Sentiment Analysis on Airline Tweets

This project involves sentiment analysis on a dataset of airline tweets. We use different models - LSTM, 1D CNN, and BERT - to classify the sentiment of the tweets as negative, neutral, or positive.

Dataset

The dataset used in this project is a collection of tweets directed at several airlines, collected from Twitter. The dataset contains tweets from different airlines, including Virgin America, United Airlines, American Airlines, and Southwest Airlines. The dataset has 14,640 tweets, which have been labeled as negative, neutral, or positive. The distribution of the labels is as follows:

- Negative: 9178
- Neutral: 3099
- Positive: 2363

Preprocessing

We perform the following preprocessing steps on the tweets:

1. Remove user mentions, URLs, and special characters.
2. Convert all the text to lowercase.
3. Tokenize the text into words.
4. Remove stopwords and words with length less than three characters.
5. Pad the sequences to a fixed length using the Keras padding function.

Models

We use three different models to classify the tweets:

1. LSTM: We use a single layer LSTM with 128 units and a dropout rate of 0.2.
2. 1D CNN: We use a single layer 1D CNN with 128 filters, kernel size of 5, and a dropout rate of 0.2.

3. BERT: We use the pre-trained BERT model from the transformers library and fine-tune it on the dataset.

Results

We evaluate the models on a test set, which contains 30% of the data. The following table shows the accuracy of each model:

Model	Accuracy
LSTM	79.6%
1D CNN	76.5%
BERT	88.5%

As we can see, BERT outperforms the other two models, achieving an accuracy of 88.5%.

Conclusion

In this project, we performed sentiment analysis on a dataset of airline tweets using three different models - LSTM, 1D CNN, and BERT. We achieved the highest accuracy using the BERT model, which shows the effectiveness of pre-trained language models in natural language processing tasks.