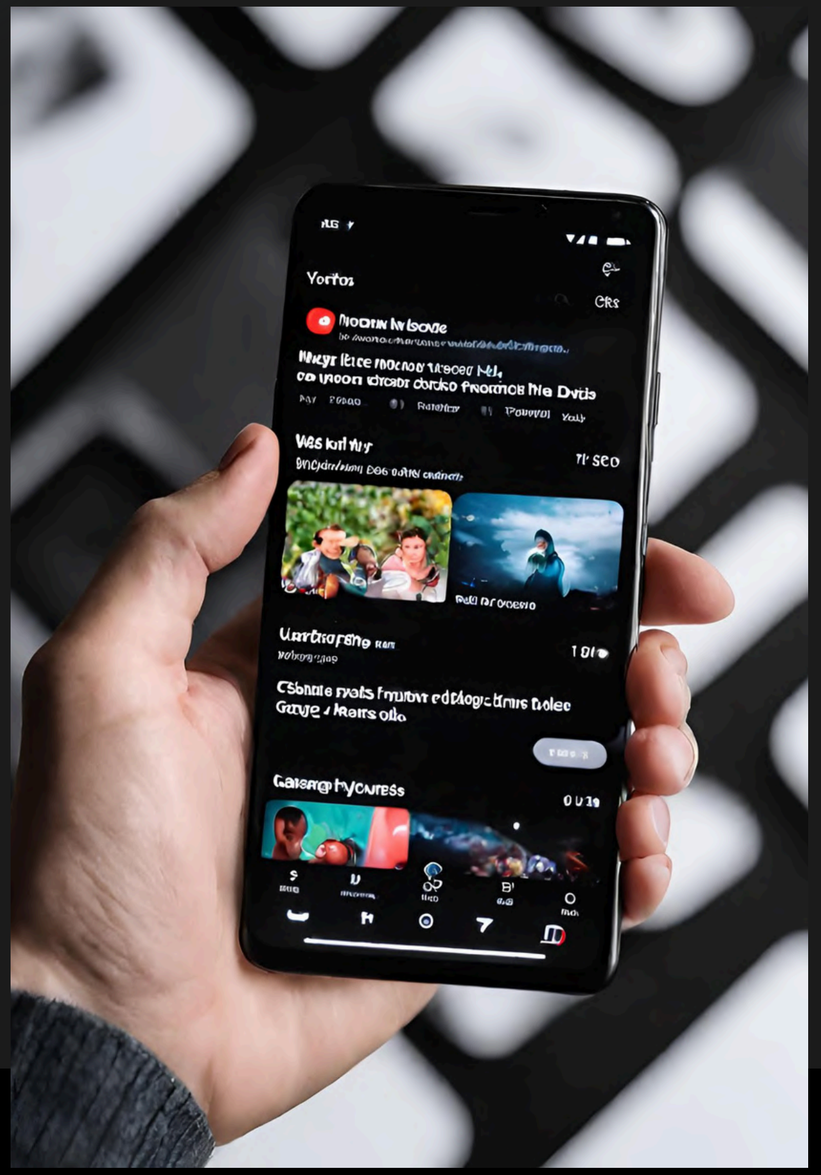
YouTube Comments Sentiment Analyzer



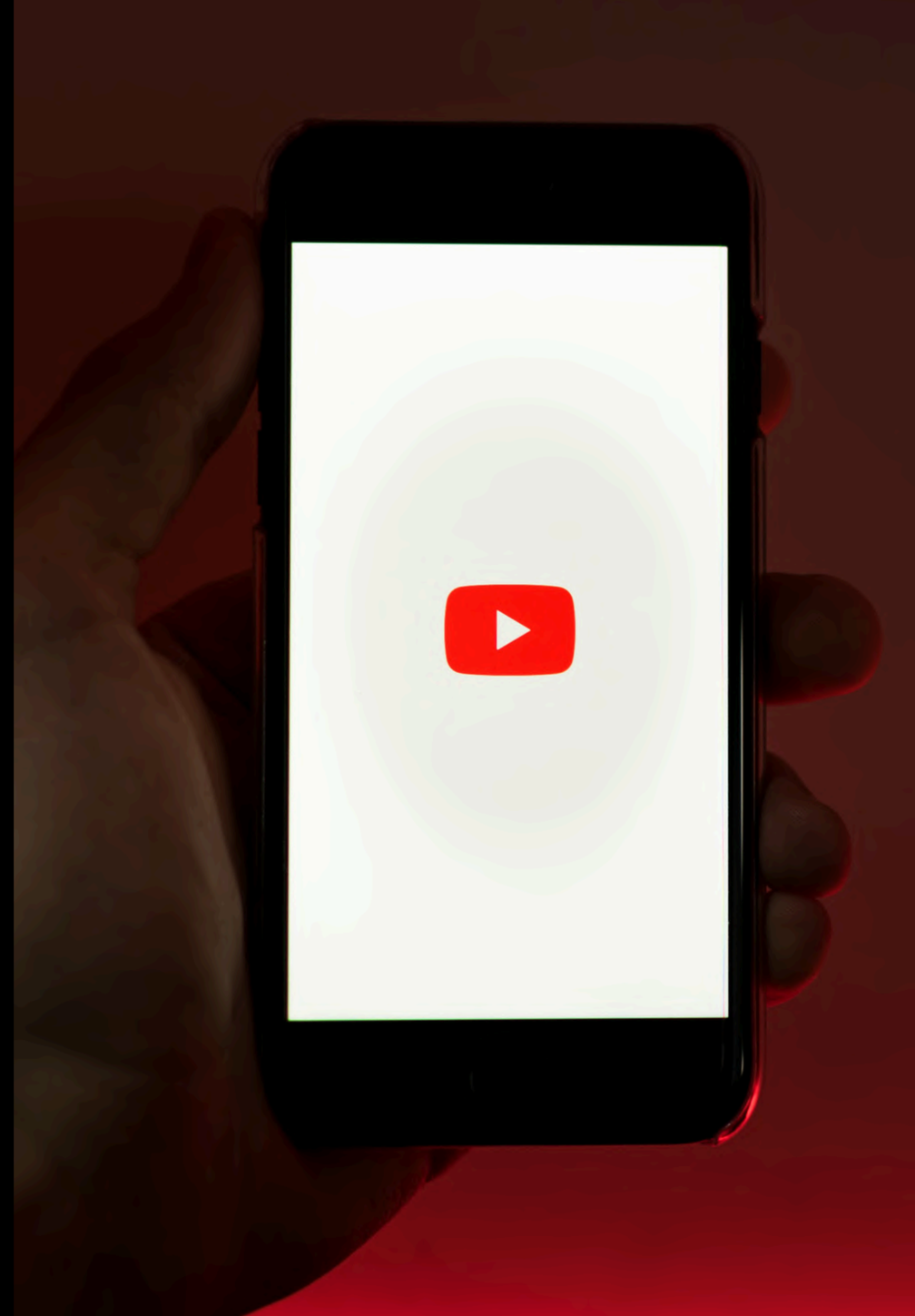
ARYA P G
PRATHAM K
MUGDHA S
SHREYAS S

Outline

- Project Objective
- Scraping Module
- Sentiment Analysis Model
- NLP Techniques
- Machine Learning Training
- Sentiment Distribution
- User Interface
- Scalability and Performance
- Results Analysis
- Future Enhancements

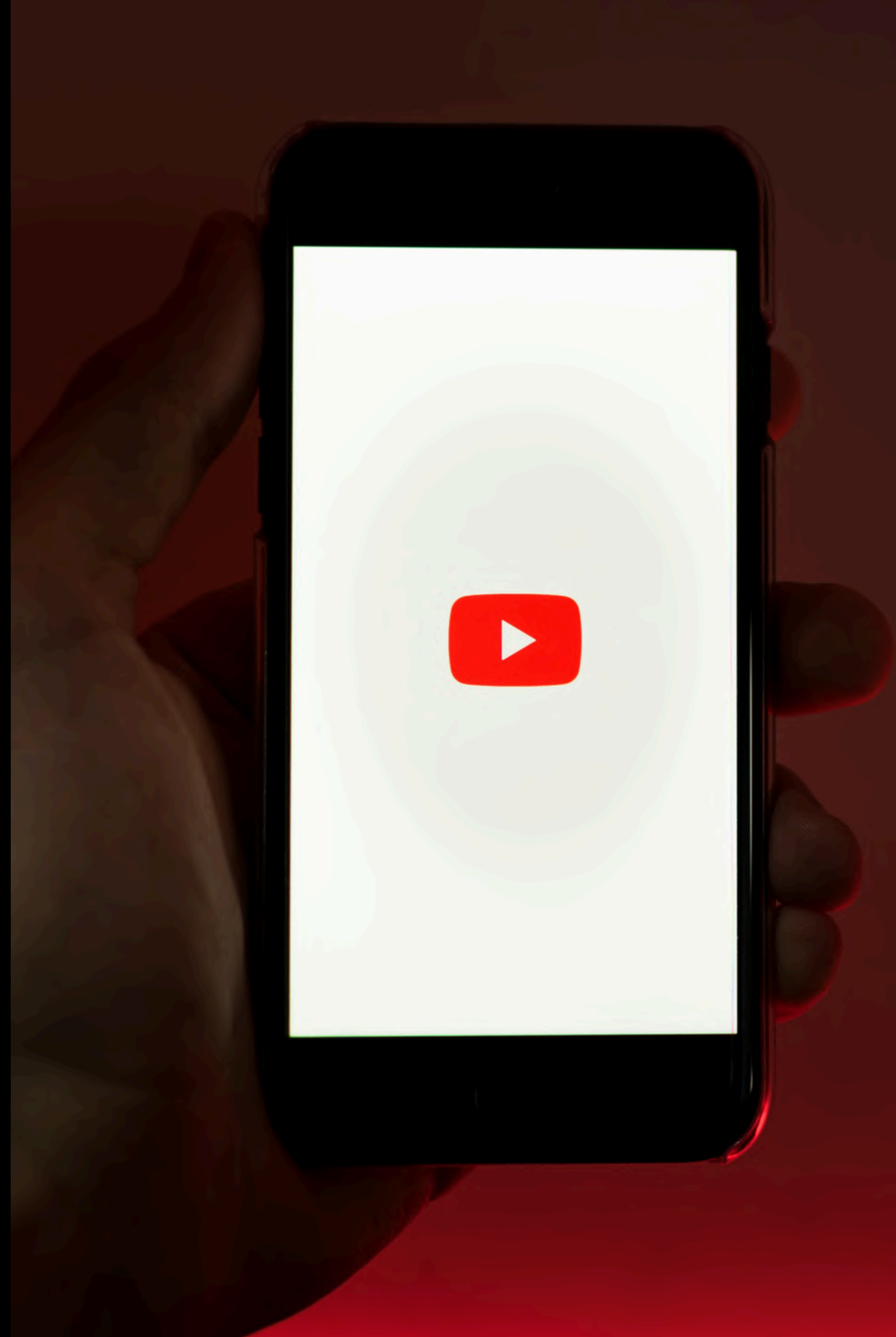
Problem Statement

- Creating a Comment Sentiment Analyzer for monitoring of youtube video comments.
- This tool predicts individual comment sentiment while also analyzing the overall sentiment trend over time, identifying key events, and providing insights into the evolving sentiment landscape.
- Ideal for businesses and individuals seeking immediate and comprehensive public sentiment understanding.



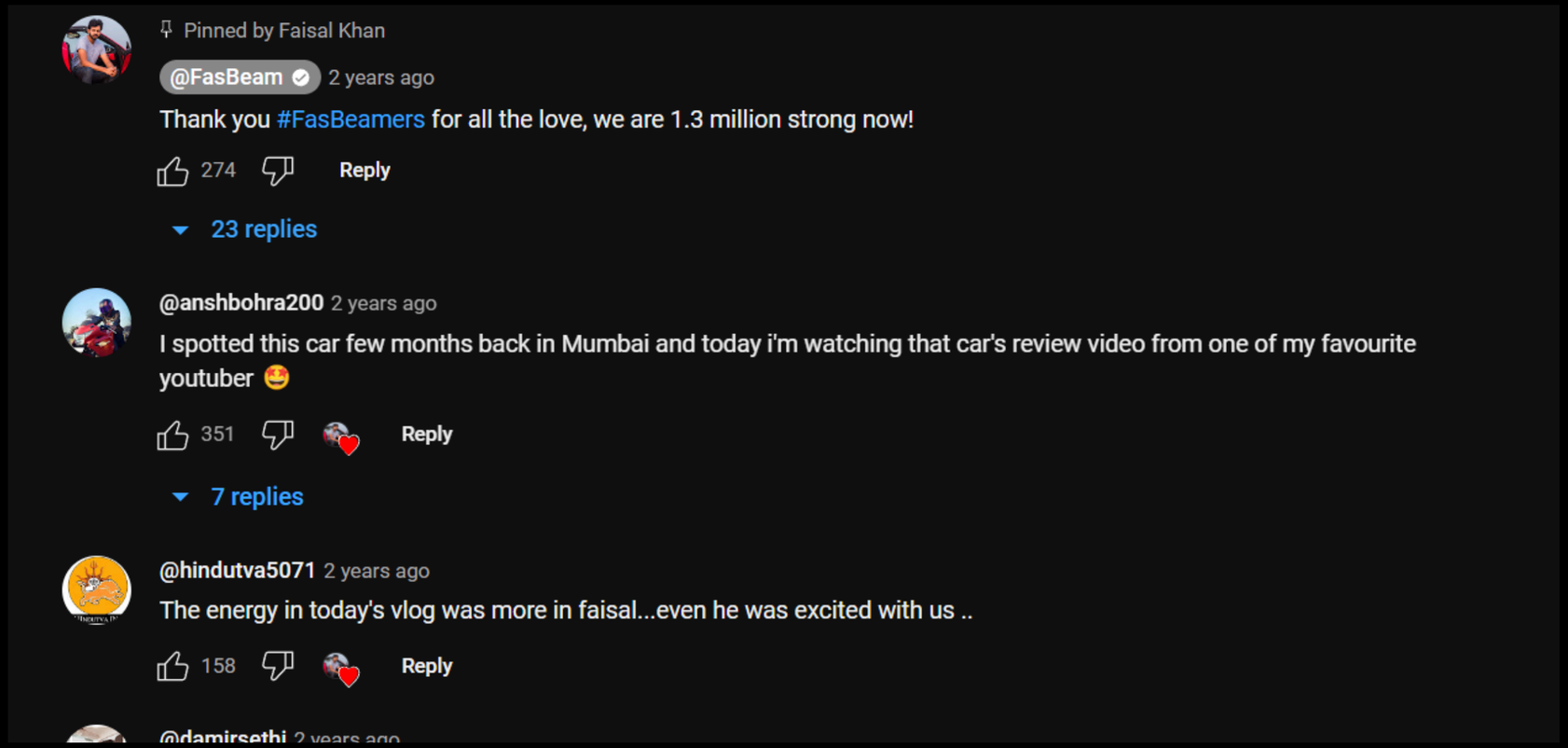
Introduction

- YouTube serves as a significant platform where billions engage daily, shaping public opinion through comments, reactions, and discussions.
- The tool predicts individual comment sentiment and analyzes overall sentiment trends over time, providing insights into the evolving sentiment landscape.
- It offers a sophisticated understanding of sentiment by leveraging cutting-edge natural language processing techniques.
- Beyond merely categorizing comments as positive, negative, or neutral, this analyzer offers a sophisticated understanding of the evolving sentiment landscape. Objectives include data collection, preprocessing comments, implementing machine learning models for sentiment analysis, and creating user-friendly interfaces and dashboards.
- This study unveils the inner workings of the tool, detailing its methodology, capabilities, and implications for decoding public opinion on YouTube.



Comments Sentiment Analyzer

It classifies the youtube comments as positive,negative and neutral.



ML MODEL

The Model used is Random Forest Classifier

Handling Non-Linearity:

- Random Forest is a powerful ensemble learning method that can effectively capture non-linear relationships and interactions between features in the data.
- This is particularly beneficial for sentiment analysis, as sentiment expression in text data often involves complex and non-linear patterns.

Robustness Against Overfitting:

Random Forest models are less prone to overfitting compared to some other complex models like deep neural networks.

This is crucial when dealing with sentiment analysis tasks where the model needs to generalize well to unseen data and avoid memorizing noise or specific instances from the training set

Efficiency and Scalability:

Random Forest can handle large datasets efficiently and can scale well with increasing data size.

Dataset details

Inputs:

- Training - https://www.youtube.com/watch?v=qjw4esSW-W4
- Testing - https://youtu.be/3tCBIDzF34M?si=0nxhHfzq9FOXZ-IN

Outputs:

- Sentiment of the particular YT video’s comments

Training Dataset Details

- Number of samples in the training set is 2268
- The independent variable: Comments column
- It is categorized into three classes
  - 585 - Positive ⊕
  - 183 - Negative ⊖
  - 1499 - Neutral 😐

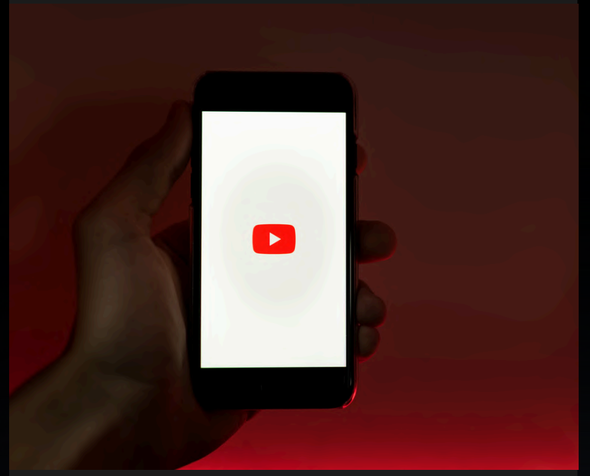
Testing Dataset Details

- Number of samples in the training set is 8420
- The independent variable: Comments column
- It is categorized into three classes
  - 1849 - Positive ⊕
  - 291 - Negative ⊖
  - 6280 - Neutral 🙂

Training Workflow

SCRAPE YOUTUBE COMMENTS

Defined a function (**get_video_comments**) to fetch comments for a specified YouTube video ID using youtube API key and store it in a list



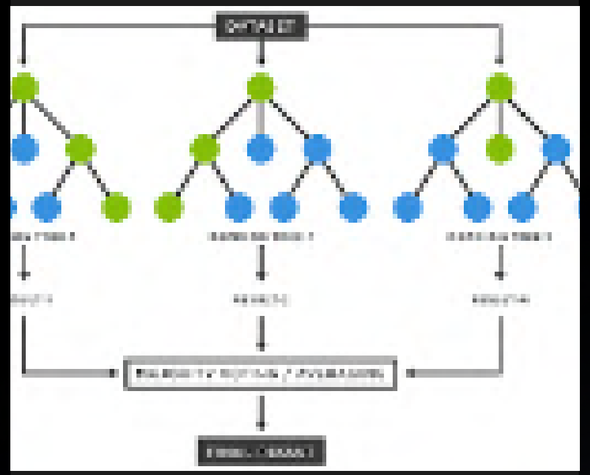
PREPROCESSING TEXT DATA

- Defined a function (**remove_emojis**) to remove emojis from text using regex.
- Applied emoji removal to the comment data obtained from YouTube.
- Lowercase the text, remove special characters, and apply word tokenization using NLTK.

PREPARING DATA FOR CLASSIFICATION

- Splitting the data into features (**X**: comments) and labels (**y**: sentiment labels).
- Vectorize the text data using TF-IDF Vectorizer to convert text into numerical features.

Training Workflow

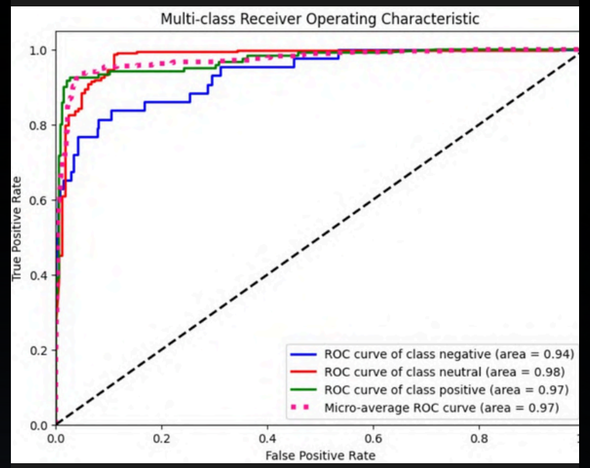


TRAINING OF RANDOM FOREST CLASSIFIER

- Splitting the data into training and testing sets.
- Initialize and train a Random Forest Classifier using **sklearn**.

EVALUATION OF THE MODEL

- Make predictions on the test set using the trained classifier.
- Calculate accuracy and display a classification report (precision, recall, F1-score).



ROC CURVE AND AUC CALCULATION

- Computed ROC curves and AUC for each sentiment class using **sklearn**.
- Plotted ROC curves for each class and the micro-average ROC curve.
- Print AUC values for each class and the micro-average AUC.

Accuracy details

Before Hyperparameter tuning

- Accuracy of the model obtained for training is 92%
- Accuracy of the model obtained for testing is 88%

- Precision:
  - Negative: 0.95
  - Neutral: 0.86
  - Positive: 0.96

- Recall:
  - Negative: 0.44
  - Neutral: 0.99
  - Positive: 0.78

- F1-score:
  - Negative: 0.60
  - Neutral: 0.92
  - Positive: 0.86

- AUC using the RoC curve:
  - AUC for class Negative: 0.92
  - AUC for class Neutral: 0.97
  - AUC for class Positive: 0.97
  - Micro-average AUC: 0.96

After Hyperparameter tuning

- Accuracy of the model obtained for training is 97%
- Accuracy of the model obtained for testing is 92%

- Precision:
  - Negative: 0.97
  - Neutral: 0.92
  - Positive: 0.96

- Recall:
  - Negative: 0.68
  - Neutral: 1.00
  - Positive: 0.87

- F1-score:
  - Negative: 0.80
  - Neutral: 0.96
  - Positive: 0.92

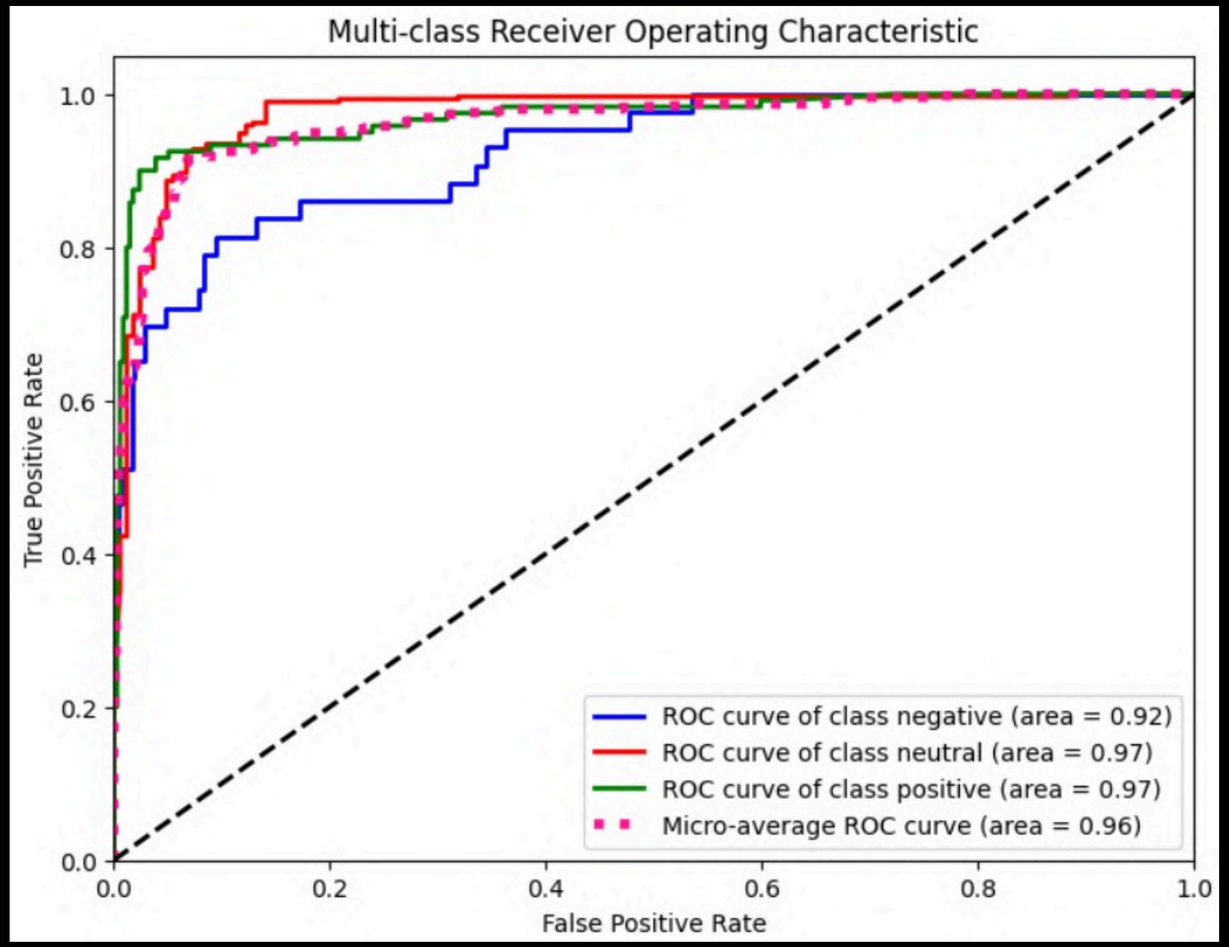
- AUC using the RoC curve:
  - AUC for class Negative: 0.94
  - AUC for class Neutral: 0.98
  - AUC for class Positive: 0.97
  - Micro-average AUC: 0.97

- Accuracy of the model obtained for training of testing dataset is 97%

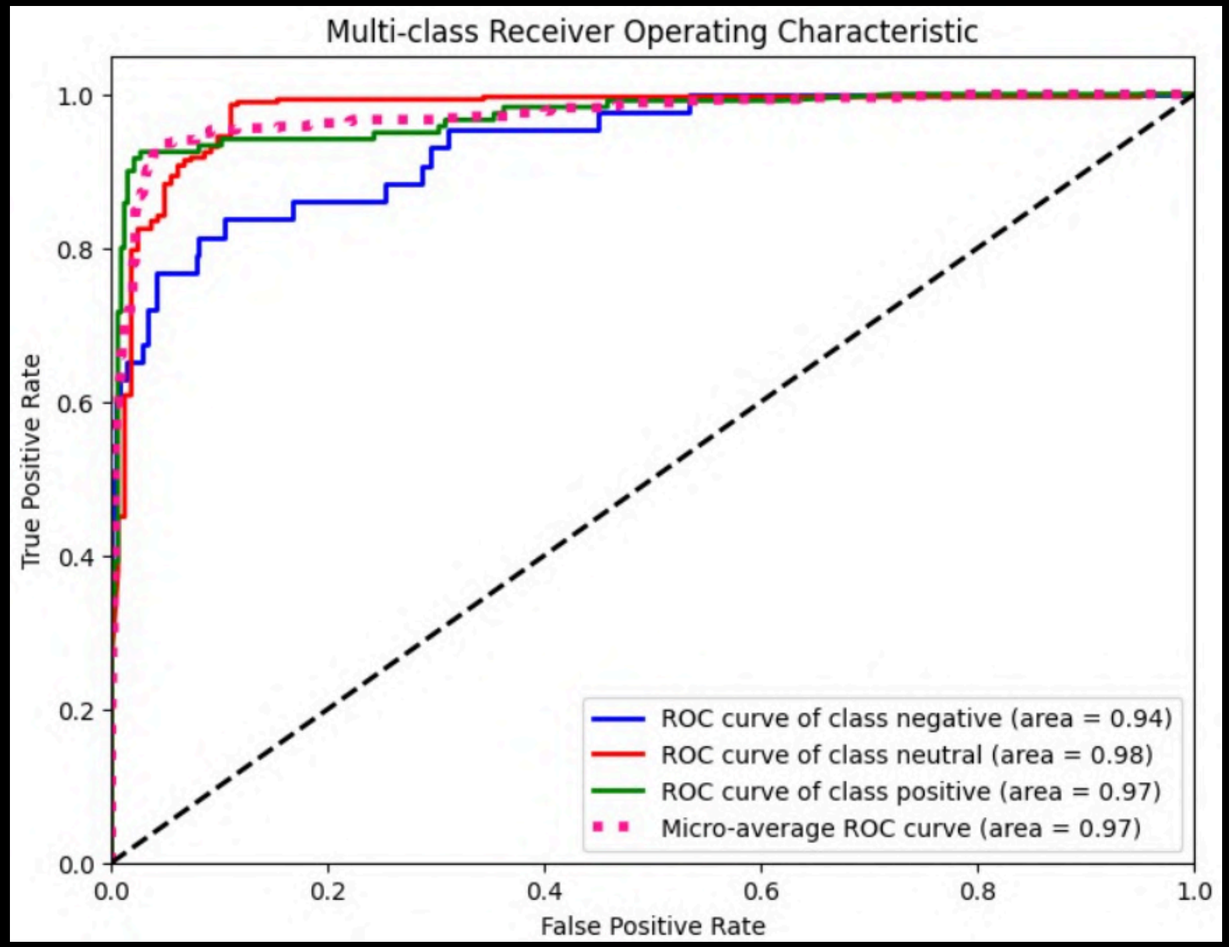
- Accuracy of the model obtained for testing of testing dataset is 96%

ROC Curves

BEFORE HYPERPARAMETER TUNING



AFTER HYPERPARAMETER TUNING



Result Analysis

Begins by vectorizing text data using TF-IDF. Then, it splits the data into training and testing sets, followed by cross-validating a Random Forest Classifier. It further implements ensemble methods: Bagging Classifier and AdaBoost Classifier, both utilizing Random Forest as base estimators.

Accuracy Obtained

- RF Classifier Cross-Validation Scores: [88.98, 91.46, 89.25, 93.38, 92.54]
- Mean Accuracy: 91.125
- Bagging Classifier Score: 91.629
- AdaBoost Classifier Score: 94.052

Website Description

- Our project encompasses a user-friendly webpage designed for analyzing sentiments within YouTube video comments.
- Users can seamlessly share the YouTube video link on our platform to receive sentiment scores reflecting the overall sentiment of the video's comments.
- Leveraging React for frontend development and Flask for backend development, our webpage ensures a smooth and responsive user experience.
- The sentiment scores encompass three categories: positive, neutral, and negative.
- Upon analysis, our webpage dynamically changes its color scheme to visually represent the sentiment distribution.

Higher Positive Comments Weightage

YouTube Comment Sentiment Analyzer

Enter the YouTube link:

https://www.youtube.com/watch?v=ZUHsAlKcMOM

Analyze

Sentiment Analysis Results

Positive 😊:

48.41%

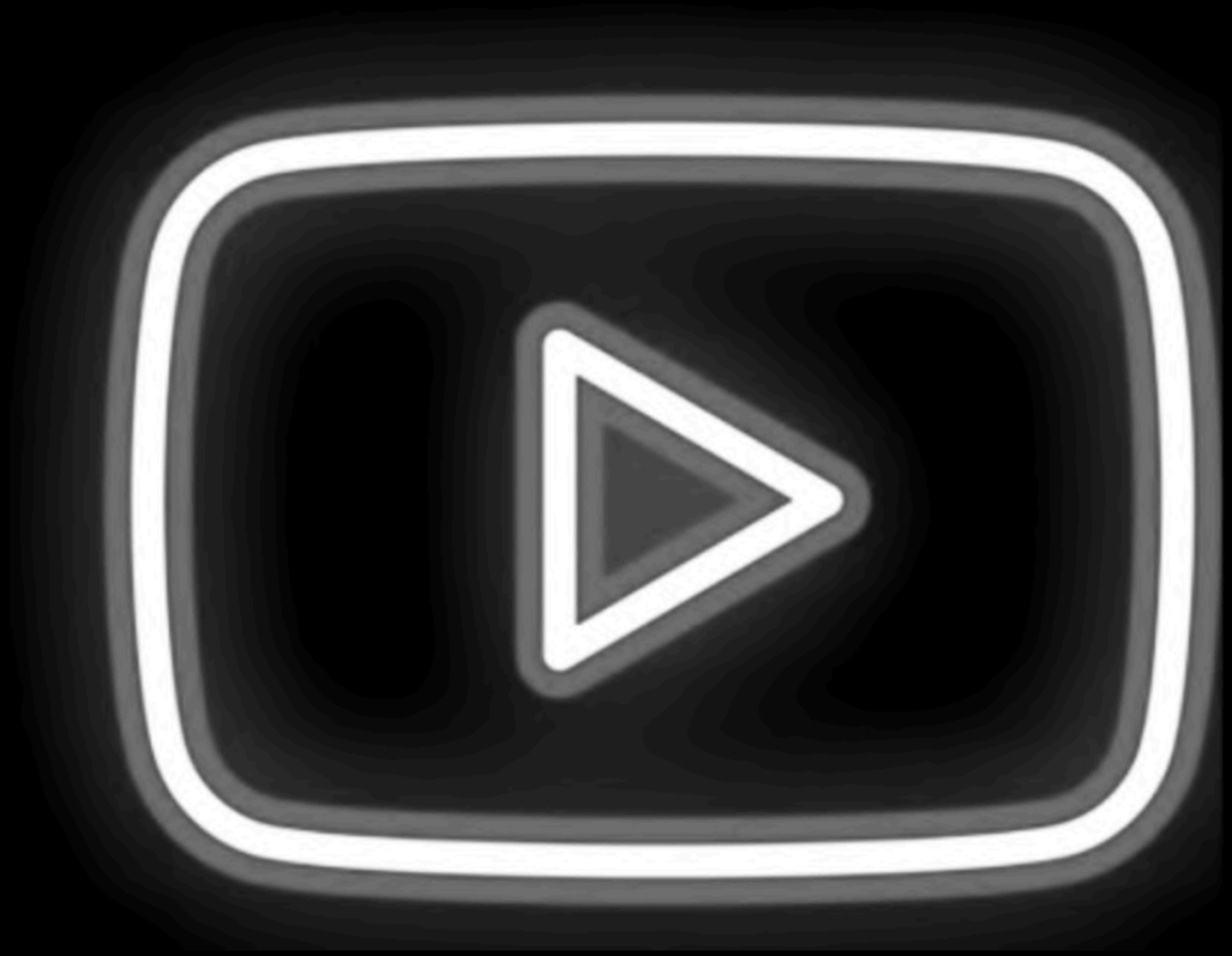
Neutral 😐:

36.98%

Negative 😟:

14.60%

Higher Neutral Comments Weightage



YouTube Video Comment Sentiment Analyzer

Enter the YouTube link:

https://youtu.be/3tCBIDzF34M?si=0nxhHfzq9FOXZ-IN

Analyze

Sentiment Analysis Results

Positive 😊:

22.49%

Neutral 😐:

73.97%

Negative 😠:

3.54%

Higher Negative Comments Weightage

YouTube Video Comment Sentiment Analyzer

Enter the YouTube link:

https://youtu.be/cng69luclUg?si=iWuq61DYZGggf4Zp

Analyzing...

Sentiment Analysis Results

Positive 😊:

30.25%

Neutral 😐:

34.64%

Negative 😠:

35.10%

Conclusion

This research underscores the effectiveness of machine learning, specifically the Random Forest algorithm with fine-tuned parameters, in accurately categorizing sentiments expressed in YouTube comments.

Through sophisticated natural language processing techniques and ensemble learning, the model achieved impressive accuracy, precision, recall, and differentiation among positive, negative, and neutral sentiments.

Nonetheless, it's important to note that the study's scope was confined to comments from specific YouTube videos, implying that performance might differ in other domains or languages. Expanding the analysis to encompass multiple languages and examining the influence of data quantity and quality would further solidify the conclusions.

Despite these considerations, this endeavor lays a sturdy groundwork for sentiment analysis of user-generated content, with potential applications in fields such as product evaluations, brand surveillance, and gauging public sentiment on diverse subjects.

Future Work

- Language Translation: The current model is trained on English comments. To make it more versatile and applicable to a global audience, language translation capabilities could be added. This would allow the model to analyze comments in multiple languages, increasing its reach and relevance.

- Real-Time Sentiment Tracking: The report discusses analyzing sentiment trends over time. Building a real-time sentiment tracking system could provide valuable insights into how sentiment evolves dynamically, enabling quicker response to emerging trends or events.

- Integration with Video Content Analysis: Combining sentiment analysis with video content analysis (e.g., speech recognition, visual analysis) could provide a more comprehensive understanding of the overall sentiment expressed in a YouTube video, accounting for both the comments and the video content itself.

- Explainable AI: Incorporating explainable AI techniques could help users understand the reasoning behind the model's sentiment classifications, increasing transparency and trust in the system.

References

YouTube Video for dataset

- https://www.youtube.com/watch?v=qjw4esSW-W4
- https://youtu.be/3tCBIDzF34M?si=0nxhHfzq9FOXZ-IN

Images

- https://www.androidcentral.com/apps-software/android-tv/the-google-tv-app-gets-a-much-needed-redesign-but-theres-one-missing-piece
- https://onezero.medium.com/the-one-rule-of-content-moderation-that-every-platform-follows-ab6323e0e293
- https://www.linkedin.com/pulse/one-minute-overview-random-forest-algorithm-saulius-dobilas?utm_source=share&utm_medium=member_android&utm_campaign=share_via
- https://saasmetrics.co/helpful-tips-to-grow-fast-your-youtube-channel/
- https://turbofuture.com/internet/How-to-Become-a-Better-YouTuber-in-10-Steps
- https://gshiftlabs.com/how-to-calculate-growth-rate-in-excel/
- https://www.freepik.com/premium-photo/business-team-present-professional-investor-working-new-start-up-project-digital-tablet-laptop-computer-finance-managers-meeting-with-digital-marketing-media-virtual-icon_25850172.htm
- https://divakarkumarp.medium.com/how-did-the-random-forest-algorithm-work-in-machine-learning-9e044573898b

Thank you!