INFO6105 – Data Science Engineering Methods and Tools

Semester- Spring 2024

Title:- A Movie Review Sentiment Analysis Web App

By

Deepak Kumar

Northeastern University

Overview

The scope of this project is to create a web tool and deploy on cloud. The main objective is to analyze the sentiment (positive or negative) of movie reviews by applying machine learning techniques. This analysis helps in understanding public opinion and can be used in various applications like improving movie recommendations.

Dataset:

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided.

Methods

The methodology involved several key steps:

- Data Acquisition: The dataset used is the IMDB movie reviews dataset, which consists of 50,000 reviews split evenly into training and testing sets.
- Data Preparation: The raw text data was preprocessed to remove HTML tags, convert to lowercase, and remove stop words and punctuation. This was followed by tokenization and padding.
- Model Building: A recurrent neural network (RNN) with LSTM (Long Short-Term Memory) units was utilized to perform sentiment analysis.
- Training: The model was trained using the prepared dataset on AWS SageMaker, leveraging its powerful computational capabilities.
- Deployment: The trained model was deployed as a web application using AWS SageMaker and AWS Lambda for real-time sentiment analysis.

Results

The LSTM model achieved promising results, demonstrating high accuracy in classifying the sentiments of movie reviews. The model's performance was evaluated based on its accuracy in predicting the sentiment on the test dataset.

- Model Testing (Model Validation on Test Data)
    - The model's accuracy was evaluated using the test dataset.
    - The resulting accuracy from this testing was 0.8556 (or 85.56%).

- Inference Testing (Practical Application Test)
    - The model was further tested by applying it to actual movie reviews.
    - This practical testing yielded an accuracy of 0.872 (or 87.2%).

Discussion

The project highlighted the effectiveness of LSTM models in handling sequence data like text for sentiment analysis. Challenges encountered included handling the large dataset size and optimizing the model's performance. The choice of AWS SageMaker facilitated efficient handling of these challenges due to its scalable environment.

Conclusion

The sentiment analysis model developed in this project proved to be effective in analyzing movie reviews. Future work could involve refining the model by exploring different neural network architectures or incorporating more nuanced sentiment analysis to capture emotions beyond positive and negative. Additionally, expanding the dataset and including multilingual reviews could broaden the applicability of the model.

Reference

- https://ai.stanford.edu/~amaas/data/sentiment/
- https://pytorch.org/docs/stable/index.html
- https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html
- https://docs.aws.amazon.com/lambda/
- https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html
- https://github.com/udacity/sagemaker-deployment/tree/master/Project
- https://docs.aws.amazon.com/apigateway/latest/developerguide/welcome.html

----------Thank You----------