Emotion Classification from Text

Chaudhari Khushi Ganesh
21084

Introduction

This project classifies emotions (Joy, Fear, Anger, Sadness, Surprise) from text using Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). The dataset comprises of the text snippets labeled with multiple emotions.The objective is to predict the correct emotion for each snippet.

Data Preprocessing

Missing values are removed from the dataset. The text was preprocessed using tokenization, lowercasing, and lemmatization to standardize the input. Non-alphabetical characters are removed and the stopwords are retained to preserve context. `TfidfVectorizer` is applied with n-grams to convert the text into numerical feature vectors.

Emotion labels are extracted from the dataset by identifying the emotion with the highest value for each text snippet. The labels are encoded using `LabelEncoder`. Although class imbalance existed, we use the original dataset without resampling due to moderate imbalance.

Model Training

Four classifiers are trained: Naive Bayes, Logistic Regression, Random Forest, and SVM. Each model's performance was evaluated using confusion matrices, classification reports, and accuracy scores.

Results Overview

Model	Accuracy	Fear (F1)	Joy (F1)	Anger (F1)	Surprise (F1)
Naive Bayes	52.75%	0.67	0.37	0.11	0.00
Logistic Regression	56.50%	0.69	0.46	0.09	0.08
Random Forest	56.25%	0.71	0.13	0.00	0.00
Support Vector Machine	56.50%	0.67	0.51	0.00	0.00

Table 1: Performance of different models on emotion classification

Conclusion

The SVM and Logistic Regression models achieved the highest accuracy (56.50%), but both showed limitations in classifying certain emotions like "Anger" and "Surprise." Overall, "Fear" was the most accurately predicted emotion across all models. Class imbalance problem needs to be considered for future work.