**Sentiment Analysis of Google Play reviews using LLM BERT and HuggingFace**

# 1. Introduction and Literature Review

Sentiment analysis is a subsidiary field of Natural Language Processing (NLP),which handles the identification and classification of subjective text information. It is extensively utilized in customer opinion analysis, brand monitoring, and public sentiment monitoring. With app stores and social networking sites witnessing rapid growth in user-generated content, the need for proper sentiment classification models has increased immensely.

Classic sentiment analysis techniques have utilized machine learning algorithms like Naive Bayes and Support Vector Machines, which tended to leverage hand-crafted features or bag-of-words models. Helpful in some contexts, they generally do not appreciate fine-grained linguistic structure, thus being poorly suited to rich linguistic structures like sarcasm, ambiguity, or specialist jargon (Hoang, et al., 2019).

Recent improvements in deep learning, especially transformer models, have achieved noteworthy performance improvements in sentiment classification. Perhaps some of the most popular models include Devlin's BERT (Bidirectional Encoder Representations from Transformers). BERT sets bidirectional training and uses masked language modeling to learn context more effectively than all its earlier predecessors. Fine-tuned on a task, BERT has established the state-of-the-art in text classification, question answering, and numerous other NLP tasks.

The transformers library by Hugging Face has played a key role in the making of BERT generally accessible for applied research and development, through pre-trained models and easy fine-tuning tools. It has been proven through experiments that BERT is effective for sentiment classification across domains and languages (Chiorrini, et al.,2021).

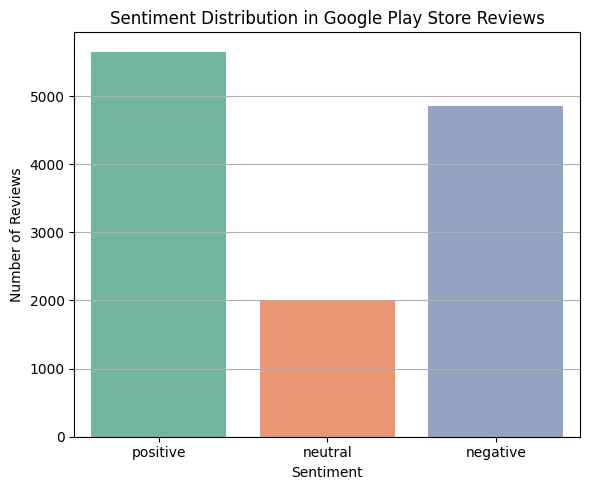
This paper presents a fine-tuned BERT model for sentiment classification in Google Play Store reviews. The aim is to categorize the reviews into positive, neutral, or negative sentiment classes and measure the model's performance based on standard classification measures and evaluation.

# 2. Methodology

**Dataset Overview and Labeling**

The dataset that has been taken from Kaggle, are Google Play Store user reviews with 11 columns of metadata like user ID, rating (score), and review text (content). For this sentiment analysis, only 2 columns score and content(review) were kept. Ratings were reclassified into following three sentiment classes: 1–2 stars → Negative ,3 stars → Neutral and 4–5 stars → Positive.

Following the process of cleaning and filtering, the final dataset contained 12,495 reviews. Class distribution, as shown in Figure 1, shows a fairly balanced dataset with a positive review dominance.



**Figure 1: Distribution of review labels in dataset**

**Model Architecture and Training Setup**

The architecture used in the current work is from the bert-base-uncased version of BERT (Bidirectional Encoder Representations from Transformers), which is a pre-trained language model. BERT is a transformer-based architecture, which is able to learn bidirectional contextual representations of text by jointly considering both left and right context in pre-training. The base model has 12 layers of encoder with self-attention mechanisms and a hidden dimension of 768, with a total of about 110 million parameters (Hoang, et al., 2019).

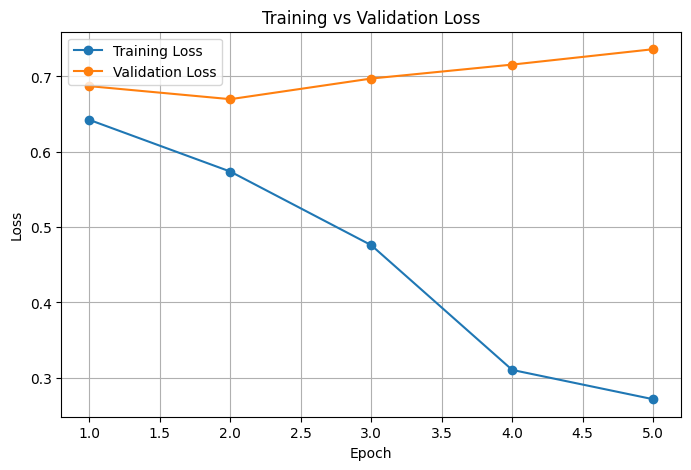
To adapt BERT for sentiment classification, a dense classification head was added to the final hidden state of the [CLS] token. This output layer projects the contextual embedding into three dimensions, corresponding to the sentiment classes: positive, neutral, and negative. The model outputs raw scores (logits), which are converted into probabilities using a softmax function.

Text inputs were tokenized using the Hugging Face AutoTokenizer, which performs lowercasing, WordPiece tokenization, and appends special tokens such as [CLS] and [SEP]. Each input was padded or truncated to 128 tokens, and attention masks were generated accordingly (Huggingface, 2025).

Fine-tuning was performed using the Hugging Face Trainer API with the AdamW optimizer, a learning rate of 2e-5, batch size of 8, and five training epochs. Cross-entropy loss was used as the objective function. The dataset was split into 80% training and 20% validation. Model performance was evaluated using accuracy, precision, recall, and F1-score, with the best model saved based on validation F1. Training was conducted on Google Colab using GPU acceleration to enhance efficiency.

# 3. Results and conclusion

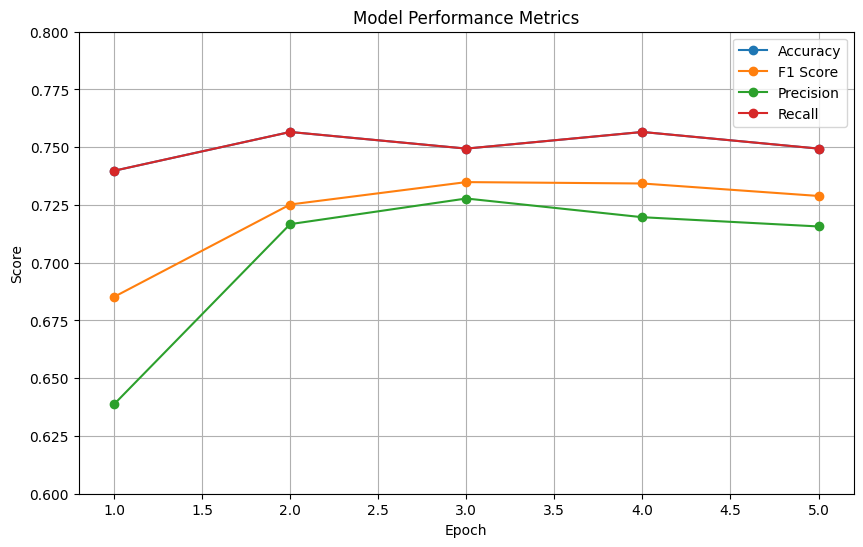
The fine-tuned BERT model demonstrated stable and consistent performance throughout the five training epochs. Training and validation losses exhibited divergent trends, as shown in Figure 2. While the training loss decreased steadily from 0.64 to 0.27, the validation loss fluctuated slightly, rising from 0.68 to 0.73 across the epochs. This trend indicates that the model was effectively learning from the training data, although mild overfitting began to appear in later epochs. However, despite this overfitting, evaluation metrics on the validation set remained relatively stable and high.



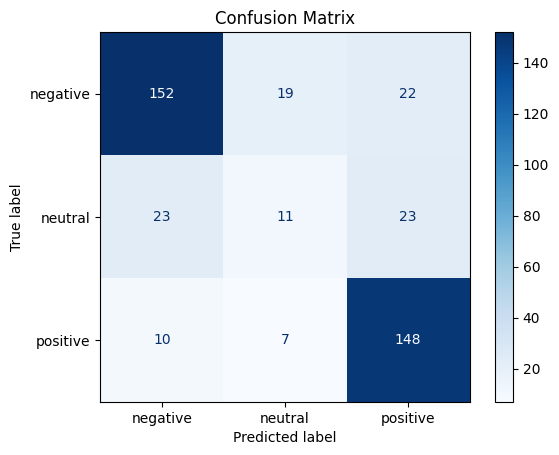
**Figure 2: Training vs validation loss per epoch**

Model performance was evaluated using accuracy, F1-score, precision, and recall. The accuracy peaked at 75.6% in epochs 2 and 4, while the F1-score reached a maximum of 0.7349 in epoch 3. As shown in Figure 3, precision and recall also followed a similar pattern, stabilizing above 0.71 after epoch 2. These results indicate that the model was not only accurate but also balanced in terms of minimizing both false positives and false negatives across the three sentiment classes (Chiorrini, et al.,2021)

The confusion matrix in Figure 4 provides insight into the classification behavior. The model correctly predicted the majority of both positive and negative sentiments. However, it struggled with neutral reviews, often misclassifying them as either positive or negative. This is to be expected due to the subjectivity and vagueness inherent in common expressions of neutral sentiment as well as the lesser representation of neutral samples in the dataset.



**Figure 3: Model performance matrix**



**Figure 4: Confusion matrix of the model perforamnce**

The last evaluation resulted in an accuracy of 74.9%, F1-score of 0.7349, precision of 0.7278, and recall of 0.7494. From these metrics, it can be seen that the fine-tuned BERT model performs well in doing multiclass sentiment classification on user reviews. Upon saving the model, the model was employed for new unseen reviews and model could predict the sentiment of all sample reviews accurately (Hoang, et al., 2019).

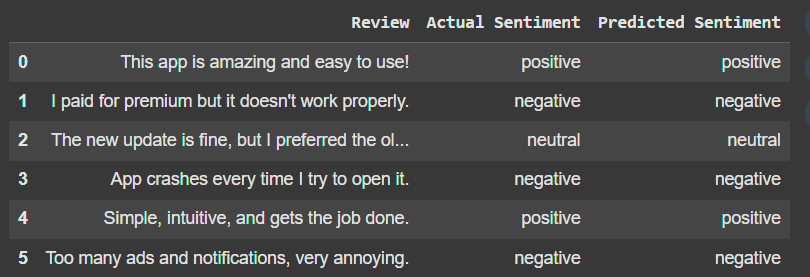


Figure 5: Predicted sentiment for unseen data

Future enhancements could involve utilizing a larger, more diversified dataset and testing other transformer architectures like RoBERTa or DistilBERT to improve generalization and lower overfitting

# References

Hoang, M., Bihorac, O.A. and Rouces, J., 2019. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics* (pp. 187-196).

Chiorrini, A., Diamantini, C., Mircoli, A. and Potena, D., 2021, March. Emotion and sentiment analysis of tweets using BERT. In *Edbt/icdt workshops* (Vol. 3, pp. 1-7).

Huggingface, 2025 Accessed online using [[Fine-tuning a masked language model - Hugging Face LLM Course](https://huggingface.co/learn/llm-course/chapter7/3?fw=pt)]