Project Report: Sentiment Classification Using BERT

# Introduction

This project focuses on sentiment classification, which is a Natural Language Processing (NLP) task where the goal is to determine whether a given piece of text (in this case, movie reviews) expresses a positive or negative sentiment. The classification is carried out using BERT (Bidirectional Encoder Representations from Transformers), a powerful language model developed by Google.

The strength of BERT lies in its ability to understand the context of words from both directions—left-to-right and right-to-left—simultaneously, which helps it to grasp the full meaning of sentences better than traditional models.

# Project Objective

The main aim of this project is to:

* Accurately classify IMDB movie reviews as positive or negative
* Use pre-trained BERT and fine-tune it for this binary classification task
* Leverage the power of deep contextual understanding to improve sentiment prediction accuracy

# Dataset Overview

The dataset used is the IMDB movie reviews dataset, which is a well-known benchmark in sentiment analysis. It consists of:

1. 50,000 reviews, equally split between:
2. 25,000 for training
3. 25,000 for testing
4. Each review is labeled as either positive (1) or negative (0)
5. The dataset is balanced, meaning both classes have an equal number of samples
6. Reviews are written in natural language and vary in length, tone, and complexity

# What is BERT?

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based machine learning model for NLP tasks. It was introduced by Google in 2018 and has since become the foundation for many advanced language understanding models.

Key Features of BERT:

* Bidirectional Understanding: Unlike earlier models that read text from left to right or right to left, BERT reads in both directions at the same time, allowing it to capture full context.
* Pre-training and Fine-tuning: BERT is pre-trained on massive datasets like Wikipedia, and can be fine-tuned for specific tasks like sentiment analysis, question answering, etc.
* Transformer Architecture: It is built entirely on the Transformer model, which relies on self-attention mechanisms rather than recurrence (used in RNNs).

For this project, the BERT Base Uncased version was used, which ignores the case of words (treating "Movie" and "movie" the same).

# Workflow and Methodology

1. Data Preprocessing

The raw IMDB reviews are often noisy and inconsistent. To prepare them for analysis:

* + HTML tags (like <br>) are removed
  + All characters are converted to lowercase to ensure uniformity
  + Special characters and numbers are stripped out to focus only on text
  + Texts are cleaned and normalized to remove irrelevant patterns

2. Tokenization

Since BERT does not understand plain text directly, the cleaned text must be tokenized—converted into a form the model can process.

The BERT tokenizer:

* Splits sentences into word pieces
* Adds special tokens such as [CLS] at the start and [SEP] at the end
* Pads shorter sequences and truncates longer ones to ensure all input sequences have the same length (usually 512 tokens)
* Produces attention masks to distinguish real tokens from padding

3. Model Architecture

The model used is a pre-trained BERT model with a classification layer added on top.

* The base BERT model provides deep representations of text
* A fully connected layer outputs the prediction: positive or negative
* The model is trained to minimize binary cross-entropy loss

4. Training

The BERT model is fine-tuned on the training portion of the IMDB dataset.

* A small learning rate is used
* Batch size is chosen to balance speed and memory
* Trained for a few epochs
* A validation set is used to monitor performance

5. Evaluation and Results

After training, the model is evaluated on the test dataset using:

* + Accuracy
  + Precision
  + Recall
  + F1-score
  + Confusion matrices are used to visualize classification performance.

The BERT model generally achieves high accuracy (>90%), significantly outperforming older approaches like Naive Bayes or LSTM models.

# Visualization and Interpretation

* Word Clouds are created for positive and negative reviews to identify commonly used words
* Graphs and plots (like accuracy over epochs) are generated to monitor training performance
* Confusion matrices help identify where the model tends to make mistakes

# Advantages of BERT in This Project

* Contextual Understanding: BERT understands complex patterns like sarcasm and negation better than traditional models
* Transfer Learning: Saves time by reusing learned knowledge
* Minimal Task-Specific Tuning: Only the final layers need to be trained from scratch

# Limitations and Future Work

Limitations:

* Computationally Intensive: BERT models are large and require powerful hardware (GPU/TPU)
* Sequence Length Constraint: Maximum input size is 512 tokens; longer texts must be truncated
* Interpretability: Deep models like BERT are often black boxes and harder to interpret

Future Improvements:

* Use lighter models like DistilBERT or TinyBERT for faster inference
* Extend to multi-class sentiment classification (e.g., very negative to very positive)
* Implement attention visualization tools to better understand the decision-making process