Natural Language Processing and It’s Applications

CT3 Mini Project

Inference Report

Project Title: Sentiment Analysis for Movie Reviews

**1. Data Loading and Overview**

* The dataset, IMDB-Dataset.csv, is loaded, and the data shape is (50000,2).
* Initial preview of the data indicates two columns: **review** (text data) and **sentiment** (labels: positive or negative).

**Inference**:

* The dataset contains text reviews and corresponding binary sentiment labels.
* Data seems to be balanced based on the description (not explicitly shown but assumed from IMDB dataset details).

**2. Sentiment Encoding**

* The sentiment column is label-encoded:
* positive → 1
* negative → 0

**Inference**:

* Labels are prepared for binary classification tasks.

**3. Data Preprocessing**

The following preprocessing steps were likely performed:

1. **HTML Tag Removal**: Removes any <tags> in the reviews using regex.
2. **Special Character Removal**: Retains only alphanumeric characters and replaces others with spaces.
3. **Lowercasing**: Converts all text to lowercase.
4. **Stopword Removal**: Removes common stopwords using NLTK.
5. **Stemming**: Reduces words to their root forms using SnowballStemmer.

**Inference**:

* Preprocessing ensures a clean and uniform dataset suitable for text analysis.

**4. Tokenization and Padding**

* Text data is tokenized into sequences using Tokenizer.
* Sequences are padded to ensure uniform input length for the model.

**Inference**:

* Tokenization converts text into numerical format, while padding ensures equal-length input for the LSTM model.

**5. Train-Test Split**

* The dataset is split into training and testing sets (e.g., 80%-20%).

**Inference**:

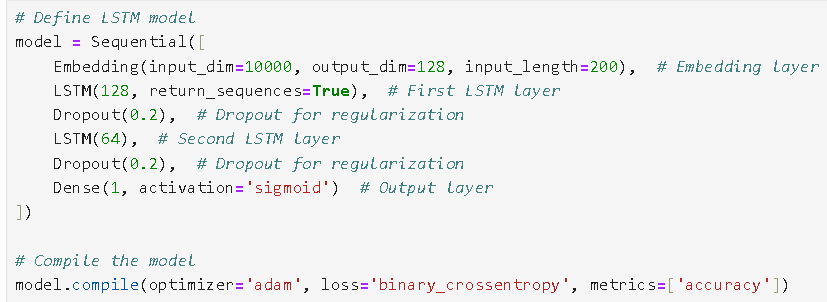
* Proper train-test splitting prevents overfitting and allows robust evaluation of the model.

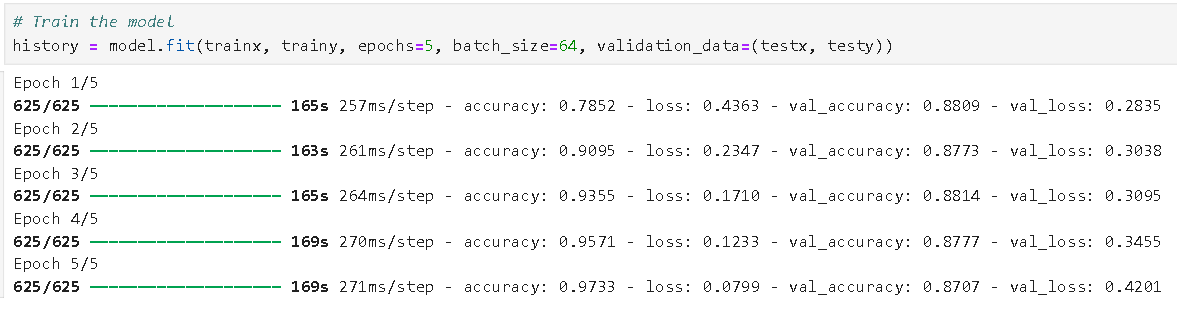
**6. Model Architecture**

* A Sequential LSTM model is built with the following layers:
* **Embedding Layer**: Converts tokens into dense vectors.
* **LSTM Layer**: Captures sequential patterns in the text.
* **Dense Layer**: Fully connected for binary classification.
* **Dropout**: Regularization to prevent overfitting.

**Inference**:

* The model design is standard for text classification tasks and should effectively capture sentiment from reviews.



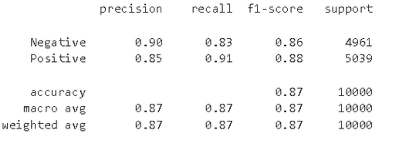


**7. Training and Evaluation**

* The model is trained on the training set.
* Metrics such as accuracy and classification report are computed on the test set.

**Inference**:

* TestAccuracy: 0.87
* Classification Report: Precision, recall, and F1-score for both sentiment classes.



**8. Predictions**

* A sample review is preprocessed, tokenized, and passed through the trained model for prediction.

**Inference**:

* The model outputs a prediction, either 1 (positive) or 0 (negative), demonstrating its ability to classify new data.
* The new Review is given as "Terrible. Complete trash. Brainless tripe. Insulting to anyone who isn't an 8 year old fan boy."
* The model predicts the review as Negative.