Intensity Analysis: A Sentiment and Emotion Prediction Model

Project Overview

The "Intensity Analysis" project aims to analyse and predict emotions in text, such as happiness, sadness, or anger. With the rise of online reviews and social media, it's important for businesses to understand how people feel in their messages.

This project will create a system that uses Natural Language Processing (NLP) and machine learning models, specifically Support Vector Machines (SVM) or a Voting Classifier, to identify emotions and their intensity in text.

It follows a step-by-step process, including data cleaning, extracting features using TF-IDF, and choosing and training models. Both the SVM and Voting Classifier were evaluated for their performance, with the goal of creating a reliable and user-friendly tool for real-time emotional analysis.

Steps Followed

1. Data Collection

- Loaded three datasets containing text reviews categorized by emotional intensity: happiness, angriness, and sadness. (Provided by upgrad)

2. Data Preprocessing

- Combined the three datasets into a single Data Frame.
- Cleaned the text data by:
    - Converting text to lowercase.
    - Removing special characters and extra spaces.
- Created a new column with the cleaned text.

3. Data Splitting

- Split the data into training and testing sets, using 80% for training and 20% for testing.

4. Feature Extraction

- Used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert the cleaned text data into numerical features suitable for modeling.

5. Model Selection and Training

- Implemented multiple models:
  - **Support Vector Machine (SVM)**
  - **Logistic Regression**
  - **Naive Bayes**
  - **Random Forest**
  - **Gradient Boosting**
  - **Voting Classifier** (combination of SVM, Random Forest, and Gradient Boosting)
- Trained each model using the training dataset.

6. Model Evaluation

- Evaluated the performance of each model using accuracy scores and classification reports on the testing dataset.
- Recorded the following test accuracies:
  - SVM: 79.17%
  - Logistic Regression: 79.17%
  - Naive Bayes: 69.36%
  - Random Forest: 78.68%
  - Gradient Boosting: 72.30%
  - Voting Classifier: 80.64%

7. Model Saving

- Saved the **top two models (Voting Classifier and SVM)** as pickle files for future use.

Performance Evaluation

In this section, we assessed how well each model performed on the test dataset. We looked at accuracy scores and classification reports for each model. The Voting Classifier achieved the highest accuracy of 80.64%, indicating it is the most effective at predicting emotional intensity. Other models, like SVM and Logistic Regression, also

performed well, both with an accuracy of 79.17%. The evaluation showed that while some models are strong, there is room for improvement in others, like Naive Bayes and Gradient Boosting.

Discussion of Future Work

For future improvements, we could:

- Explore more advanced models, such as deep learning techniques (e.g., LSTM, BERT), to potentially increase accuracy.
- Gather more diverse datasets to enhance the model's ability to generalize across different text inputs.
- Implement additional text preprocessing steps, like lemmatization or using more sophisticated tokenization methods, to improve the quality of input data.
- Test the models in real-world applications to evaluate their performance in practical scenarios.

Conclusion

In conclusion, the Intensity Analyzer project successfully built a text classification system to determine emotional intensity from reviews. By comparing multiple models, we identified the Voting Classifier as the best performer. This project highlights the importance of model evaluation and the potential for further enhancements, making it a solid foundation for future developments in sentiment analysis.