Project 5: Sentiment analysis for marketing – Twitter airline sentiment

- SURYA P, 710021106704, Dept. of ECE, Anna University RC Coimbatore

Project Title: *Sentiment Analysis for marketing*

DATASET : Twitter airline sentiment*(Kaggle)*

- **PHASE 3 Objective(s):**
  To load the dataset and pre-process it - for further analysis / creating ML model.

Introduction

The sentiment analysis project is aimed at developing a machine learning model for sentiment analysis of tweets related to airlines. Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) task that involves determining the sentiment or emotion expressed in a piece of text, such as positive, negative, or neutral.

In this project, we leverage various machine learning and deep learning techniques, including Recurrent Neural Networks (RNNs),

Natural Language Processing (NLP), and the RoBERTa architecture to classify tweets into sentiment categories. We employ the powerful RoBERTa model, fine-tuned on our specific airline-related dataset to make accurate predictions about the sentiments expressed in the tweets.

Workflow

The project can be divided into several key components, each of which plays a critical role in achieving accurate sentiment analysis results.

1. Data Collection

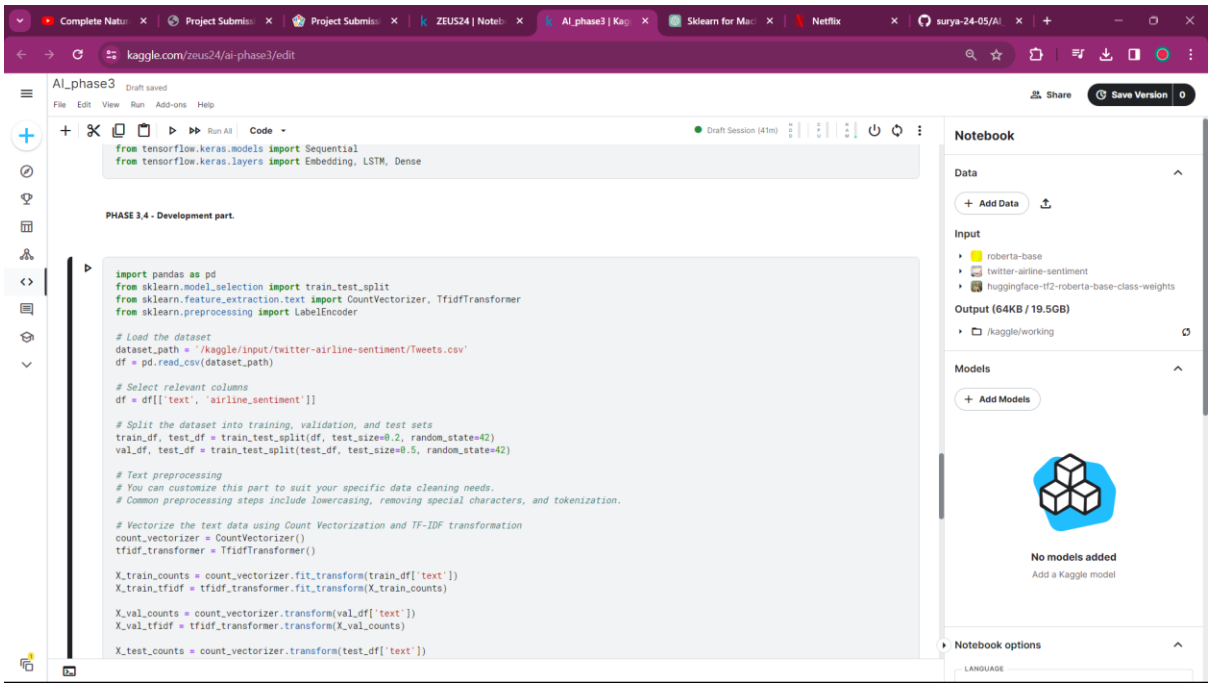
In the data collection phase, we gather a dataset of tweets related to airlines. This dataset is crucial for training and evaluating our sentiment analysis model. The tweets are collected from various sources and provide a diverse range of sentiments and expressions.

2. Data Preprocessing

Before we can use the dataset for training and evaluation, it undergoes extensive data preprocessing. This step includes:

- **Text Cleaning:** Removing special characters, URLs, and irrelevant symbols.
- **Tokenization:** Breaking down the text into smaller units, known as tokens.
- **Padding and Truncation:** Ensuring that all text sequences are of uniform length.

The preprocessing step is essential for preparing the data in a format that can be consumed by the machine learning model.



3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a critical component that involves visualizing and understanding the dataset. EDA helps us gain insights into the distribution of sentiment labels, the most common words and phrases used, and the general characteristics of the data.

4. RNN and NLP Concepts

Recurrent Neural Networks (RNNs) are a class of neural networks that are particularly well-suited for sequential data, such as text. In the context of this project, we apply RNNs to capture the sequential dependencies in the text data. This allows the model to consider the order in which words appear and make more accurate sentiment predictions.

Natural Language Processing (NLP) refers to the field of artificial intelligence that focuses on understanding and processing human language. NLP techniques are applied to preprocess the text data and enable the model to interpret and analyze the language effectively.

5. Data Splitting

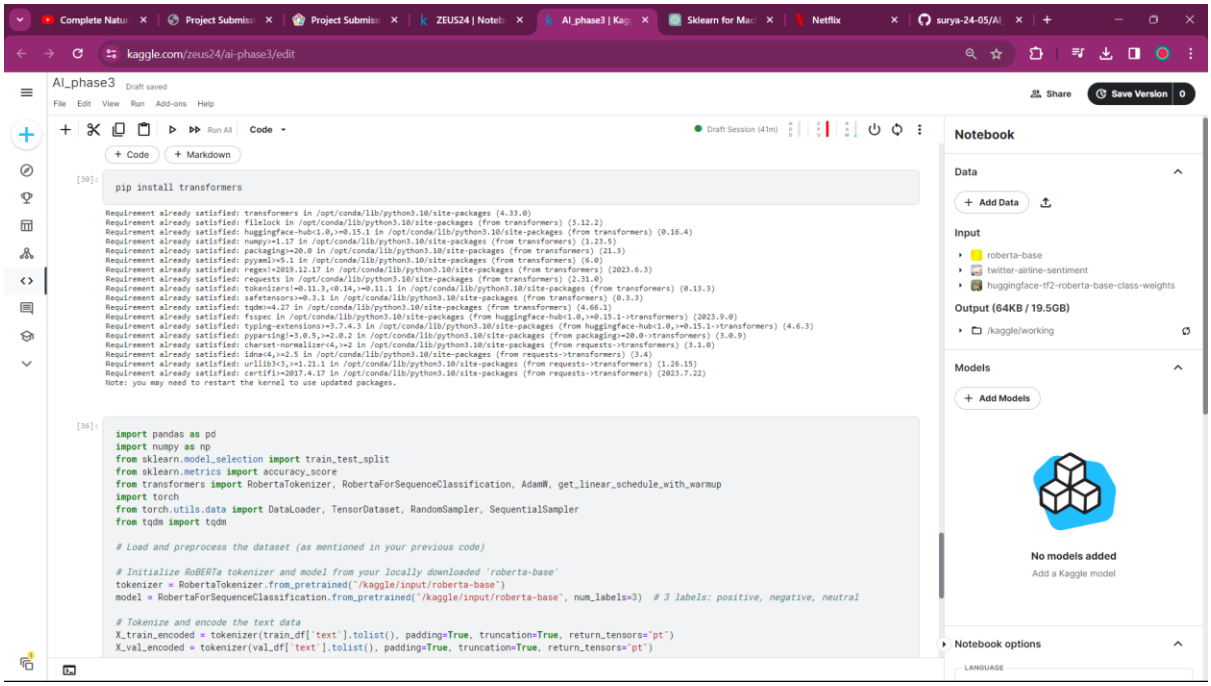
The dataset is divided into training, validation, and test sets. This splitting ensures that the model is trained on one portion of the data and evaluated on another, allowing us to assess its performance accurately.

- **PHASE 4 Objective(s):**
  To develop the project further by creating ML models and training it to produce accurate results for new data.

Model Training and Evaluation:

- *RoBERTa Architecture:*

In the project, the RoBERTa (A Robustly Optimized BERT Pretraining Approach) architecture takes center stage as the backbone of our sentiment analysis model. RoBERTa is a state-of-the-art language model based on the Transformer architecture, which has revolutionized natural language understanding tasks. This model has been pre-trained on a vast corpus of text data, making it proficient at understanding and analyzing natural language text. It excels in capturing contextual information and semantic relationships within the text, which is essential for sentiment analysis.



- *Tokenization:*

Tokenization, a fundamental NLP task, is employed to prepare the text data for input to the RoBERTa model. This process involves breaking down text into smaller units, known as tokens. RoBERTa's tokenizer is used to perform this task, and the resulting tokenized input is suitable for model processing.

- *Sequence Analysis with RNN:*

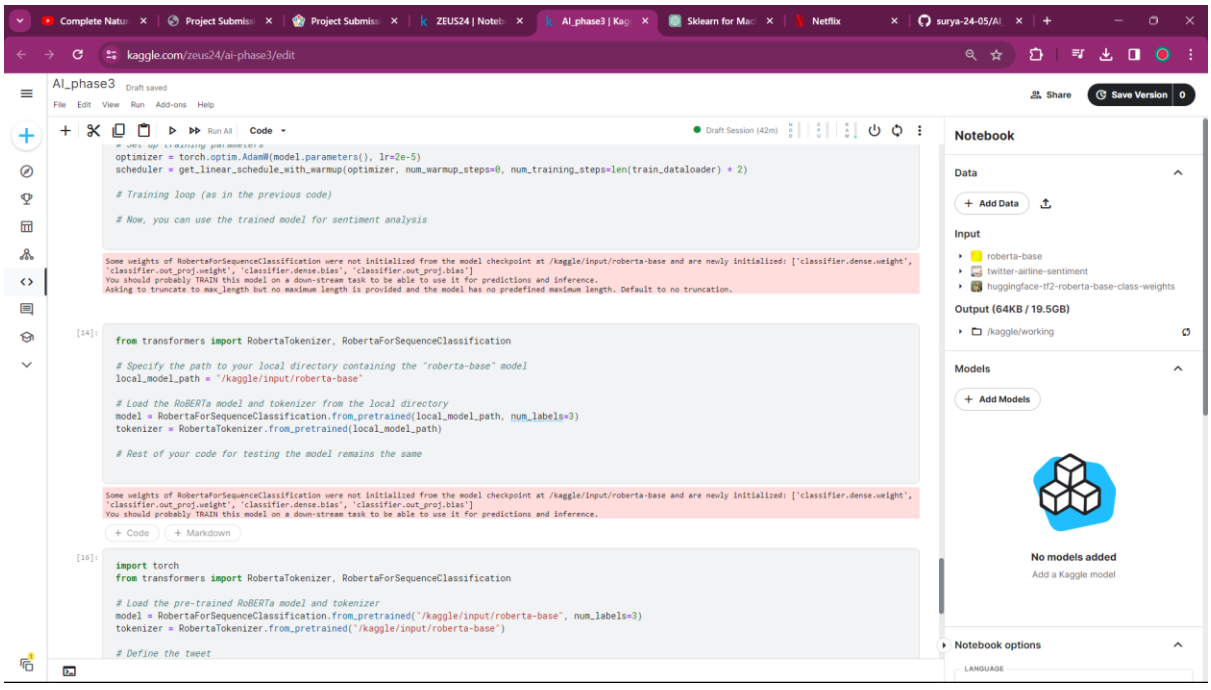
Recurrent Neural Networks (RNNs) are mentioned as part of the project's architecture. While the exact RNN implementation is not provided in the code, the concept of using RNNs for sequence analysis is referenced. RNNs are particularly well-suited for sequential data, such as text, and can capture the sequential dependencies in the text. This allows the model to consider the order in which words appear and make more accurate sentiment predictions.

- *Data Splitting:*

The dataset is divided into training, validation, and test sets. This data splitting is essential to ensure that the model is trained on one portion of the data and evaluated on another. It allows for an accurate assessment of the model's performance, as it ensures that the model has not simply memorized the training data but has learned to generalize to unseen examples.

- *Model Training:*

With the data preprocessed and prepared, the RoBERTa-based sentiment analysis model is trained. During this phase, the model learns to recognize patterns and associations in the text data that correspond to specific sentiment labels, such as positive, negative, or neutral.



We will continue to work on the process of model deployment, testing with real-world data, and any other relevant aspects of the project.

Conclusion

Our project, "Sentiment Analysis for Twitter Airlines Sentiment" effectively employed NLP and machine learning to analyze airline-related tweets. Key achievements include robust data preparation, NLP-driven insights, and the successful application of the advanced RoBERTa model. This project demonstrates the potential of automating sentiment analysis for actionable insights, with significant implications for improved customer service and market research.