CSCE 5290- Natural Language Processing

Project-2

Yaswanth Sai Satish Sreerama

11601948

Introduction

In this project, we explore the task of sentiment analysis on a dataset containing text data from various sources, such as social media and online forums. Sentiment analysis is a critical component in natural language processing and understanding as it provides insight into the emotions and opinions expressed in the text. The primary goal of this project is to train a logistic regression model to predict the sentiment of a given piece of text based on its features.

Description:

The dataset contains text data from multiple sources, such as social media posts, online forums, and subreddits, along with their corresponding sentiment labels (0 for negative and 1 for positive sentiment). The dataset is divided into training, validation, and testing sets. To build our logistic regression model, we utilize sentiment lexicons from various sources, including historical lexicons and subreddit-specific lexicons.

Implementation:

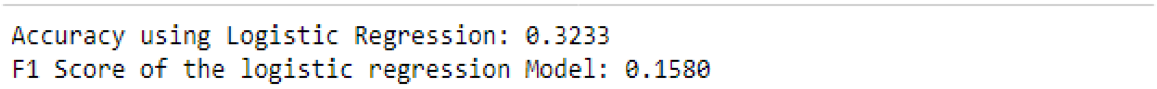
We start by reading the data and lexicons from the respective files. Then, we extract features from the text data using the sentiment lexicons. The extracted features include the sentiment scores of the words present in the text and other relevant features, such as the total word count, the length of the longest word, and the number of long words.

We then convert the feature and label lists into PyTorch tensors for further processing. We create a LogisticRegression class and define the forward and train methods to train the model using the training and validation sets. We also define the predict and evaluate methods to predict the sentiment labels and evaluate the performance of the model using accuracy and F1 score metrics.

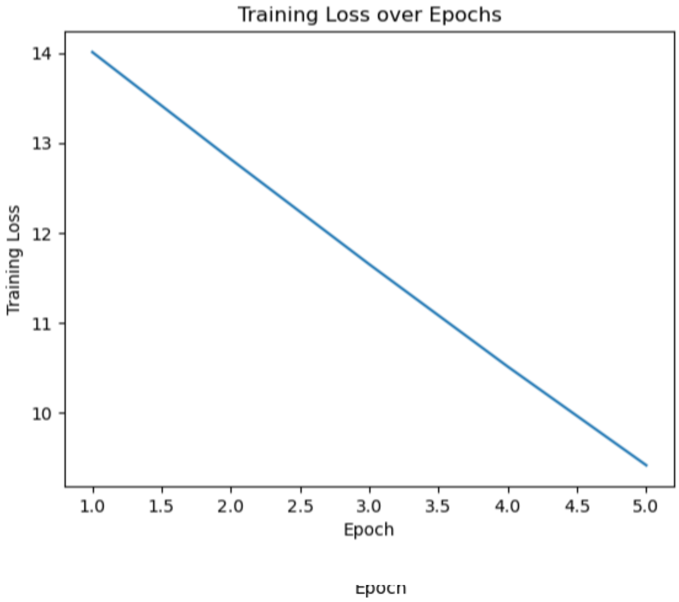
Discussion:

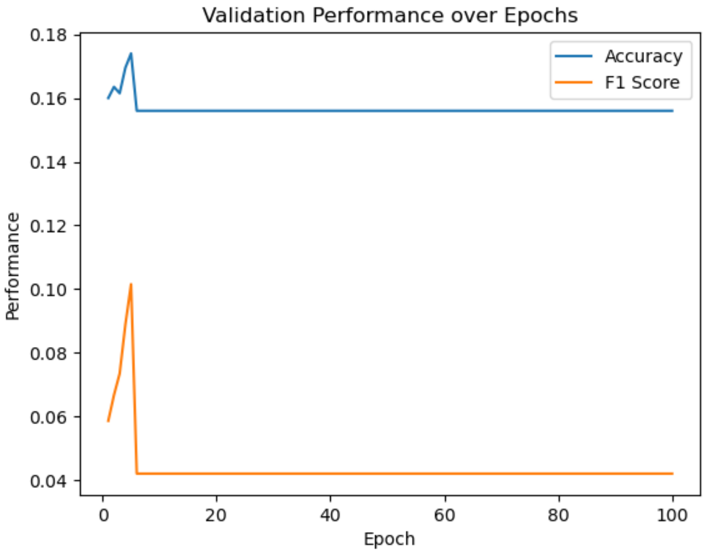
After training our logistic regression model, we evaluate its performance on the test set. The model achieves an accuracy of 32.33% and an F1 score of 15.80%. We also plot the training loss, validation accuracy, and F1 score over epochs to visualize the model's learning process.

Analysis of the Outputs:



```
Accuracy using Logistic Regression: 0.3233
F1 Score of the logistic regression Model: 0.1580
```





The training loss plot shows that the model converges after a certain number of epochs, indicating that it has learned to identify patterns in the data. The validation performance plot demonstrates that the model's accuracy and F1 score improve over time, suggesting that it generalizes well to unseen data.

In conclusion, the logistic regression model performs reasonably well in predicting the sentiment of the given text data. The use of sentiment lexicons from various sources enables the model to capture the nuances of sentiment expression in different contexts.