Summary Report

Categorizing YouTube Comments Using Machine Learning

1. Introduction

A. Objective and Context

In this project, we aim to explore the fascinating world of YouTube video data. Our central focus is to understand the intricate relationship between video characteristics and viewer engagement. Specifically, we'll analyze comments and sentiments associated with top YouTube videos to gain insights into viewer behavior. The objective is to glean insights into viewer behavior and preferences, aiding content creators and platform managers in optimizing content strategy and user experience.

B. Data Source

For this project, we were able to source our data from Kaggle. We took the “Trending Youtube Video Statistics” data set by Mitchell J and then went forward in our project with the data relevant to the US.

We work with two primary datasets:

- Video Information Dataset: Contains details about YouTube videos, including metadata such as title, description, view count, and upload date.
- Comments Dataset: Captures comments posted by viewers on these videos.

2. Initial Steps: Data Loading and Exploration

A. Data Acquisition

We sourced our data from CSV files, ensuring that we had a comprehensive representation of YouTube content. These files were meticulously cleaned and preprocessed to remove any inconsistencies or missing values.

B. Exploratory Data Analysis (EDA)

Our initial exploration involved:

- Checking for missing values
- Examining summary statistics
- Visualizing key features (e.g., view counts, comment counts)

3. Diving Deep: Sentiment Analysis

A. Understanding Sentiment in Comments

We focused on the top 1000 videos (presumably sorted by view count) and performed sentiment analysis on their comments. The TextBlob library was our tool of choice for this task. Here's what we did:

B. Sentiment Polarity Calculation:

- For each comment, we calculated its sentiment polarity score using TextBlob. The polarity score ranges from -1 (negative) to 1 (positive), with 0 indicating neutrality.
- The distribution of these scores was visualized using a violin plot. This plot combines a box plot and a kernel density plot, providing insights into the density of comments at different polarity values.

C. Correlation with Video Views:

- We computed the Pearson correlation coefficient between sentiment polarity and video views.
- Surprisingly, the correlation was weak (with a value of 0.017). In other words, the sentiment expressed in comments does not strongly correlate with video popularity (views).
- However, remember that correlation doesn't imply causation. Other factors, such as video content, audience demographics, or engagement metrics, may play a more significant role.

D. Takeaways

- Sentiment in comments alone isn't a reliable predictor of video success.
- Further investigation is needed to uncover hidden patterns.

4. Data Preprocessing

1. Cleaning and Tokenization

We start by cleaning the text data in the **comments_text** variable. Common cleaning steps include:

- Converting all text to lowercase.
- Remove Non-Alphanumeric Characters:
- Non-alphanumeric characters (such as punctuation marks, symbols, etc.) can be distracting and may not contribute much to the analysis. Removing them simplifies the text data.
- Remove Stop Words and Emojis.
- Stop words (common words with little semantic meaning) are often removed to focus on more meaningful content.
- Emojis, while expressive, can be challenging to handle in text analysis. Removing them ensures a cleaner dataset.
- Tokenize the Comments Using word_tokenize: Tokenization broke down the text into individual words, facilitating further analysis.

2. Feature Extraction

- We used the CountVectorizer to convert text data into a matrix of token counts. This matrix served as input for our machine-learning models.
- Additionally, we applied Principal Component Analysis (PCA) using TruncatedSVD to reduce dimensionality while preserving variance.

5. Model Training and Evaluation

- We split the data into training and testing sets using k-fold cross-validation (stratified to maintain class distribution).
- A common choice for text classification is logistic regression, but other models can also be used.

Classification Models

We trained several classification models using GridSearchCV to find the best hyperparameters:

- Random Forest
- Logistic Regression
- Multinomial Naive Bayes
- SVC
- KNN Classifier

Used the best parameters after some tuning,

```
#train multiple classification models and obtain the accuracy scores

rf = RandomForestClassifier(n_estimators=200, max_depth=20, max_features='log2', criterion='gini', random_state=42)

log_reg = LogisticRegression(C = 0.001, max_iter=1000)

nb = MultinomialNB(alpha = 0.5, fit_prior = True)

svc = SVC(C=0.1, gamma=0.1, kernel='rbf')

knn = KNeighborsClassifier(n_neighbors=3)
```

Performance Evaluation

- Cross-validation was used to assess model performance.
- Metrics included accuracy, precision, and recall.
- Bar plots helped compare model performance.

	Accuracy	**Precision**	**Recall**
RandomForestClassifier	0.202718	0.999283	0.202718
LogisticRegression	0.355738	0.656595	0.355738
MultinomialNB	0.589688	0.609525	0.589688

6. Unveiling Insights

1. **Key Findings**
    a. The analysis revealed insights into the relationship between sentiment in comments and video views, indicating a weak correlation.
    b. The word cloud provided insights into the most frequent words used in comments, highlighting topics of interest among viewers.
    c. Model performance varied across different classification algorithms, with some models achieving higher accuracy and precision in categorizing videos based on comments.
2. **Implications**
    a. Content creators can tailor video content based on sentiment insights.
    b. Platform managers can categorize videos for better recommendations.

7. Wrapping Up: Conclusion and Recommendations

Conclusion

- The analysis provided valuable insights into viewer engagement and sentiment on YouTube videos, offering potential strategies for content creators to enhance viewer interaction and optimize video content. Further analysis and experimentation with advanced neural networks could provide deeper insights and improve classification accuracy.

Recommendations

- Content creators could leverage insights from sentiment analysis to tailor video content and engage with their audience more effectively.
- Platform managers could utilize classification models to categorize videos based on comments, enabling better content recommendation and user experience customization.

8. Looking Ahead: Future Work

Next Steps

- Future analysis could explore additional features such as video metadata and user engagement metrics to enhance classification accuracy further.
- Experimentation with advanced machine learning techniques such as deep learning could provide deeper insights into viewer behavior and sentiment.