TVS Training & Services

PROJECT REPORT – ANALIZATION OF YOUTUBE VIDEO COMMENTS

GROUP NAME : GCEE – SCIPY

GROUP MEMBERS :

1. **21IMT25 - MANOJ V**
2. **21IMT16 - KABILAN T**
3. **21ECE16 - GOWTHAM R**
4. **21ECE08 - DHANUSRI T**
5. **21ECE52 - THARANI M**

1. ABSTRACT :

This project focuses on analyzing and visually representing YouTube comments to positive, negative and neutral sentiments. Using advanced natural language processing (NLP) techniques and sentiment analysis algorithms, the project aims to categorize comments into positive, negative, and neutral sentiments.

The results are then depicted through intuitive visualizations such as Bar chart or pie chart, which provide insights into viewer opinions and engagement. This approach not only helps in understanding audience reactions but also offers valuable feedback for content creators and marketers to tailor their strategies effectively.

2. MODEL IMPLEMENTATION :

The model implementation of a machine learning pipeline that leverages Support Vector Machine (SVM), Naive Bayes, and Decision Tree classifiers to categorize comments. First, collect a large dataset of comments by google API client, pre-process them by removing noise and performing text normalization by re, emoji modules and then convert the text data into numerical features using technique TF-IDF. Split the dataset into training and testing sets. Train the SVM, Naive Bayes, and Decision Tree models separately on the training data. Evaluate each model's performance on the testing set using metrics like accuracy, precision, recall, and F1-score. Finally, compare the results to determine which model performs best for this task, and consider ensemble methods to combine their strengths for improved classification accuracy.

3. ALGORITHM IMPLEMENTATION :

3.1. Data Preparation

- **Load Data:**
  The dataset should be in a format where each comment is paired with a label. For instance, you might have a list for storing comments and labels or store like a CSV file.

- **Preprocess Data:**
  Preprocess the data by cleaning it (removing punctuation, special characters, and typos) and tokenizing it (splitting sentences into words).

3.2. Feature Extraction

- **Text Vectorization:**
  Convert the text comments into numerical features. A common technique is to use Term Frequency-Inverse Document Frequency (TF-IDF), Count Vectorizer which transforms the text into a matrix of features.

3.3. Model Training

- **Support Vector Machine (SVM):**
  Train an SVM model with a linear kernel. SVMs are effective in high-dimensional spaces and are well-suited for text classification tasks.

- **Naive Bayes:**
  Train a Naive Bayes classifier, particularly the Multinomial Naive Bayes, which is effective for text data and works well with word counts or TF-IDF features.

- **Decision Tree:**
  Train a Decision Tree classifier, which is simple yet powerful for classification tasks. It works by learning decision rules from the features.

3.4 Model Evaluation

- **Predict and Evaluate:**
  Use the trained models to make predictions on the test set. Evaluate the performance of each model using metrics like accuracy and classification report (which includes precision, recall, and F1-score).

4. PREDICATION COMPARISON REPORT :

To compare how well three different models – Support Vector Model(SVM), Naïve Bayes and Decision Tree perform in classifying Youtube video comments.

Sample Accuracy scores :

- **Naïve Bayes :** Achieved an accuracy score of 85%.
- **SVM :** Achieved an accuracy score of 100%.
- **Decision Tree :** Achieved accuracy score of 55%.

Visualization of Result :

A bar chart displays the accuracy of each classifier (SVM, Naïve Bayes and Decision Tree). The SVM model is the best at correctly predicting comment categories, followed by Naïve Bayes and then Decision Tree.

5. FINAL PREDICTIONS :

To evaluate and predict the most effective classification model for YouTube video comments using Support Vector Machine (SVM), Naive Bayes, and Decision Tree classifiers.

- **SVM :**

  Given its high accuracy and balanced performance metrics, SVM is the recommended model for classifying YouTube video comments. It is likely to provide the most reliable results in identifying comment sentiments or categories.

- **NAÏVE BAYES :**

  Naive Bayes is a viable alternative, especially if computational resources are limited or if interpretability and simplicity are prioritized over maximum accuracy.

- **DECISION TREE :**

  While Decision Trees offer interpretability and can highlight key features, they are less effective compared to SVM and Naive Bayes for this task. They may be used for exploratory analysis or when feature importance is a critical factor.

6. CONCLUSION :

Our study investigated the prediction of like proportions of trending Youtube videos using a simple prediction based on the number of comments classified as positive, neutral and negative. Some correlation was found between the like proportions of Youtube videos and the predicted like proportions based on the sentiment of their comments classified by the classifiers used in this study (SVM, multinomial NB, and Decision Tree). But due to the high prediction errors and standard deviations obtained this type of prediction we used is not of great use in predicting like proportions for trending Youtube videos in real-world scenarios, such as if dislike counts were to be hidden. The classifiers performed better with only Youtube comments as training compared to a larger dataset of Youtube comments and tweets.