BERT for SMS Spam Detection

GitHub - Link

1. Introduction

Large Language Models (LLMs) have revolutionised NLP by enabling tasks such as summarisation, translation, and classification. Among these, BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2018), employs a transformer-based architecture to capture context from both directions in text. This project focuses on fine-tuning BERT for SMS spam detection, a task critical for safeguarding users from fraud and unsolicited messages. While spam detection is well-studied, modern LLMs deliver significant improvements over traditional methods.

2. Background and Literature

Earlier spam detection relied on techniques like Naïve Bayes and SVMs, which struggled with context and linguistic nuances (Meyer et al., 2004). BERT, pre-trained on massive corpora with masked language modelling, allows fine-tuning for specific tasks with high accuracy (Devlin et al., 2018). Hugging Face's Transformers library (Wolf et al., 2020) has simplified such work, enabling rapid experimentation. Studies highlight that cleaning text, balancing classes, and tuning thresholds are vital for handling imbalanced datasets, a key issue in spam classification (Liu et al., 2019).

3. Data and Pre-processing

This work used the SMS Spam Collection Dataset from Kaggle, which contains 5,572 labelled messages. Although not new, it remains a widely accepted benchmark for text classification tasks. Figure 1 shows the original class imbalance, where spam comprises around 13% of messages.

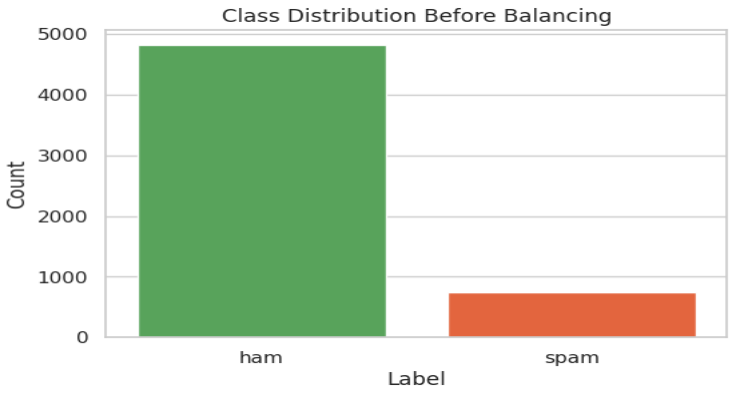

Figure 1. Class distribution before balancing.

To address this, spam messages were oversampled so that ham and spam classes were equally represented, as shown in Figure 2.

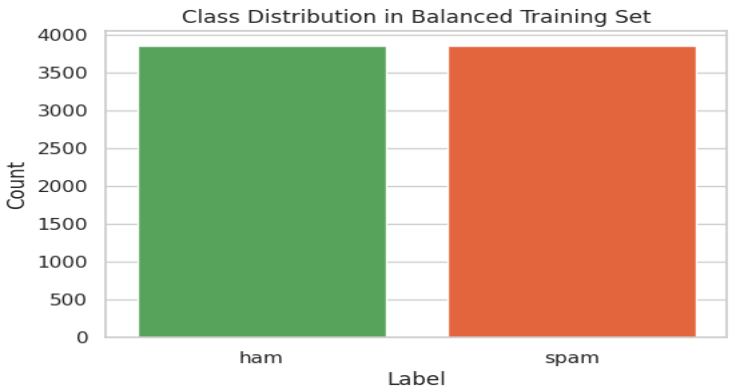

Figure 2. Class distribution after balancing.

Pre-processing involved several steps: converting text to lowercase, removing non-alphabetic characters, applying lemmatization via spaCy, removing stop words, and truncating any message exceeding 512 characters. These steps reduced noise and improved the efficiency of the BERT tokenizer. Figure 3 illustrates the distribution of message lengths after cleaning.

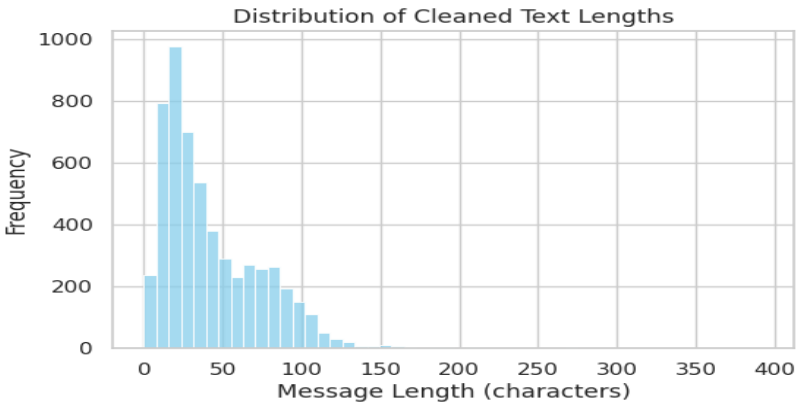

Figure 3. Distribution of cleaned text lengths.

4. Model and Training

The BERT-base-uncased model was fine-tuned using Hugging Face's library. The model used the WordPiece tokenizer and a maximum sequence length of 64 tokens. Training employed the AdamW optimiser with a learning rate of 2e-5, and class imbalance was addressed via weighted CrossEntropyLoss, assigning higher weight (3.0) to the spam class to improve its recall. The model was trained for three epochs with validation checks after each epoch to mitigate overfitting.

5. Results

5.1 Metrics

On the validation set:

Metric	Ham	Spam
Precision	0.99	0.98
Recall	1.00	0.91
F1-score	0.99	0.94

Overall accuracy was **98.6%**.

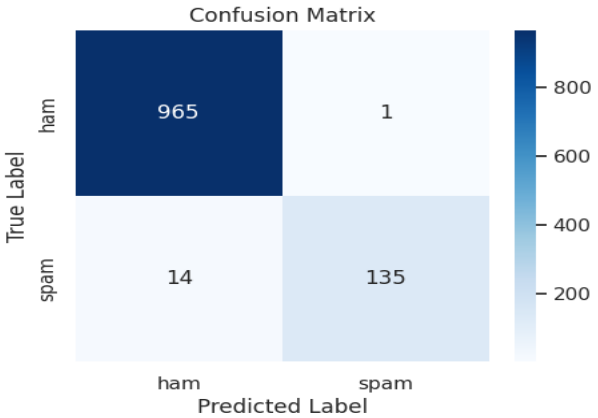

Figure 4. Confusion matrix for the validation set.

The confusion matrix indicates very few spam messages were misclassified as ham, a typical challenge in this task.

5.2 Learning Curves

Loss decreased steadily, while validation accuracy remained stable across epochs.

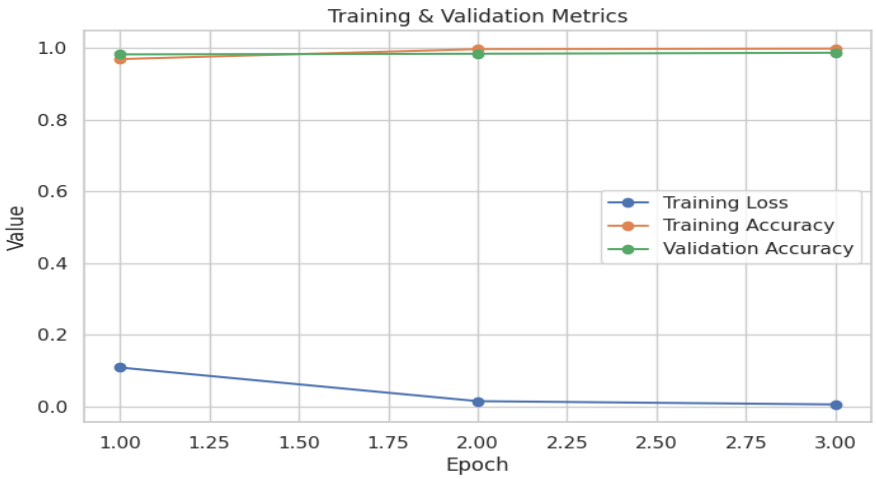

Figure 5. Training and validation loss and accuracy over epochs.

5.3 Sample Predictions

Examples from the validation set include "ya told she ask wat matter," correctly predicted as ham with a confidence of 0.9991, and "reason team budget available buy unsold player base rate," which was misclassified as spam with a lower confidence of 0.6663. Although most predictions were correct, occasional false positives highlight the challenge of ambiguous or context-dependent text.

6. Discussion

The fine-tuned BERT model demonstrated excellent accuracy and effective spam detection, benefiting from comprehensive pre-processing and class balancing. Hugging Face's tools significantly streamlined experimentation and deployment. However, the dataset remains relatively small and may not fully reflect modern spam characteristics, such as the presence of emojis, URLs, or multimedia content. Some semantic overlap between ham and spam messages caused occasional misclassifications. Future work could explore larger and more diverse datasets, alternative models like RoBERTa, or deploying the model in production systems for real-time filtering.

7. Conclusion

This project shows that fine-tuning BERT on SMS data is a highly effective approach for spam detection, achieving strong performance metrics. It underscores the benefits of transfer learning, text pre-processing, and handling class imbalance. Modern LLMs like BERT hold significant promise for enhancing digital security and real-world text classification applications.

8. References

- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. Available at: https://arxiv.org/abs/1810.04805
- Liu, Y., Ott, M., Goyal, N. et al. (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint. Available at: https://arxiv.org/abs/1907.11692
- Wolf, T., Debut, L., Sanh, V. et al. (2020) 'Transformers: State-of-the-art natural language processing', Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). Available at: https://aclanthology.org/2020.emnlp-demos.6