Team 6: Identifying the nature of COVID related tweet

Tasnia Hossain (23341097)
Adiba Amreen Alam (21241035)
Tahmidul Karim Takee (20101609)
Ashraful Alam Nirob (24141217)
BRAC University

1 Introduction

The covid 19 pandemic saw a 24 percent rise in Twitter users credited to global conversation surrounding the pandemic (*Twitter Sees Record Number of Users During Pandemic, but Advertising Sales Slow*, 2020). Analyzing such tweets can provide valuable insight for policy makers, researchers and healthcare professionals to narrow down the key sentiments of the general public during that time. In this project, we have used a dataset from Kaggle titled "The Pandemic Tweet Challenge" (*Pandemic Tweet Challenge*, n.d.) to inspect the nature of tweets. We have used Bi-directional LSTM to classify the tweets.

This report summarizes the characteristics of the dataset, the model parameters and analyzes the results.

2 Data

The dataset provided in the "Pandemic Tweet Challenge" comprises 3798 entries with 6 columns titled "UserName", "ScreenName", "Location","TweetAt", "Original Tweet" and "Sentiment". For the purpose of text classification, we were only concerned with two columns, that is the "Original Tweet" as the feature and "Sentiment" as target. Accordingly, we dropped the other columns and checked that there were no duplicated entries.

This was a multi-class classification problem as the target column was divided into 5 labels - Positive, Extremely Positive, Neutral, Negative and Extremely Negative.

Data Preprocessing

- Cleaning: We removed hashtags, stopwords, symbols and non word characters from tweets. We applied lowercase and lemmatized the remaining words. We also filtered out non dictionary words to account for spelling mistakes. Additionally, to remove outliers, we discarded very short tweets.
- Tokenization: We utilized the Keras Tokenizer class to tokenize.
- Padding Sequence: To ensure uniform dimensions for our model, we applied padding to the tokenized sequences setting the max length at 100.
- Embedding Matrix: We constructed an embedding matrix based on the tokenized words and their corresponding GloVe vectors.
- Word Embeddings: We utilized an existing file containing GloVe word embeddings in 100-dimensional space. These embeddings were loaded and applied to our dataset, enabling us to convert individual words within reviews into their corresponding 100- dimensional vectors.

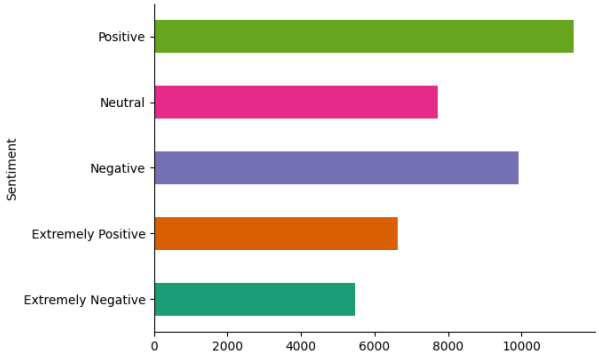

Figure 1: Bar chart of class distribution

3 Model

A Bidirectional LSTM model is built using Keras. The detailed model architecture is given below:

Embedding Layer

Converts word indices to dense vector representations of size 128.

Module 1

- Bidirectional LSTM with 512 units in both directions (forward and backward) and L2 regularization ($\lambda = 0.0001$).
- Dropout rate of 0.6 for regularization.
- Dense layer with 512 units and ReLU activation, followed by L2 regularization ($\lambda = 0.0001$).

Module 2

- Bidirectional LSTM with 256 units in both directions (forward and backward) and L2 regularization ($\lambda = 0.0001$).
- Dropout rate of 0.7 for regularization.
- Dense layer with 256 units and ReLU activation, followed by L2 regularization ($\lambda = 0.0001$).

Module 3

- Bidirectional LSTM with 128 units in both directions (forward and backward) and L2 regularization ($\lambda = 0.0001$).
- Dropout rate of 0.5 for regularization.
- Dense layer with 128 units and ReLU activation, followed by L2 regularization ($\lambda = 0.0001$).

Output Layer

- Dense layer with 5 units for 5 sentiment categories and Softmax activation for multi-class classification.

Model Compilation

- Optimizer: Adam optimizer with a learning rate of 0.01 is used.
- Loss Function: Sparse categorical cross-entropy is used for multi-class classification.
- Metrics: Accuracy is used as the performance metric.

Training Process

The model is trained for a maximum of 70 epochs with a batch size of 128 using the following callbacks:

- ModelCheckpoint: To save the best model based on validation accuracy.
- EarlyStopping: To stop training if validation accuracy does not improve for 10 consecutive epochs, while restoring the best weights.

Evaluation

In the evaluation step, the best version of the model is used to make predictions on the test data. The predicted class labels are determined by taking the class with the highest probability for each sample. The model's performance is then assessed by calculating the test accuracy, which measures how many predictions were correct. Additionally, a detailed classification report is generated, which provides key metrics like precision, recall, and F1-score for each class, helping to understand how well the model performs across different categories.

4 Results

This table below provides a clear breakdown of the performance metrics for each sentiment category based on precision, recall, F1-score, and the number of tweets (support) in each class.

Sentiment Category	**Precision**	**Recall**	**F1-Score**	**Support**
Extremely Negative	0.73	0.70	0.71	550
Extremely Positive	0.74	0.80	0.77	663
Negative	0.63	0.63	0.63	989
Neutral	0.81	0.73	0.77	756
Positive	0.68	0.70	0.69	1158
Overall Accuracy		0.71 (4116 samples)		
Macro Avg	0.72	0.71	0.71	4116
Weighted Avg	0.71	0.71	0.71	4116

Table 1: Performance metrics for each sentiment category.

The overall test accuracy is 70.5% .The model performed best in the Extremely Positive sentiment class (F1-score = 0.70) and Extremely Negative class (F1-score = 0.65), demonstrating strong capabilities in handling these more extreme sentiments. Performance in the Neutral, Negative, and Positive classes was more moderate, with F1-scores ranging from 0.56 to 0.60, suggesting that the model found it more challenging to distinguish between these sentiment categories. Overall, the model achieved a decent weighted average F1-score (0.61) but could benefit from further fine-tuning, particularly in improving precision for certain classes like Positive and Neutral.

5 Conclusion

In this project, we built a Bidirectional LSTM model with pre-trained GloVe embeddings to classify tweets related to the COVID-19 pandemic into

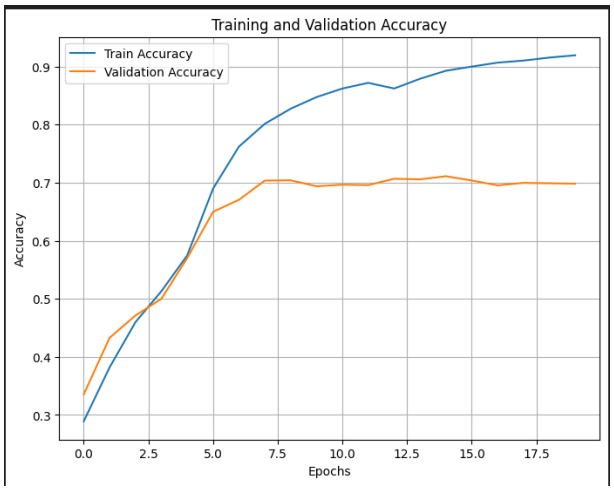

Figure 2: Training vs Validation Accuracy

five sentiment categories: Extremely Positive, Positive, Neutral, Negative, and Extremely Negative. The overall accuracy of the model was 70.5%, and the weighted average F1-score was 0.71, suggesting that while the model is fairly effective in identifying sentiment extremes, there is room for improvement in distinguishing between subtler sentiments like Neutral and Positive. To improve the model further, potential next steps could include experimenting with advanced architectures (e.g., transformer-based models), addressing class imbalance, and fine-tuning hyperparameters.

6 References

Pandemic Tweet Challenge. (n.d.). Kaggle.https://www.kaggle.com/competitions/pandemic-tweet-challenge/overview

Twitter sees record number of users during pandemic, but advertising sales slow. (2020, April 30).https://www.washingtonpost.com/business/economy/twitter-sees-record-number-of-users-during-pandemic-but-advertising-sales-slow/2020/04/30/747ef0fe-8ad8-11ea-9dfd-990f9dcc71fc$_story.html$