Transformer-based Text Classification: A Comprehensive Analysis

Authors:
Biswadip Bhattacharyya
Malika Hafiza Pasha

Agenda

- Introduction
- Problem Statement
- Dataset Overview
- Transformer Model Architecture
- Training Process
- Evaluation Metrics
- Visualizations
- Conclusion and Future Work

Introduction

TEXT CLASSIFICATION PLAYS A CRUCIAL ROLE IN NATURAL LANGUAGE PROCESSING. IN THIS PROJECT, WE FOCUS ON CLASSIFYING TEXT WITH A SPECIFIC EMPHASIS ON HATE SPEECH DETECTION USING A TRANSFORMER-BASED ARCHITECTURE.

THE MOTIVATION IS TO AUTOMATE THE DETECTION OF HARMFUL CONTENT ON PLATFORMS.

Problem Statement

- The goal of this project is to build a robust model to classify text into categories (e.g., hate speech vs. non-hate speech).
- We aim to leverage the power of transformers to accurately detect harmful content in user-generated text.

Dataset Overview

Number of records: 31925

Number of classes: Hate Speech, Non-Hate Speech

Features:

Text content

Preprocessing steps:

Tokenization, Padding, Tensor Conversion

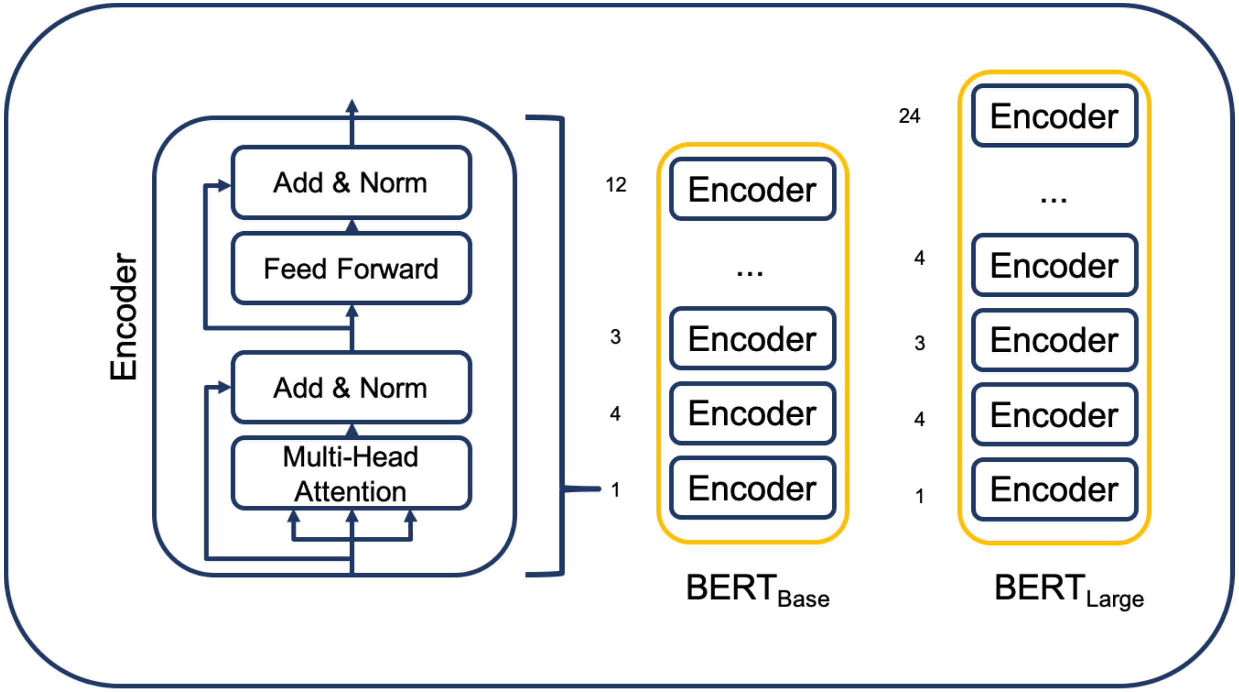
Transformer Model Architecture

Why Transformers?

- Contextual Understanding of Language
- Transfer Learning with Pretrained Models

Model Structure:

- Pretrained Model: BERT (Bidirectional Encoder Representations from Transformers)
- Fine-tuning the last few layers for classification
- Input: Tokenized sentences, Output: Class probabilities



Training Process

Training Strategy:

- - Optimizer: AdamW
- - Learning Rate: 1e-5
- - Epochs: 3
- - Batch Size: 16
- - Loss Function: Cross-Entropy Loss
- - Hardware: GPU for faster training

Hyperparameter Tuning:

- - Adjusted batch size and learning rate for optimal performance

Evaluation Metrics

Results Summary →

Accuracy: 99.28%

Precision: 98.59%

Recall: 100%

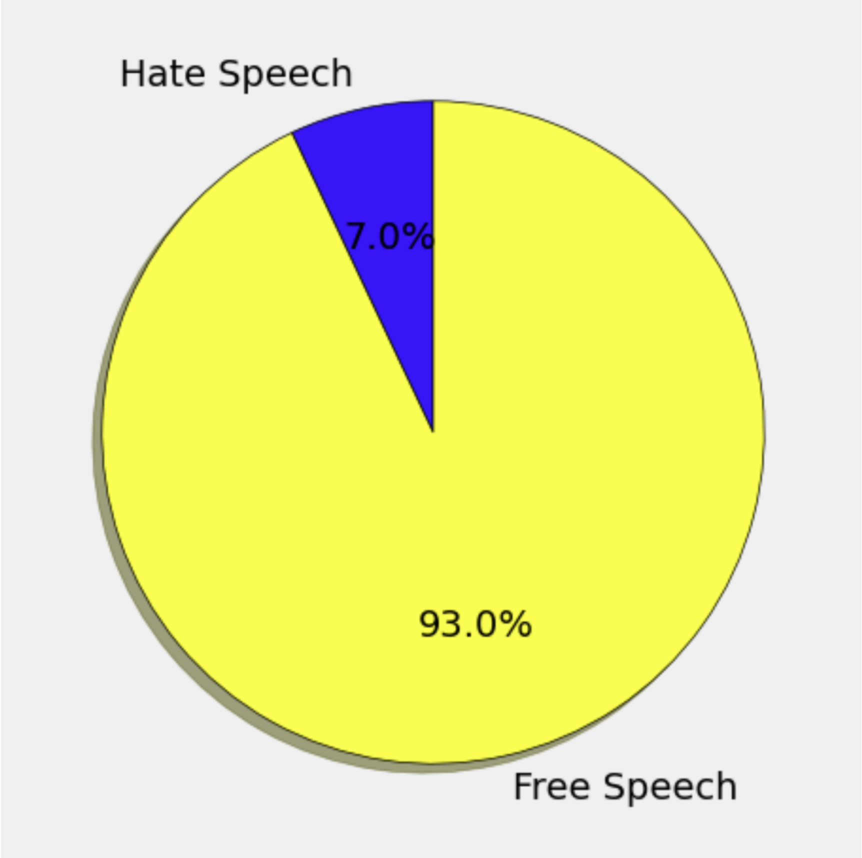
F1 Score: 99.29%

Confusion Matrix: Provides insight into false positives and false negatives.

Visualization

Graphs/Charts:

- Training and Validation Accuracy
- Training and Validation Loss over epochs
- Precision-Recall Curve



Conclusion & Future Work

- The transformer-based model achieved high accuracy in text classification tasks, particularly for detecting hate speech.

- However, there are areas for improvement:

- Dataset Diversity
- Training Time Reduction
- Deployment on large-scale platforms.