**Name : Cheryala Nikhil**

**Enrollment No : 21cs002369**

**Internship Batch : 2**

**Detailed Report: Sentiment Analysis Using LSTM on Twitter Data**

**1. Introduction**

This report describes a comprehensive sentiment analysis project, involving exploratory data analysis (EDA), text preprocessing, model building using LSTM (Long Short-Term Memory) neural networks, and model deployment. The objective is to classify tweets as either positive or negative.

### 2. Data Description

The dataset used for this project is a CSV file named training data.csv, containing the following columns:

* **label**: Sentiment of the tweet (0 = negative, 4 = positive).
* **time**: The time the tweet was posted.
* **date**: The date the tweet was posted.
* **query**: The query (if any) that was issued.
* **username**: The username of the person who posted the tweet.
* **text**: The content of the tweet.

### 3. Data Loading and Initial Exploration

The dataset was loaded and inspected to understand its structure and content. Initial exploration included viewing the first and last five records, checking the columns, length, shape, and data types, and identifying any missing values.

### 4. Exploratory Data Analysis (EDA)

#### Label Distribution

EDA revealed the distribution of the labels in the dataset. The dataset contained an equal number of positive (label 4) and negative (label 0) tweets.

#### Text Length Analysis

The length of tweets was analyzed to understand the distribution of text length. This helped in determining the appropriate maximum length for padding sequences during preprocessing.

#### Word Frequency Analysis

Common words and their frequencies were analyzed to understand the most frequently used words in positive and negative tweets. This was visualized using word clouds and bar charts.

### 5. Data Cleaning and Preprocessing

#### Selecting Relevant Columns

Only the text and label columns were selected for analysis. The labels were converted to binary: 4 (positive) became 1, and 0 (negative) remained 0.

#### Balancing the Dataset

An equal number of positive and negative tweets (20,000 each) were selected to avoid bias in the model.

#### Text Preprocessing

The following preprocessing steps were performed on the tweet text:

* Conversion to lowercase
* Removal of stopwords
* Removal of punctuations
* Removal of repeating characters
* Removal of email addresses
* Removal of URLs
* Removal of numeric numbers
* Tokenization of text
* Stemming and lemmatization

### 6. Preparing Data for Model Training

The text data was converted to sequences and padded to a maximum length of 500 words. The dataset was then split into training and test sets.

### 7. Building and Training the Model

An LSTM-based model was built using TensorFlow. The model consisted of the following layers:

* Input layer
* Embedding layer
* LSTM layer
* Fully connected (Dense) layer with ReLU activation
* Dropout layer
* Output layer with sigmoid activation

The model was compiled with binary cross-entropy loss and RMSprop optimizer. It was trained on the training data for six epochs.

### 8. Evaluating the Model

#### Accuracy

The model's accuracy on the test data was evaluated and reported.

#### Confusion Matrix

A confusion matrix was generated to show the number of true positives, true negatives, false positives, and false negatives.

#### ROC Curve

The ROC curve was plotted, and the area under the curve (AUC) was calculated to evaluate the model's performance across different thresholds.

### 9. Model Deployment

The trained model can be deployed using various methods, such as:

* Creating a web service API using Flask or FastAPI
* Deploying the model on cloud platforms like AWS, Azure, or Google Cloud
* Integrating the model into an existing application

### 10. Results

The model achieved a high accuracy on the test data. The confusion matrix and ROC curve demonstrated the model's effectiveness in distinguishing between positive and negative tweets.

### 11. Conclusion

This project successfully built and evaluated a sentiment analysis model using LSTM. The thorough EDA and preprocessing steps contributed to the model's performance. The results indicate that the model is capable of accurately classifying tweets into positive and negative sentiments.

### 12. Future Work

Future improvements could include:

* Experimenting with different model architectures and hyperparameters
* Using pre-trained embeddings such as GloVe or Word2Vec
* Expanding the dataset for better generalization
* Implementing techniques to handle imbalanced datasets more effectively

This concludes the detailed report on the sentiment analysis project using LSTM.