Yi Jen Lin 林宜蓁 110006261
Data Mining: Concepts, Techniques, and Applications

Data Mining Lab 2 Report

Data Preprocessing
The dataset consisted of tweets and their corresponding emotions with identification tags. The pre-processing phase started off with the loading of JSON and CSV files, normalizing, and mapping tweet IDs to corresponding emotions and data identification labels. Further, the whole dataset was divided into two subsets: training and test. The text data has been preprocessed using a pipeline that includes lowercasing, removing non-alphabetic characters, tokenization, stopword removal, and lemmatization to ensure a clean input for the model. To speed up preprocessing, the use of Joblib parallelization was exploited, with NLTK for stopword and lemmatization resources. A CountVectorizer has also been used to analyze the most frequent terms, which was useful for understanding the data distribution.

Exploratory Data Analysis (EDA)
The distribution of emotions in the training set was visualized as percentages, showing the imbalance in the dataset. For example, some emotions, such as "joy" and "sadness," were more represented compared to others. These insights helped in understanding potential biases the model might encounter during training.

Feature Engineering
Some important feature engineering steps for preparing the text data included tokenization and padding. Further preprocessing parameters set up included a vocabulary size of 5000, a maximum sequence length of 100, and an out-of-vocabulary token to handle unseen terms during inference. These are then transformed into padded sequences for training, validation, and test datasets to maintain uniformity in the length of input.

Model Architecture
A Bidirectional LSTM-based architecture was developed to capture the contextual relationships in the text. The model included:

1. An embedding layer (dimension 128) to learn dense vector representations of words.
2. Two Bidirectional LSTM layers (64 and 32 units) for sequential data modeling.
3. Dropout layers (0.3 and 0.2) to mitigate overfitting.
4. A dense layer (64 units) with ReLU activation for feature extraction.
5. A final dense layer with softmax activation for multi-class emotion classification.

Mixed precision training was enabled using TensorFlow's API to improve computational efficiency.

Training and Evaluation
This data was then transformed into TensorFlow datasets with batching and prefetching to optimize the training pipeline better. Early stopping was implemented to stop the training when the validation loss stopped improving to avoid overfitting. The model was trained on three epochs with a batch size of 64 and evaluated on the validation set for accuracy and loss.

Challenges and Insights
Initial experiments were overfitting due to imbalanced data and limited regularization. Higher dropout rates and reducing the number of LSTM units in subsequent layers improved generalization. Lastly, the lack of pre-trained embeddings, such as GloVe or Word2Vec, reduced the performance of this version slightly, although the custom embeddings worked reasonably well.

Results and Submission

The trained model was tested on the test set, and predictions were generated. Results were mapped to their emotion labels and saved as a CSV for submission. In future improvements, it can consider pre-trained embeddings, augmentation for class imbalance, and tuning of hyperparameters.
The final submission was saved as submission.csv.