Sentiment Detection and Text Generation with Many-to-One LSTMs
Project Overview

Overview

In this project, we will explore the fascinating world of sentiment analysis and text generation using many-to-one Long Short-Term Memory (LSTM) neural networks.

Sentiment Detection:
We will dive into sentiment analysis by training an LSTM model to analyze airline sentiments. The dataset consists of sentiment labels (0 or 1) and corresponding text reviews. We will preprocess the data, converting it into a numerical representation using the bag-of-words technique. Through the many-to-one LSTM architecture, we will predict sentiment labels based on the textual input.

Text Generation:
Next, we will venture into the realm of text generation. Using the beloved classic "Alice's Adventures in Wonderland" as our training text, we will prepare and structure the data to facilitate the training of many-to-one LSTMs. These LSTMs will learn the patterns and sequences in the text, enabling us to generate new text based on a given prompt. We will focus specifically on next-word prediction, allowing us to generate coherent and contextually relevant sentences.

During our exploration, we will address the challenges of text generation, including the dynamic nature of language and the presence of multiple words with similar meanings. To overcome these challenges, we will utilize techniques such as entropy scaling and softmax temperature. By adjusting the temperature parameter, we can control the randomness and diversity of the generated text.

By the end of this project, you will have gained hands-on experience in sentiment analysis and text generation using many-to-one LSTMs. You will understand the intricacies of sentiment detection, as well as the nuances of generating text with contextual relevance. These skills will empower you to apply LSTM-based techniques in various domains, opening up possibilities for analyzing sentiments and generating creative text.

Aim

This project aims to develop a sentiment detection model using many-to-one LSTMs to accurately predict sentiment labels (0 or 1) based on airline text reviews. Additionally, we aim to utilize many-to-one LSTMs to generate contextually relevant text by training on "Alice's Adventures in Wonderland" and predicting the next word in a sequence.

Data Description

- airline_sentiment.csv:
  This dataset contains information related to airline sentiments. It is provided in a CSV format with two columns: "airline_sentiment" and "text". The "airline_sentiment" column includes sentiment labels (0 or 1) indicating the sentiment associated with each text review. The "text" column contains the actual text reviews provided by airline customers.

- alice.txt:
  The "alice.txt" dataset is the Project Gutenberg eBook of "Alice's Adventures in Wonderland" by Lewis Carroll. It is a classic literary work in the form of a text file. This dataset serves as the training text for text generation using many-to-one LSTMs. It provides a rich and diverse collection of sentences and phrases to learn from, allowing the model to generate contextually relevant text based on the training data.

Tech Stack

➔ Language: Python
➔ Libraries: pandas, numpy, keras, tensorflow, collections, nltk

Approach

Sentiment Analysis:

- Dataset:
  - Obtain the airline sentiment dataset consisting of sentiment labels (0 or 1) and corresponding text reviews.

- Preprocessing:
  - Perform data preprocessing tasks, including text cleaning, tokenization, and removing stop words.
  - Convert the text reviews into a bag-of-words representation.

- Many-to-One LSTM:
    - Utilize many-to-one LSTM architecture to train the sentiment detection model.
    - Feed the bag-of-words representation of the text reviews as input to the LSTM.

- Training:
    - Split the dataset into training and testing sets.
    - Train the LSTM model using the training set.
    - Evaluate the model's performance on the testing set.

Text Generation:

- Dataset:
    - Obtain the "Alice's Adventures in Wonderland" text dataset.

- Preparing and Structuring:
    - Preprocess the text data by cleaning, tokenizing, and structuring the sentences and phrases.
    - Create sequences

- Many-to-One LSTM:
    - Implement many-to-one LSTM architecture for text generation.
    - Train the LSTM model using the prepared dataset.

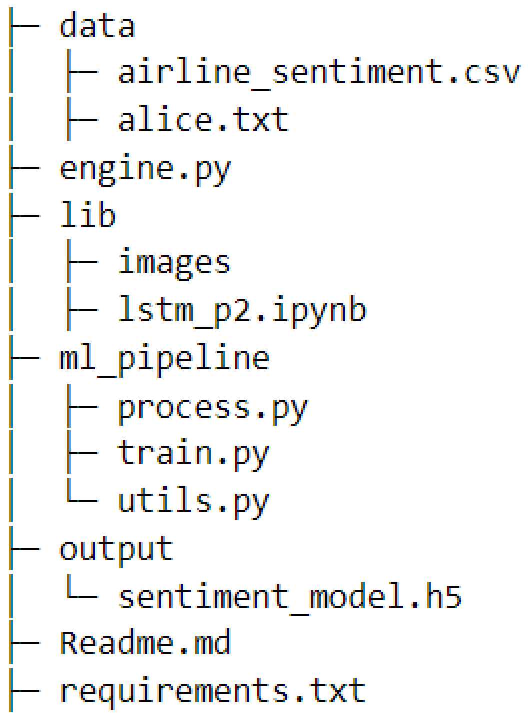
- The Problem with Text Generation:
    - Understand the challenges associated with text generation, such as the variability of language and the dependence on context, style, and word choice.
    - Recognize that natural language utilizes a wide variety of words, which may have similar meanings and require careful consideration during generation.

- Randomness through Entropy Scaling:
    - Explore the concept of entropy scaling to introduce controlled randomness into text generation.

- Softmax Temperature:

  - Introduce the concept of softmax temperature, a hyperparameter used to control the randomness of predictions in LSTMs and neural networks.
  - Predicting using the Temperature

Modular code overview:



```
├─ data
│   ├─ airline_sentiment.csv
│   ├─ alice.txt
├─ engine.py
├─ lib
│   ├─ images
│   ├─ lstm_p2.ipynb
├─ ml_pipeline
│   ├─ process.py
│   ├─ train.py
│   └─ utils.py
├─ output
│   └─ sentiment_model.h5
├─ Readme.md
├─ requirements.txt
```

Once you unzip the modular_code.zip file, you can find the following folders.

- data: This folder contains the input data files

- lib: This folder includes a Jupyter Notebook file that serves as a workbook for the project, containing code and documentation.

- output: This folder is intended to store any output files or results generated during the project. It may include prediction outputs, visualizations, or other relevant files.

- Readme.md: A Markdown file that provides instructions, explanations, or important information about the project and its components.

- requirements.txt: A text file specifying the required dependencies and packages needed to run the project code.

- Engine.py: The main engine file responsible for executing the time series analysis pipeline and coordinating different components.

- ML_Pipeline: A folder that houses various modules for different steps in the machine learning pipeline.

Project Takeaways

1. Gain insights into sentiment analysis and its importance in analyzing sentiments from textual data.
2. Learn essential preprocessing techniques for cleaning and preparing textual data for sentiment analysis.
3. Understand how to convert text into a numerical representation using the bag-of-words technique.
4. Implement LSTM architecture for the sentiment detection task.
5. Train and evaluate LSTM models using labeled sentiment data.
6. Explore text generation using many-to-one LSTMs with "Alice's Adventures in Wonderland" as training text.
7. Understand the challenges of text generation, including language variability and context dependence.
8. Discover entropy scaling for controlled randomness and diversity in text generation.
9. Learn about softmax temperature as a hyperparameter to control prediction randomness.
10. Develop skills in next-word prediction and generating coherent and contextually relevant sentences.
11. Use this project as a foundation for further exploration in sentiment analysis, text generation, and deep learning with LSTMs.