Twitter Sentiment Analysis using EDA and Deep Learning

Student Name:

Roll No:

Department: Artificial Intelligence and Data Science

Course: Exploratory Data Analysis and Visualization (U21ADP05)

Date of Submission:

Abstract (100–150 words)

This project performs exploratory data analysis and sentiment classification on Twitter data. The dataset contains tweets labeled with sentiments (positive, neutral, negative). We perform thorough preprocessing (cleaning text, tokenization, stopword removal, lemmatization), exploratory visualizations (sentiment distribution, wordclouds, tweet-length distribution, hashtag analysis, sentiment by category), and build a deep learning model (Bidirectional LSTM) to classify tweet sentiment. Model training includes hyperparameter tuning and evaluation using accuracy, confusion matrix, ROC/AUC, and loss/accuracy curves. The repository contains the code notebook, dataset link, trained model, and a README describing how to reproduce results. Key findings and suggestions for future improvement are provided.

---

1. Introduction & Objective

Objective: Analyze Twitter data to extract insights and build a DL model to classify sentiment (positive/neutral/negative). The goal is to demonstrate EDA, preprocessing, visualization, modeling, and interpretation.

---

2. Dataset Description

- **Source:** (e.g., Kaggle — Twitter US Airline Sentiment)
- **Format:** CSV
- **Important fields:** `tweet_id`, `text`, `airline_sentiment` (label), `airline`, `tweet_created`, `tweet_location`, `user_timezone`, etc.
- **Size:** (fill after downloading)
- **Number of features:** >= 15 (if using enriched dataset or adding derived features like `text_length`, `num_hashtags`, etc.)

---

3. EDA & Preprocessing

3.1 EDA Steps

- Missing values and duplicates check
- Distribution of sentiment labels
- Tweet length statistics and distribution
- Top words overall and by sentiment (wordclouds and bar plots)
- Top hashtags and mention counts

3.2 Preprocessing Steps

- Lowercasing
- Removing URLs, mentions (@), hashtags (#), punctuation, numbers
- Expand/consolidate emojis or remove them
- Tokenization, stopword removal, lemmatization
- Create features: `text_length`, `num_tokens`, `num_hashtags`, `num_mentions`
- Convert labels to categorical
- Train/validation/test split (e.g., 70/15/15)

---

4. Data Visualization (Minimum 5 visuals)

(Include the generated plots here — each with a short explanation and insight)

1. **Sentiment Distribution** — shows class imbalance and overall polarity.
2. **WordCloud (All tweets)** — common words across dataset.
3. **Top words per sentiment (bar charts)** — reveals distinct vocabulary.
4. **Tweet Length Distribution** — helps set max sequence length for model.
5. **Sentiment by Category (airline or topic)** — shows which categories have more negative tweets.

---

5. Deep Learning Model

5.1 Architecture

- Embedding layer (pretrained embeddings optional)
- Bidirectional LSTM with dropout
- Dense layers with ReLU and final softmax

5.2 Hyperparameters

- `embedding_dim` = 100 (or use GloVe 100d)
- `maxlen` = (set from tweet length distribution)
- `batch_size` = 32

- `epochs` = 8–15
- `optimizer` = Adam

5.3 Training details

- Loss: categorical_crossentropy
- Metrics: accuracy
- Callbacks: ModelCheckpoint, EarlyStopping

---

6. Result Visualization & Interpretation

- **Accuracy vs Epoch** — check for under/overfitting
- **Loss vs Epoch** — training/validation loss behavior
- **Confusion Matrix** — per-class performance
- **ROC & AUC** — multiclass ROC (one-vs-rest)

Interpret the model performance, precision/recall per class, and dominant error types.

---

7. Conclusion & Future Scope

- Summarize achievements and main insights from EDA
- Model performance summary
- Limitations (class imbalance, sarcasm detection, domain shifts)
- Future work: transformer models (BERT), data augmentation, multilingual support

---

8. References (APA/IEEE format)

1. Kaggle dataset — Twitter US Airline Sentiment. (Year). URL
2. Chollet, F. (2018). Deep Learning with Python. (Manning.)
3. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python.

---

9. Appendix — Code

- Link to `twitter_sentiment_analysis.ipynb` in repository

Prepared by: (Your name)