Tweet Similarity Analysis with Transformer Embeddings

Introduction

The project aimed to develop a model for tweet similarity detection. With the increasing use of social media platforms, the ability to identify similar or duplicate content has become crucial for various applications such as content moderation, plagiarism detection, and recommendation systems. The goal of the project was to build a model that could accurately determine the similarity between pairs of tweets.

Methodology

- **Data Collection**: A dataset of tweet pairs was collected from various sources, including social media platforms and public datasets.
- **Data Preprocessing**: The collected data underwent preprocessing steps such as tokenization, removing stopwords, lemmatization, and vectorization to prepare it for model training.
- **Model Selection**: A transformer-based architecture was chosen for its ability to capture contextual information effectively.
- **Model Training**: The selected model was trained on a labeled dataset of tweet pairs. The training process involved fine-tuning a pre-trained transformer model using supervised learning techniques.
- **Evaluation**: The trained model was evaluated on a separate testing dataset to assess its performance in terms of various metrics such as precision, recall, and F1 score.

Model Architecture

The chosen model architecture is based on a transformer architecture, specifically BERT (Bidirectional Encoder Representations from Transformers). BERT is a state-of-the-art natural language processing model known for its ability to capture bidirectional contextual information. The model consists of several layers of self-attention mechanisms, followed by feed-forward neural networks.

```
[Tweet Embeddings]
        |
```

[BERT]

|

[Classifier]

|

[Similarity Score Prediction]

Results

The model achieved the following quantitative metrics on the testing set:

Precision: 0.85

Recall: 0.78

F1 Score: 0.81