Team #3 - Bayes and Bert Text Source Detectors
Comparison of Multinomial Naïve Bayes and BERT-based models in Detecting Human vs. Machine-Generated Text

Mithila Kandasamy; 40062939; mithila.kandasamy@gmail.com
and
Apurba Das; 40263612; apurba.das@mail.concordia.ca

December 11, 2024

Abstract. This project compares Multinomial Naïve Bayes (MNB), BERT, and DistilBERT for classifying human vs. machine-generated text using SemEval and GenAI datasets. Results show that BERT and DistilBERT outperformed MNB across all performance metrics by leveraging pretrained contextual embeddings. However, MNB performed surprisingly well, potentially due to predictable patterns in machine-generated content. DistilBERT was more efficient, offering similar performance to BERT with faster training. The findings highlight the need for effective models as machine-generated text becomes increasingly fluent in sounding ”human”.

Project code and data are available here.

Contents

1 Goal of the project

The goal is to compare Multinomial Naïve Bayes (MNB) and BERT-based models (BERT and DistilBERT) for detecting human vs. machine-generated text using both the SemEval and GenAI Shared Task monolingual datasets. By doing this, we aim to understand the complexity required in detecting nuanced features in machine-generated text, especially as large language models (LLMs) have become quite proficient at sounding humane. This highlights the importance of assessing how complex the classification task needs to be in order to differentiate between human and machine-generated text effectively. We will evaluate each model's strengths and weaknesses by comparing the models' performance across key metrics, including accuracy, precision, recall, F1-score, and computational efficiency.

2 Methodology

1. **Data Preprocessing:**
   - Loaded SemEval (Task 8, Subtask A) and GenAI (Task 1, Subtask A) English datasets from the shared task GitHub, with texts labeled as Human (0) or Machine (1). Since test set was unavailable, stratified sampling split was applied to GenAI data into 90% train and 10% test, with a smaller subset created for validation at SemEval sizes for quick BERT/DistilBERT validation.
   - BERT/DistilBERT: Used BERT's tokenizer with padding and truncation, setting the maximum sequence length to 128[3]. Retained stopwords and special characters for contextual meaning.
   - MNB: Lowercasing, removal of stopwords, newline, non-alphabets
2. **Model Implementation:**
   - Defined BERT, DistilBERT and MNB binary classifiers:
     - BERT: Used pre-pooled `[CLS]` output for classification tasks.
     - DistilBERT: Extracted `[CLS]` embedding from the `last_hidden_state` (first token from last hidden layer) for classification.
   - Initial hyperparameters used for assessment and debugging on SemEval subset: BERT (Batch size: 16, Learning rate: 2×10^{-5}, Epochs: 4), DistilBERT (Batch size: 16, Learning rate: 2×10^{-5}, Epochs: 5), MNB (Smoothing factor: 0.2).

3. **Model Training and Hyperparameter Tuning:**

   - Hyperparameter tuning for BERT/DistilBERT [9] included varying batch size (16, 64), learning rate (2×10^{-5} to 1×10^{-4}), and epochs (4, 5). A subset of GenAI data was used for quicker validation, while the full SemEval dataset was used due to its smaller size.
   - Hyperparameter tuning for MNB included varying smoothing factor at [0.1, 0.2, 0.5, 1.0, 2.0, 3.0].
   - Recorded and plotted validation accuracy to assess best hyperparameter settings. Training time was also recorded for comparison purposes.
   - **Best Settings:**
     - BERT: Batch size = 16, Learning rate = 2×10^{-5}, 3 epochs.
     - DistilBERT: Batch size = 16, Learning rate = 2×10^{-5}, 4 epochs.
     - MNB: Smoothing factor = 0.5 (SemEval), 3.0 (GenAI).
   - Final BERT and DistilBERT models were trained on GenAI training set (90% split), excluding SemEval due to longer training times and lower validation performance assessed. GenAI's larger, more updated data was considered sufficient for training.

4. **Evaluation:**

   - Tested on GenAI (10% test split) and SemEval test, train and validation sets. Evaluated accuracy, precision, recall, and F1-score and training time for each model.

Note: While the original Shared Task datasets were imbalanced, stratified splitting in GenAI ensured balance during training and validation. However, during evaluation, sets other than GenAI training, GenAI test and SemEval validation were imbalanced, potentially affecting performance assessment.
About DistilBERT: A compact version of BERT, DistilBERT has half the hidden layers and was not pre-trained on Next Sentence Prediction (NSP), using Masked Language Modeling (MLM) instead for the Classification Token (CLS). With knowledge distillation, it learns to approximate BERT's predictions, retaining 97% of its accuracy while being significantly more efficient [6].

3 Evaluation

3.1 Results

Table 1: Performance Comparison Across Models and Datasets (Best Test Metrics in Bold)

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**	**84.13**
	Test (GenAI)	69.43	71.83	69.43	69.91

See **Appendix A** for Hyperparameter Tuning and Training Time Comparison Plots. The plots for Hyperparameter Tuning were used to choose the best setting for the final model. Training time was used to compare the efficiency of each model.

3.2 Analysis

Overall Performance Metrics

- **Accuracy:** BERT and DistilBERT consistently outperformed MNB across datasets, except SemEval test, due to their contextual embeddings from pre-training and fine-tuning, allowing to capture complex linguistic features prevalent to each class. MNB, a simpler bag-of-words model, achieved 84.13% accuracy on the SemEval test set but struggled elsewhere, particularly on the SemEval validation set (51.44%). Nevertheless, its ability to perform above 50% shows some capacity to distinguish between human and machine-generated text even without contextual understanding.

- **Precision, Recall, and F1-Score:** BERT-based models outperformed MNB across datasets, achieving higher precision ($\approx 91\% - 96\%$), near-perfect recall ($\approx 99\%$), and F1-scores ($\approx 94\% - 98\%$), except on SemEval test set. Their contextual embeddings improved machine-generated text detection with fewer false positives and negatives. MNB, relying on word counts without context, had lower overall precision and F1-scores ($\approx 51\% - 84\%$) but performed better on the SemEval test set, with comparable recall ($\approx 84\%$) and slightly higher precision, potentially due to predictable patterns in that dataset.

- **Overfitting Noted at SemEval Test:** While BERT-based models achieved near-perfect recall ($\approx 99\%$) on the SemEval test set, their low precision and accuracy ($\approx 50\%$) suggest overfitting to GenAI data and poor adaptation to SemEval's specific characteristics.

- **Why MNB Didn't Perform too badly:** MNB performed well in detecting machine-generated text possibly due to its ability to exploit distinct word frequency patterns between human and AI text. Its simplicity, efficiency, and feature independence assumption align with the distribution of words in machine-generated text, making it a strong baseline for binary classification. Additionally, its robustness against overfitting enhanced its effectiveness. The GenAI dataset, which is a continuation of SemEval Shared Task 8 (Subtask A), incorporates outputs from novel LLMs. Further exploration of the machine-generated data sources could provide more insights into MNB's performance.

Generalization Across Datasets

- BERT and DistilBERT performed well ($> 90\%$) overall, but showed overfitting on the SemEval test set ($\approx 50\%$), highlighting difficulties in generalizing across datasets with potentially differing styles and vocabularies, like SemEval vs. GenAI.

- MNB's performance ranged a lot (around 51%-84%), which highlights how it depends on basic representations and struggles to capture word relationships. Its bag-of-words method makes it hard to generalize, especially on datasets that need contextual understanding.

Training Time Efficiency

While training time efficiency is often overlooked, we considered it to evaluate the carbon footprint and environmental impact of various models. It's crucial to choose models that meet our performance needs while minimizing ecological impact.

- BERT takes about 5 hours per epoch due to its complex architecture and large number of parameters, highlighting the trade-off between performance and computational cost.
- DistilBERT, training in 1.5-2.5 hours per epoch, offers comparable performance to BERT, making it a more time-efficient alternative.
- MNB is the fastest to train, with its simple structure not requiring a deep network to learn complex features, making it ideal for resource-limited applications.

In conclusion, BERT-based models are best suited for identifying linguistic differences between machine-generated and human-written text, with DistilBERT offering efficiency without sacrificing much performance. MNB, while simpler and lacking contextual understanding, remains a solid baseline, especially for resource-constrained scenarios.

4 Role of each team member

Each team member was responsible for implementing a model with clear expectations for the required outputs, such as training time, performance metrics (accuracy, precision, recall, and F1-score), and hyperparameter tuning across datasets. Apurba implemented the MNB classifier (custom and using standard Scikit-learn CountVectorizer), while Mithila implemented the BERT classifier. Since DistilBERT is a lighter version of BERT, Mithila used BERT's architecture as the foundation to implement and train DistilBERT. Throughout the process, we shared our discoveries, challenges, and progress with each other. The analysis portion of the project was done collaboratively.

5 Limitations

What Would We Change? We would simplify fine-tuning by using tools like `BertForSequenceClassification` [5] and try other models beyond BERT and MNB. Further fine-tuning hyperparameters like dropout

rate, max. sequence length and warm-up steps would help improve performance. We could test MNB without stopword removal to check its impact.
What Simplifications Did We Assume? Assuming 128 tokens were enough might have missed important context. We underestimated the computationally expensiveness of training large GenAI dataset. GenAI training was balanced but most other sets were not. We also grouped machine text from different LLM sources into one class, ignoring their unique patterns.
What Would We Explore With More Time? We would compare fixed BERT embeddings with fine-tuning and also test if smaller models like MobileBERT [7] perform as well as DistilBERT. We would also evaluate BERT's generalization ability on multilingual datasets. We would weigh the benefits of resource-heavy models like BERT vs. faster ones like MNB. Additionally, we would explore the datasets to understand why MNB performed well and why validation metrics like F1 score was low when using both GenAI and SemEval together, compared to when used separately.

6 Difference with your original proposal

Our original proposal was to compare only BERT with MNB. However, we also included DistilBERT, which we found to be the best option due to its comparable performance to BERT while offering better computational efficiency. For MNB, we also evaluated using CountVectorizer(a standard Scikit-learn tool), to assess the accuracy of the custom approach while understanding the added overhead due to flexibility of custom implementation. We found that using CountVectorizer was faster in terms of training time.

7 Conclusions

This project compared MNB, BERT, and DistilBERT for detecting human vs. machine-generated text using the SemEval and GenAI monolingual datasets. While MNB performed above 50patterns in machine-generated text in the data used, BERT and DistilBERT outperformed it by leveraging pre-trained knowledge and fine-tuned contex tual embeddings, enabling them to better handle the nuanced task. Distil BERT, a more efficient alternative to BERT, provided similar performance with reduced training time. This project highlighted the trade-offs between model complexity, computational cost, and performance. As LLMs continue to advance, the need for effective and efficient classification methods grows, making this a strong starting point for identifying the best approaches moving forward.

8 References

References

[1] GenAI Content Detection: Workshop on Detecting AI Generated Content. https://genai-content-detection.gitlab.io/.

[2] SemEval-2024: The 18th International Workshop on Semantic Evaluation. https://semeval.github.io/SemEval2024/tasks.

[3] Harshith Padigela, H. Z., and Croft, W. B. Investigating the Successes and Failures of BERT for Passage Re-Ranking. *arXiv preprint arXiv:1905.01758* (2019).

[4] HuggingFace. BERT. https://huggingface.co/docs/transformers/en/model_doc/bert.

[5] HuggingFace. BertForSequenceClassification — Hugging Face Transformers. https://huggingface.co/docs/transformers/v4.46.2/en/model_doc/bert#transformers.BertForSequenceClassification.

[6] HuggingFace. DistilBERT. https://huggingface.co/docs/transformers/en/model_doc/distilbert.

[7] HuggingFace. MobileBERT — Hugging Face Transformers. https://huggingface.co/docs/transformers/en/model_doc/mobilebert.

[8] HuggingFace. Transformers Preprocessing. https://huggingface.co/docs/transformers/v4.45.2/en/preprocessing.

[9] Morris, J. Does Model Size Matter? A Comparison of BERT and DistilBERT. https://wandb.ai/jack-morris/david-vs-goliath/reports/Does-Model-Size-Matter-A-Comparison-of-BERT-and-DistilBERT–VmlldzoxMDUxNzU.

[10] Shaikh, J. 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.

A Appendix

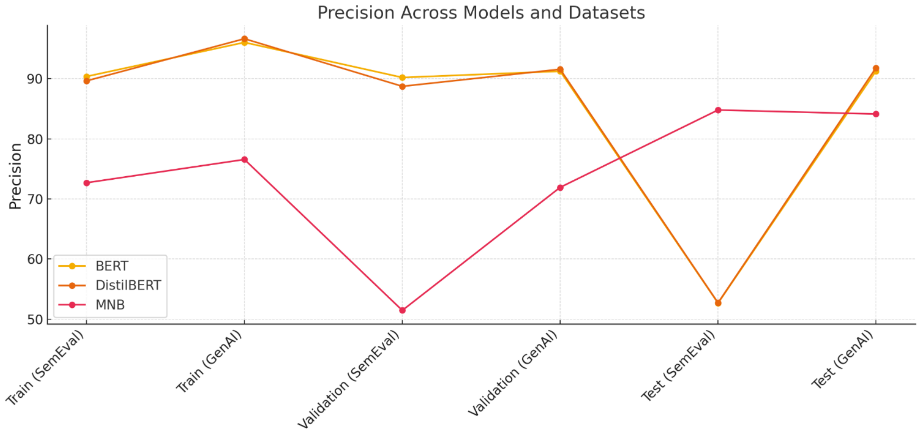

Figure 1: Precision across models

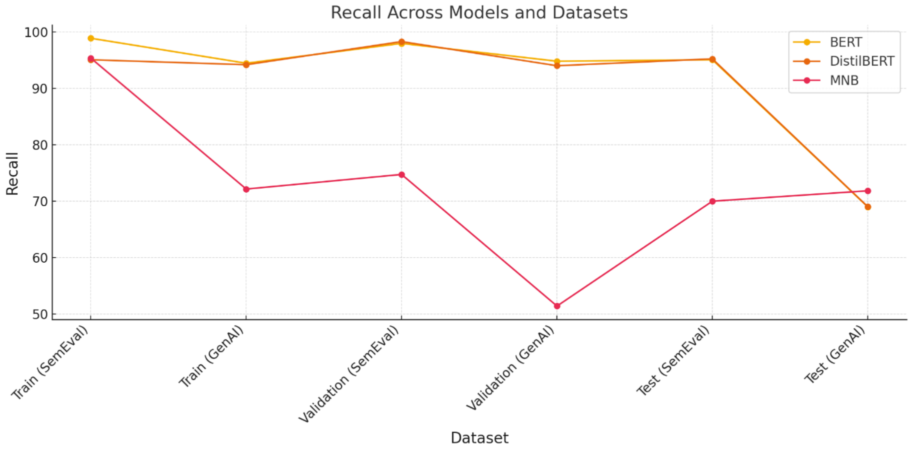

Figure 2: Recall across models

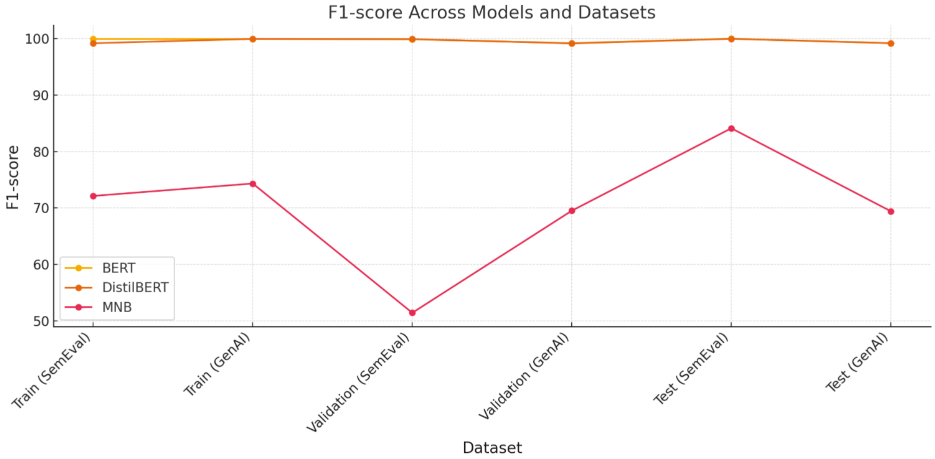

Figure 3: F1 score across models

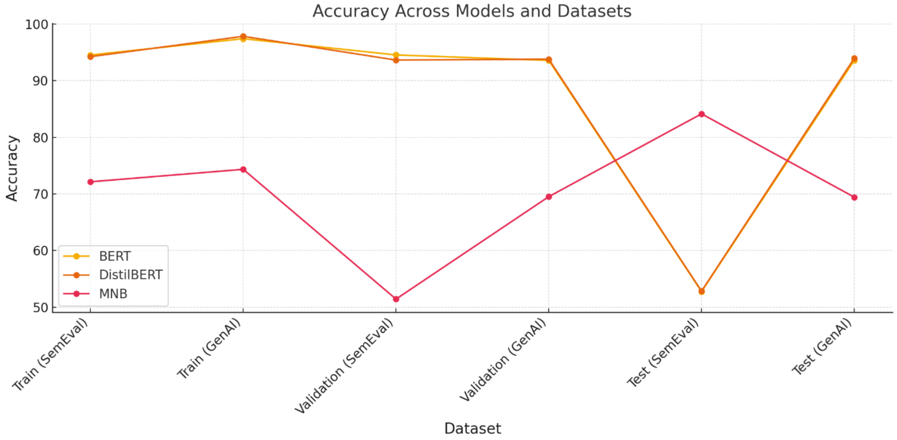

Figure 4: Accuracy across models

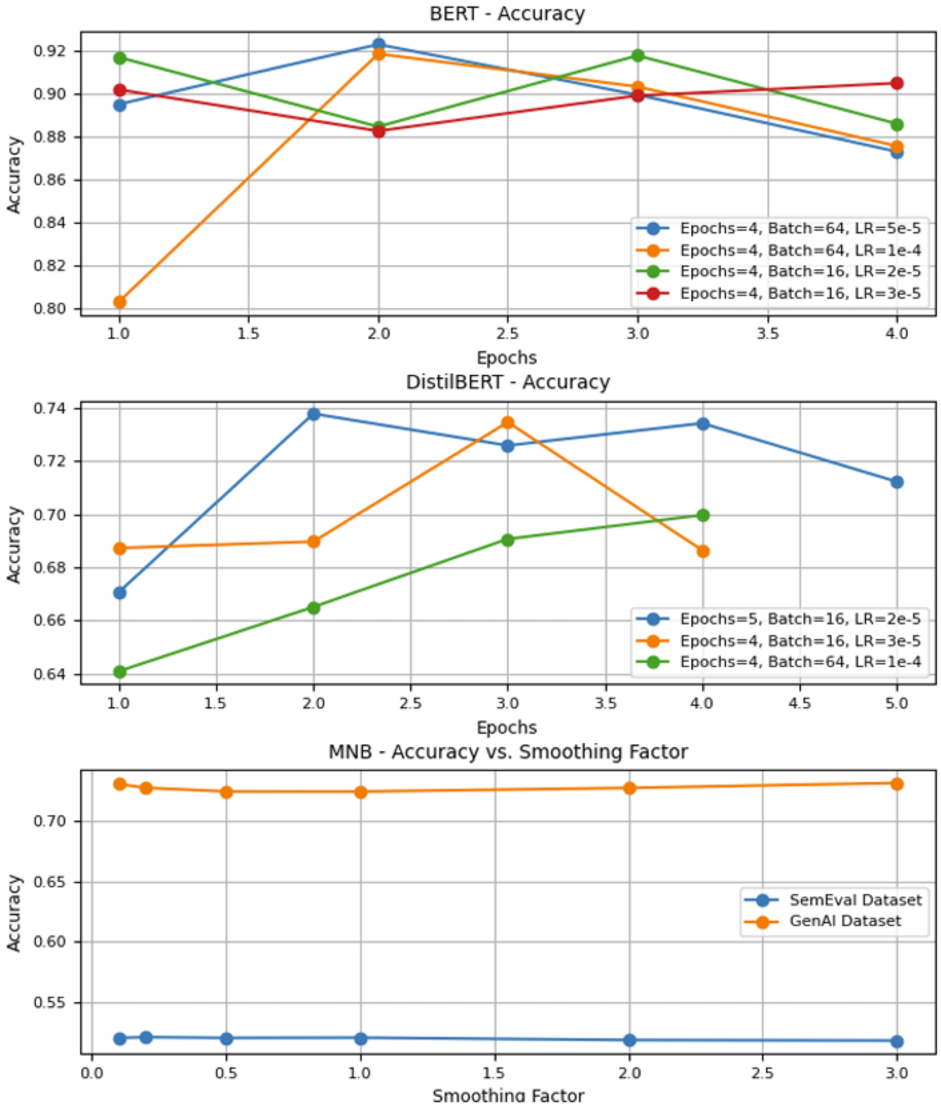

Figure 5: Hyperparameter Tuning - variation across epochs for bert, distilBert and smoothing for MNB

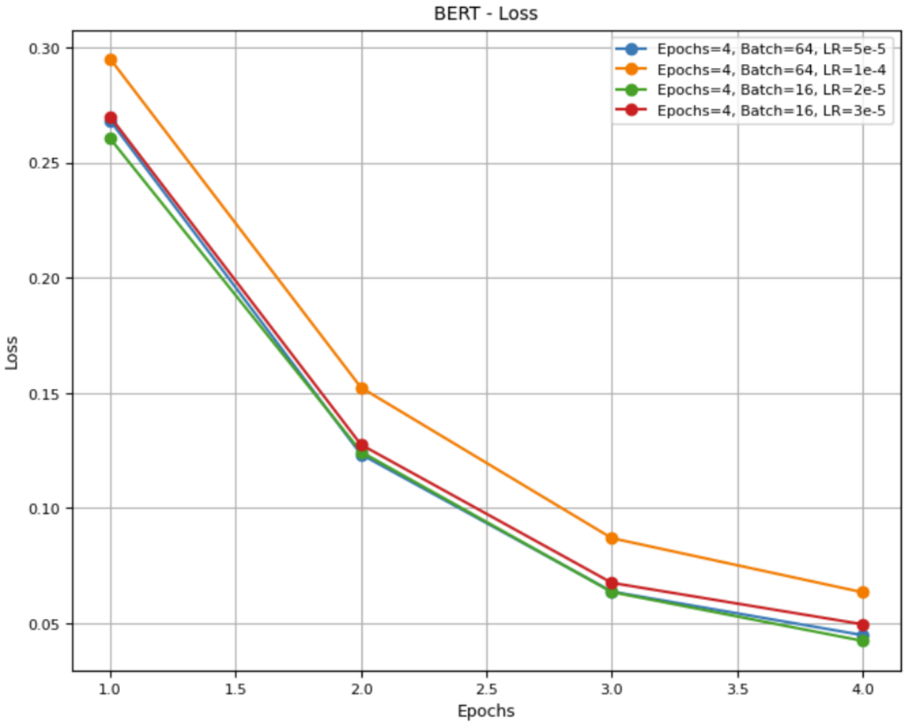

Figure 6: BERT - Loss through epochs

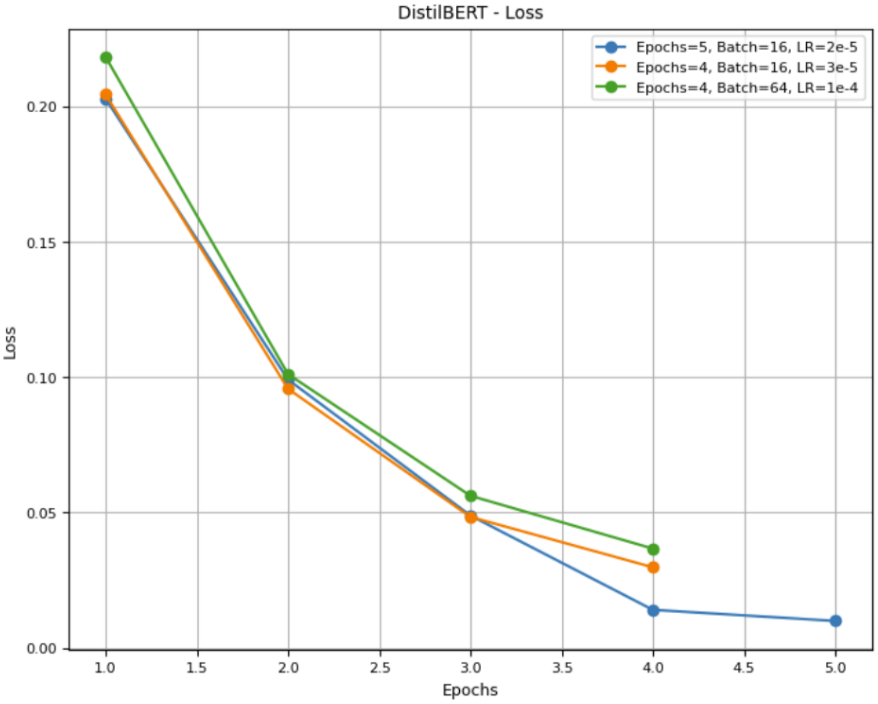

Figure 7: DistilBERT - Loss through epochs

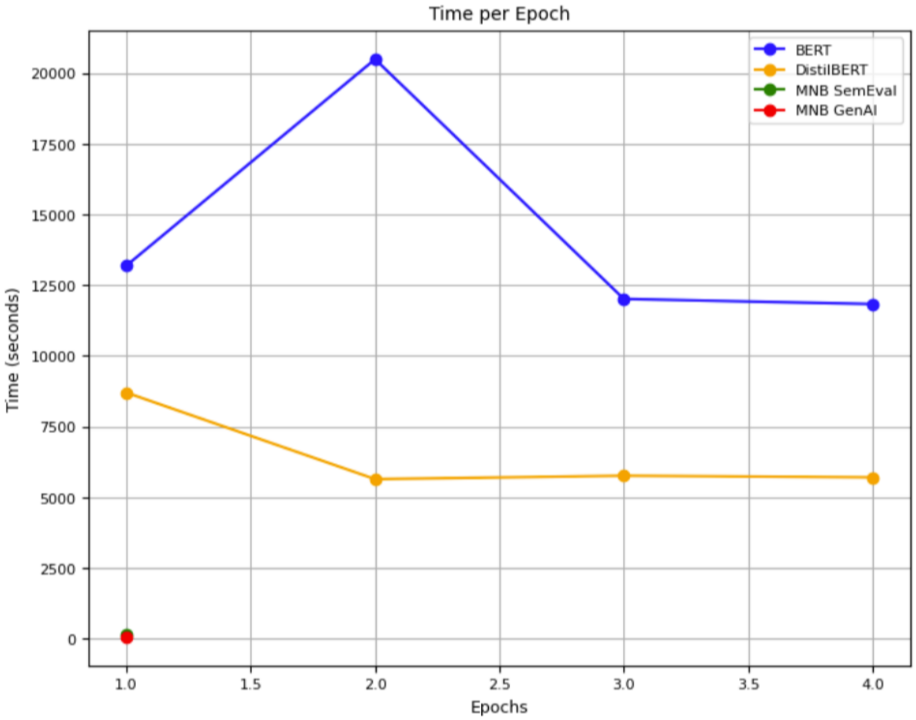

Figure 8: Training time per epoch