (99%) due to contextualized embeddings.
  - **MNB:** Strong on SemEval Test set (84.13%) but weaker elsewhere (51.44%), with lower precision (76%) but reasonable recall (84%).
- **Generalization:**
  - **BERT/DistilBERT:** Generalized well, except on SemEval Test, where high recall led to misclassifications with potential overfitting with GenAI.
  - **MNB:** Stable generalization (69%-84%), capturing basic text patterns effectively.
- **Efficiency:**
  - **BERT:** High accuracy but slow (5 hours/epoch).
  - **DistilBERT:** Faster (1.5–2.5 hours/epoch), nearly as effective.
  - **MNB:** Extremely fast (in minutes), suitable for quick results in resource-limited settings.
- **Conclusion:**
  - **BERT/DistilBERT:** Ideal for nuanced, high-accuracy tasks, leveraging pre-trained contextual embeddings to capture subtle linguistic differences; DistilBERT offers faster, more efficient performance.
  - **MNB:** Performed well but fell short of BERT in capturing complex patterns. Likely succeeded due to predictable LLM outputs (e.g., repetition, low entropy). A solid baseline for simpler tasks.

## Limitations and Future Work

- **What would we change?**
  - Simplify fine-tuning with `BertForSequenceClassification`.
  - Explore other transformer or simpler models beyond BERT and MNB.
  - Further tune learning rate, batch size, sequence length, epochs, and regularization.
  - Experiment with skip gram models instead of just CBOW for MNB.
- **What simplifications did we assume?**
  - Assumed 128 tokens were sufficient.
  - Underestimated resources for large GenAI data.
  - Did not consider the impact of data imbalance on model performance; while GenAI training was balanced, other data splits were not.
  - Treated Machine text from various sources (e.g., ChatGPT, Cohere, BLOOMz) as a single unified class, overlooking source-specific nuances or biases.
- **What research question would we address with more time?**
  - How effective are fixed BERT embeddings vs. fine-tuning for classification?
  - Can smaller architectures like MobileBERT match DistilBERT's performance?
  - How well does fine-tuned BERT generalize to multilingual datasets for this nuanced task?
  - Understand the scope better to calculate the trade off in using a higher time consuming model like BERT vs faster models like MNB. Is the cost to the environment worth the benefits?

## References


[1] J. Shaikh, *Machine learning, nlp: Text classification using scikit-learn, python and nltk*, `https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a`.

[2] H. Face, *Bert*, `https://huggingface.co/docs/transformers/en/model_doc/bert`.