Machine learning (scientific report)

Mini Project 2

Introduction:

In the era of digital communication, understanding the sentiment behind a piece of text has become increasingly important, especially for social media platforms like Twitter where millions of tweets are posted every day expressing opinions on a wide range of topics. The ability to automatically classify these tweets according to their sentiment can provide valuable insights for various applications, from market research and political analysis to customer service and public opinion monitoring. The problem we are addressing in this project is sentiment analysis on Twitter data, with the aim to build machine learning models that can accurately classify tweets from the Sentiment140 dataset into different sentiment categories. Furthermore, tweets often contain special characters, URLs, and mentions, which need to be appropriately processed to ensure the effectiveness of the models. In this project, we explore two different machine learning models for sentiment analysis and compare their performance, with the goal not only to build models that perform well on this task but also to understand the strengths and weaknesses of different approaches to sentiment analysis on Twitter data.

Data processing method:

In the data processing stage, several important choices were made to ensure the data is in a suitable format for the machine learning models. Here's an overview:

1. **Loading the Data**:The data was loaded from a TSV file into a pandas Data Frame using the pd.read_csv function with the delimiter set to '\t'. This is a standard method for loading tabular data into a format that can be easily manipulated in Python.
2. **Splitting the Data**: The data was split into training and testing sets using the train_test_split function from sklearn.model_selection. This is a common practice in machine learning to ensure that the model's performance is evaluated on unseen data. The test size was set to 0.2, meaning that 80% of the data was used for training the model and 20% was used for testing.
3. **Text Transformation**: The text data was transformed into numerical vectors using two different methods: Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW). These are standard techniques in natural language processing that convert text into a format that can be understood by machine learning models. The TF-IDF vectors were used for the Logistic Regression model, while the BoW vectors were used for the Naive Bayes model.

Modeling

Two different machine learning models were used for sentiment analysis: Logistic Regression and Naive Bayes.

Logistic Regression: The Logistic Regression model was trained using the TF-IDF vectors of the tweet text. The model achieved an accuracy of approximately 78.6% on the test set. The precision, recall, and F1-score for both classes (0 and 4) were around 0.79, indicating a balanced performance for both classes.

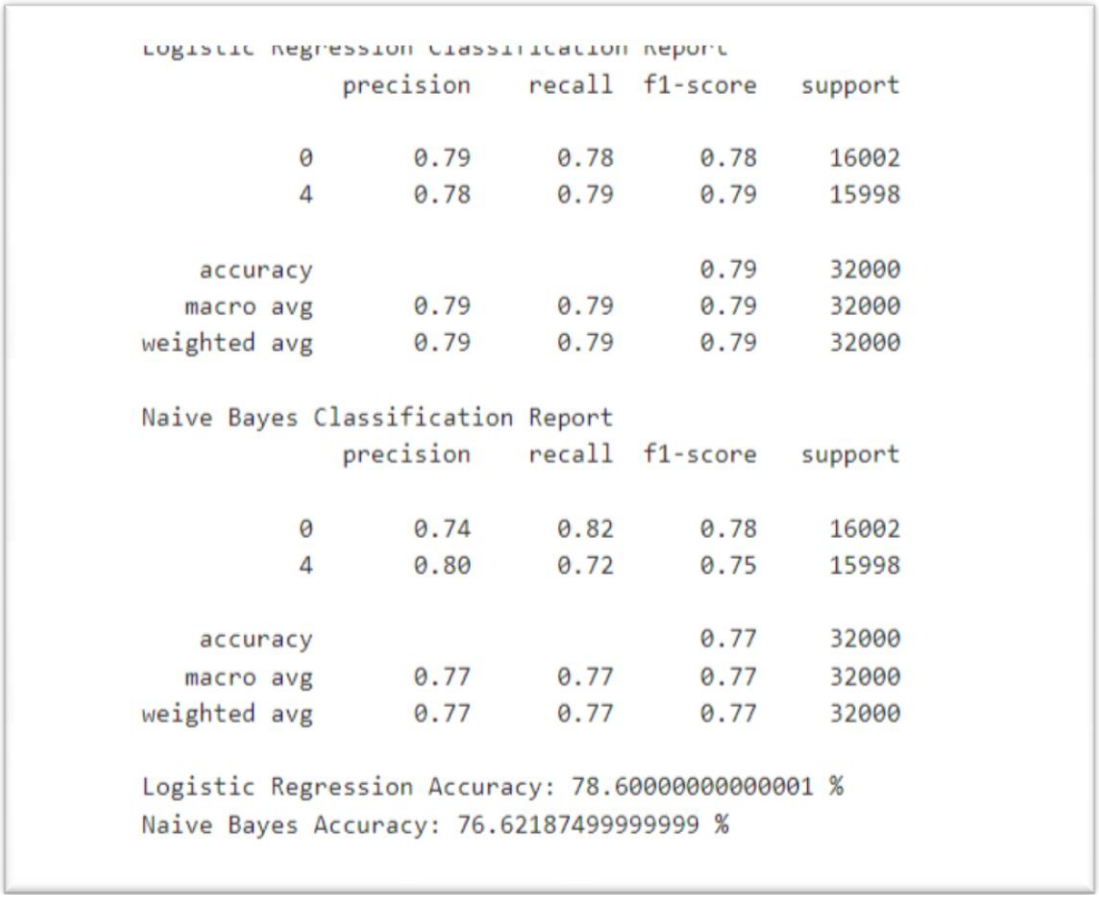

Fig 1: Two Model Accuracy

Naive Bayes: The Naive Bayes model was trained using the Bag of Words vectors of the tweet text. This model achieved an accuracy of approximately 76.6% on the test set. The precision, recall, and F1-score varied more between the two classes compared to the Logistic Regression model. Specifically, the model had a higher recall but lower precision for class 0, and a higher precision but lower recall for class 4.
The Logistic Regression model performed slightly better than the Naive Bayes model in terms of overall accuracy. This could be due to the fact that Logistic Regression makes fewer assumptions about the data and can model more complex relationships, while Naive Bayes assumes that all features are independent, which might not be the case with text data.

To improve the models, several steps could be taken:

1. **Feature Engineering:** Additional features could be created from the text data, such as the length of the tweets, the number of hash tags, mentions, etc.
2. **Parameter Tuning:** The parameters of the models could be tuned using techniques like grid search or random search to find the optimal parameters.
3. **Model Selection:** Other models could be explored, such as Support Vector Machines, Decision Trees, or even deep learning models like Convolution Neural Networks or Recurrent Neural Networks.
4. **Ensemble Methods:** Multiple models could be combined using ensemble methods to improve the performance.

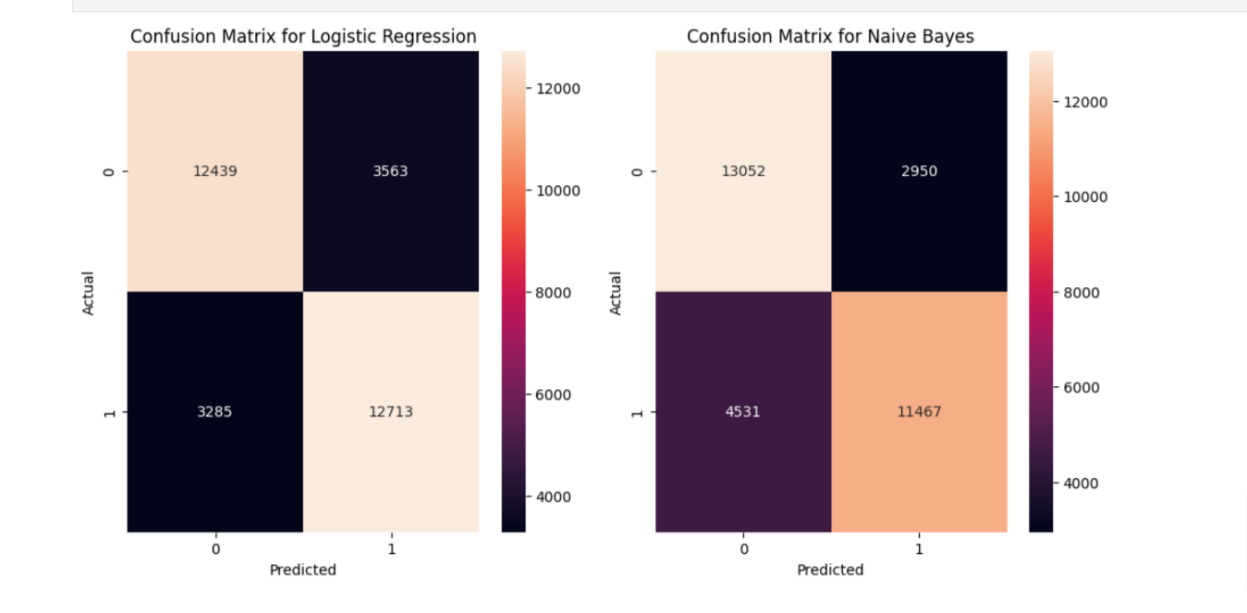

Fig 2: Confusion Matrix

Conclusion

This sentiment analysis project encountered and addressed several scientific challenges:

1. **High-Dimensional Feature Space**:

- Challenge: The bag-of-words model resulted in a high-dimensional feature space, which could lead to over fitting or poor generalization.
- Solution: We applied regularization techniques like L1 or L2 to avoid over fitting. We also considered exploring feature selection methods, such as selecting the most informative features.

2. **Imbalanced Classes**:

- Challenge: The dataset might have an imbalanced distribution of classes, for example, more negative tweets than positive ones.
- Solution: We considered techniques like oversampling the minority class or using class weights during training. We also evaluated performance using metrics beyond accuracy, such as precision, recall, and F1-score.

3. **Model Selection and Hyper parameter Tuning**:

- Challenge: The selection of the appropriate model and its hyper parameters is crucial for achieving good performance.
- Solution: We experimented with models like logistic regression and Naïve Bayes. We also considered hyper parameter tuning methods, such as grid search or random search, to further optimize model performance.

4. **Convergence Warning in Logistic Regression**:

- Challenge: The logistic regression model displayed a convergence warning.
- Solution: We increased the maximum number of iterations and scaled the data as suggested. We also considered investigating alternative solvers and preprocessing techniques.

5. **Understanding Model Predictions**:

- Challenge: Gaining insights into the reasons behind specific model predictions is crucial for building trust and interpretability.
- Solution: We can gain insights by conducting a feature importance analysis (for instance, by examining the coefficients in logistic regression) and visualizing decision boundaries.

6. **Applicability to Real-World Scenarios**:

- Challenge: The dataset used for training and testing might not accurately reflect real-world situations.
- Solution: Techniques like cross-validation and external validation on unseen data (like live tweets) are essential for evaluating the model's ability to generalize.

Warm Regards
Md Hasibul Haque Zahid
ID:2302302