ANALYSIS OF USER COMMENTS ON YOUTUBE VIDEOS USING MACHINE LEARNING

Dr. Y. Usha Rani[1], A Shiva Teja[2], A Abhishek Rao[3], M Akshitha[4], M Roshini[5]

[1]Assistant Professor, Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India [2,3,4,5]Students, Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

ABSTRACT:

The surge in user comments on YouTube channels poses a challenge for content creators seeking relevant and engaging feedback[1]. Addressing this, our project introduces a unique analysis framework utilizing machine learning. We combine sentiment analysis and sentence type classification to categorize YouTube comments, aiding creators in enhancing viewer engagement. Distinct from prior studies, we tailor our approach to the specific characteristics of YouTube comments. Leveraging machine learning models like Naive Bayes, SVM, Random Forest, Gradient Boosting, and Logistic Regression, along with established statistical measures, we evaluate their performance using cross-validation and achieve impressive accuracies of 96.14% (Logistic Regression) and 97.33% (Random Forest). This comprehensive technique provides content creators with a powerful tool to navigate and leverage the vast volume of user comments, thereby enriching their channels and viewer interaction.

Keywords: Sentiment Analysis, Naïve bayes, SVM, Random Forest, Gradient Boosting, Logistic regression, Stacking Classifier, Natural Language Processing, Machine learning

1. INTRODUCTION:

In today's world Information and opinions are like bees in a hive in today's online environment. With over a billion users actively creating content and interacting with one another on social networking sites such as Facebook, YouTube, and Instagram, these platforms help to generate this buzz [2]. YouTube is a platform that distinguishes itself from others for allowing users to share videos and interact with them by leaving opinions, favourites, comments, and shares. Especially the comment section provides an interesting forum for users to share their viewpoints, feelings, and even jokes about the film.

The quantity and proportion of likes on a video are important indicators of the creator's popularity and influence. A video that has an abnormally high dislike ratio might be a public relations nightmare for the artist. Sentiment analysis can assist with this. It's the machine's equivalent of interpreting the beehive's hum and finding out the feelings and viewpoints that are concealed in written language.

Machine learning and natural language processing may be used to evaluate YouTube comments and measure the opinions that users have expressed. Thus, we can forecast the like percentage of the video, which is a useful tool for businesses and creators seeking to negotiate the complexities of internet celebrity and income generation. Furthermore, it is even more important to be able to anticipate dislikes based on other data in light of YouTube's recent experiment with suppressing dislike statistical analysis.

We explore the fascinating world of YouTube comments, utilizing sentiment analysis to forecast likes and maybe even figure out the secrets of hidden dislikes. This trip aims to clarify the complex connection between feedback from users and online behaviour, which will eventually enable both producers and viewers to better navigate YouTube's always changing environment. We use machine learning models like Naïve Bayes, SVM, Random Forest, Gradient Boosting, and Logistic Regression in our approach. We evaluate their effectiveness using established statistical measurements and rigorous cross-validation methods. Finally, we compare the results obtained from different models in our experimental analysis.

In Section 2, we provide an overview of previous research conducted in the field. Section 4 outlines the limitations of current systems. Section 5 briefly explains our research methods, and Section 6 details the implementation of our proposed methodology. Results and evaluations are presented in a table in Section 7. Sections 8 and 9 are dedicated to discussing the conclusions derived from our study and outlining the potential future scope of the project.

2. LITERATURE SURVEY:

In 2014, Ritika singh et al.[4] conducted sentiment analysis on 1500 YouTube comments, emphasizing Decision Tree classifiers' significance but noting potential issues like overfitting and limited attribute evaluation.

In 2015, Alberto et al.[5] addressed YouTube comment spam, achieving 96.23% accuracy with TubeSpam, an effective online tool. Pros include a focused approach, but cons involve missing implementation challenges and ethical considerations.

In 2017, Mulholland et al.[6] introduced an educational video recommender system integrating sentiment analysis and gamification. Pros include innovative sentiment analysis and emotion detection, but challenges involve sentiment analysis limitations and the need for further testing.

In 2019, Abbi Nizar Muhammad et al. [7] achieved 87% accuracy in YouTube sentiment analysis using NBSVM, highlighting its precision but noting limited originality and the need for deeper model analysis.

In 2021, Rawan Fahad Alhujaili et al.[8] achieved 88.8% accuracy in Arabic YouTube comments sentiment analysis. Pros include a comprehensive overview, but cons involve limited originality and technique discussion.

In 2021, Rhitabrat Pokharel et al.[9] achieved 86% accuracy classifying YouTube comments into six categories. Pros include a clear research objective but lack detailed hyperparameter tuning and direct classifier comparison.

In 2021, Hayoung OH [10] addressed YouTube spam using ensemble models, achieving high accuracy with Support Vector Machines and Random Forests. Pros involve a three-dimensional classification approach, while cons emphasize the need for further research.

In 2023, Sainath Pichad et al. [11] achieved 95.5% accuracy in YouTube video categorization, addressing challenges in analyzing vast comments. Pros include aiding businesses, while cons involve handling informal language and sentiment analysis complexities.

3. EXISTING SYSTEM

Existing systems mainly cater to traditional text data, lacking adaptability for YouTube comments' unique traits—short, informal, and highly polarized. Limited methodologies exist for YouTube comment classification. Comment Miner, a UC Berkeley-developed system, deploys sentiment analysis and topic modeling. YouTube Lyrics, a commercial tool, offers analytics for channels, including comment trends. Social Blade, another commercial tool, provides analytics for YouTube and social media, aiding in growth tracking and comment analysis. These tools contribute to understanding and categorizing YouTube comments, enhancing viewer engagement and content strategy.

4. LIMITATIONS OF EXISTING SYSTEM:

Existing systems for analysing YouTube comments have a number of limitations. For example, these systems may not be able to accurately classify comments with short and informal text, non-standard language features, or highly polarized sentiment. Additionally, these systems may not be able to identify all of the relevant and engaging comments for a given video. Existing systems for analysing YouTube comments are not well-suited for handling the unique characteristics of YouTube comments, such as their short and informal nature, their use of non-standard language features, and their highly polarized sentiment. Additionally, these systems may not be able to identify all of the relevant and engaging comments for a given video.

5. PROPOSED METHODOLOGY:

The machine learning framework for extracting and classifying user comments on YouTube videos. Our approach combines sentiment analysis and sentence type classification to categorize comments and help YouTubers find comments that can enhance their channel's viewership. Our approach is specifically tailored to the unique

characteristics of YouTube comments. We use a variety of techniques to address the challenges of short, informal text, non-standard language features, and highly polarized sentiment.

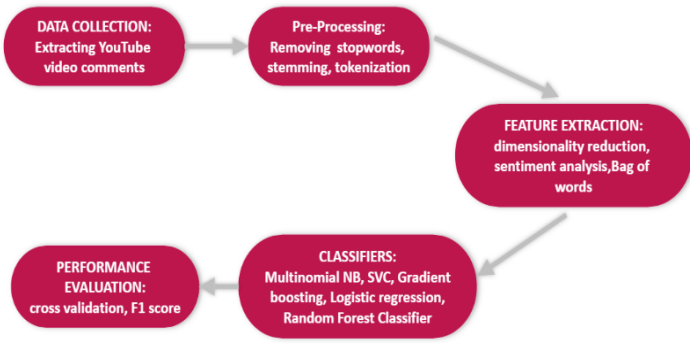

Fig 1: Flow of methodology

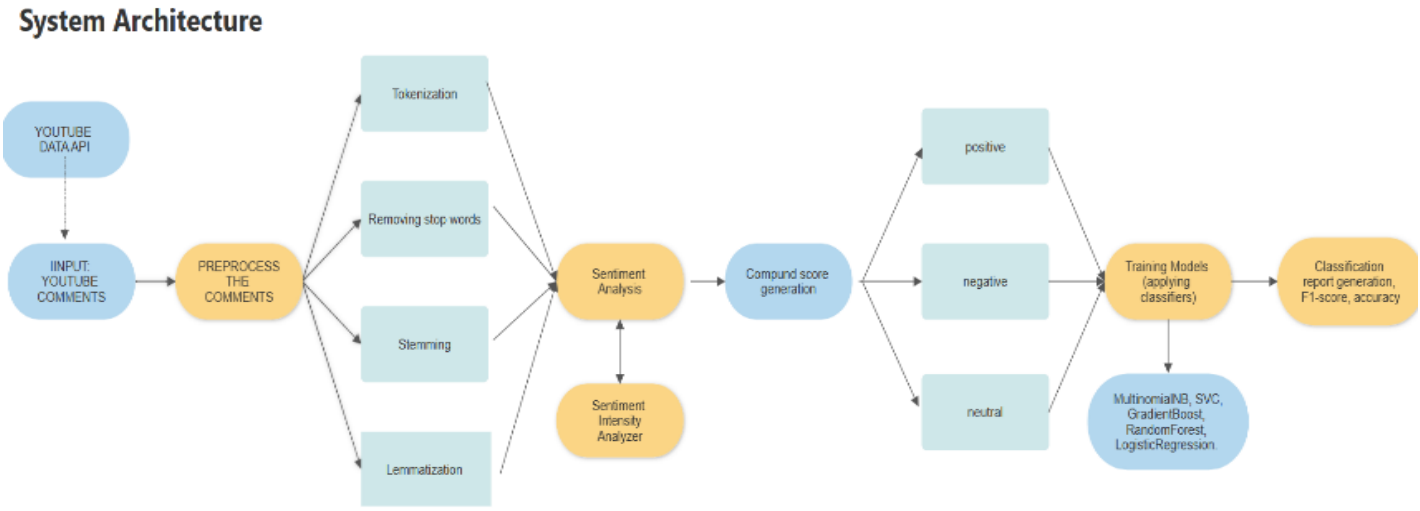

Fig 2: System Architecture .

5.1 DATA COLLECTION: Collecting dataset of YouTube comments. Using web scraping techniques to obtain comments from relevant videos or channels

5.1.1 Extracting Data Using Google Client API:

Utilizing the YouTube Data API, we prompt users for a video ID, collect up to 10,000 comments, and save data in CSV format. Respecting privacy and adhering to YouTube's terms.

5.1.2 Input YouTube Comments:

We are taking the YouTube Comments as input analysing the comments to figure out how people feel about the video and what kind of sentence they are using.

5.2 PREPROCESSING:

Text data preparation involves importing common English stop words, initializing three stemming algorithms (Porter, Lancaster, Snowball) for efficient processing.

5.2.1 Tokenization: Tokenization is the process of breaking down a text into smaller units, such as words, punctuation marks, or sub words. This is a fundamental step in many natural language processing (NLP) tasks, such as machine translation, text summarization, and sentiment analysis.

5.2.2 Removing Stop Words: Removing stop words is the process of filtering out common words from a text, such as articles, prepositions, and conjunctions. This is done to improve the performance of NLP tasks (classification, machine translation, and information retrieval.)

5.2.3 Stemming: Stemming is a technique for reducing words to their root or stem form. This is done by removing prefixes, suffixes, and other affixes from words.

5.2.4 Lemmatization: Lemmatization is a text processing technique that converts words to their base or root form. It is similar to stemming, but it is more sophisticated and accurate. Lemmatization uses a morphological dictionary to identify the part of speech and the base form of the word, which allows it to preserve more information about the word.

5.3 FEATURE EXTRACTION:

Feature extraction is a crucial step in analysing YouTube comments using machine learning. They involve selecting, creating, and transforming attributes (features) from the raw comment data to feed into your machine learning model. We used Sentiment Analysis, Bag of Words Methods.

5.3.1 Sentiment Analysis: sentiment analysis one common approach is to use a lexicon-based approach. This involves using a list of words and phrases that are associated with positive and negative sentiment. The lexicon is then used to score the text, with higher scores indicating more positive sentiment and lower scores indicating more negative sentiment.

5.3.2 Sentiment Intensity Analyser: Sentiment intensity analysers are typically trained on a dataset of labelled text, where each piece of text is labelled with its sentiment and its intensity. The analyzer learns to identify patterns in the text that are associated with different levels of sentiment intensity.

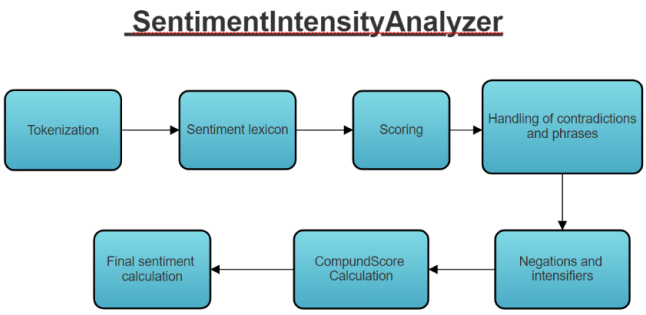

Fig 3: SentimentIntensityAnalyzer workflow

5.3.3 Compound Score generation: Identifying positive and negative reviews: Compound score generation can be used to identify positive and negative reviews. This information can be used by businesses to improve their products and services.

$$\text{Compound Score} = (\Sigma\ (\text{Valence Score} / \sqrt{(\text{Valence Score}^2 + \text{Alpha})})) / n$$

In this formula, Valence Score is the sentiment score of each word. Alpha is a normalization factor (typically a small constant like 0.1) to prevent division by zero. The summation Σ is performed over all the words in the text.

5.4 APPLYING CLASSIFIERS:

Following the pre-processing and feature selection stages, the subsequent phase involves the application of classification algorithms. Numerous text classifiers have been proposed in the existing literature. In our study, we employed five machine learning algorithms, namely Multinomial Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), gradient boosting, and random forest. Furthermore, we integrated a stacking classifier into our methodology for comprehensive analysis.

5.4.1 Multinomial NB:

Multinomial Naive Bayes is a helpful tool for sorting YouTube comments into different sentiment categories. Based on Bayes' theorem, it assumes features are independent, simplifying the handling of text-heavy data. It works well with diverse content found in YouTube comments, efficiently managing large feature spaces. Its adaptability to sparse data, computational efficiency, and suitability for multiclass classification contribute to its effectiveness. The algorithm's probabilistic nature ensures an understandable output, aiding users in gauging the likelihood of a comment belonging to a specific sentiment class. Overall, Multinomial Naive Bayes stands out for its efficiency, simplicity, and effectiveness in handling YouTube comment sentiment analysis.

5.4.2 Support Vector Machine:

Support Vector Machine (SVM) is a supervised learning algorithm for sentiment analysis in YouTube comments. It establishes an optimal hyperplane to categorize sentiments, proving valuable in discerning patterns within text data. SVM is effective in high-dimensional spaces, suitable for sentiment analysis where each word represents a dimension. It aims to find a hyperplane maximizing the margin between sentiment classes, acting as the decision boundary for positive, negative, or neutral sentiments. SVM excels in text data analysis, robust to overfitting, versatile with kernels for nonlinear patterns, and adaptable for binary or multiclass sentiment classification. Its interpretable decision boundaries aid in understanding sentiment nuances in diverse comments.

5.4.3 Logistic regression:

Using Logistic Regression in YouTube comment sentiment analysis predicts a comment's likelihood of belonging to a sentiment class. It's a supervised learning algorithm for binary and multiclass classification, applied to forecast the probability of a comment falling into a specific sentiment category. Pre-processing and feature extraction from YouTube comments involve methods like TF-IDF. Unlike linear regression, Logistic Regression predicts probabilities using the logistic function, transforming output to a probability scale. While effective for linearly separable relationships, it may struggle with complex, nonlinear sentiment expressions in YouTube comments. Advantages include interpretable coefficients for clear feature understanding and efficiency for large datasets, providing insights into confidence levels for improved interpretability in sentiment predictions.

5.4.4 Gradient boosting:

Applying Gradient Boosting, an ensemble learning method, combines weak learners like decision trees to create a robust predictive model. It minimizes errors by iteratively adding learners, focusing on residuals. While individual trees lack interpretability, the ensemble approach offers a holistic interpretation of decision-making. In YouTube comment sentiment analysis, text data is pre-processed, and features are represented using TF-IDF or word embeddings. Gradient Boosting excels in predictive accuracy, handles nonlinear patterns, and resists overfitting. Its adaptability to various scenarios and interpretability of feature importance make it widely used in machine learning, including sentiment analysis for diverse YouTube comments.

5.4.5 Random Forest:

Random Forest, a vital ensemble learning algorithm in YouTube comment sentiment analysis, constructs decision tree ensembles during training. Each tree predicts independently, introducing diversity with bootstrapped subsamples and feature randomization to prevent overfitting. The ensemble's strength lies in aggregating individual predictions, determining the final sentiment by taking the mode of predicted classes. This approach enhances predictive accuracy, generalization, and robustness against overfitting. Random Forest provides insights into significant features and excels in capturing sentiment complexities in YouTube comments. Its ensemble nature ensures reliable predictions even when individual trees overfit, showcasing its effectiveness in sentiment analysis.

5.4.6 Stacking Classifier

The Stacking Classifier, an ensemble learning technique, enhances predictive performance by combining multiple base models. Initialized with defined base models, it employs another Logistic Regression model as the final estimator to aggregate base model predictions. During training, base models make individual predictions, and the final estimator learns to effectively combine them. The Stacking Classifier leverages the diversity of base models, improving generalization and predictive accuracy. Acting as a meta-learner, the final Logistic Regression model optimally weighs predictions from base models. This meta-model excels when different models capture various data aspects, contributing to a robust ensemble approach that outperforms individual models.

5.5 PERFORMANCE EVALUATION:

The implemented classifiers demonstrate distinct levels of accuracy in their predictions. The efficiency of each classifier is gauged based on its accuracy, with the one achieving the highest accuracy being deemed the most effective. Additionally, the evaluation using Root Mean Squared Error (RMSE) provides insight into the precision of the models. In this context, the classifier that attains the lowest RMSE score is considered more accurate, reflecting a smaller margin of error in its predictions. Hence, accuracy and RMSE serve as pivotal benchmarks for the performance assessment of classifiers, guiding the identification of the most reliable and precise predictive models.

6. IMPLEMENTATION:

6.1: Data Extraction

Retrieves comments from YouTube videos using the YouTube API,[12] extracting and organizing them into a CSV file named youtube_comments.csv. This process employs API functionality to acquire comments systematically, facilitating subsequent analysis.

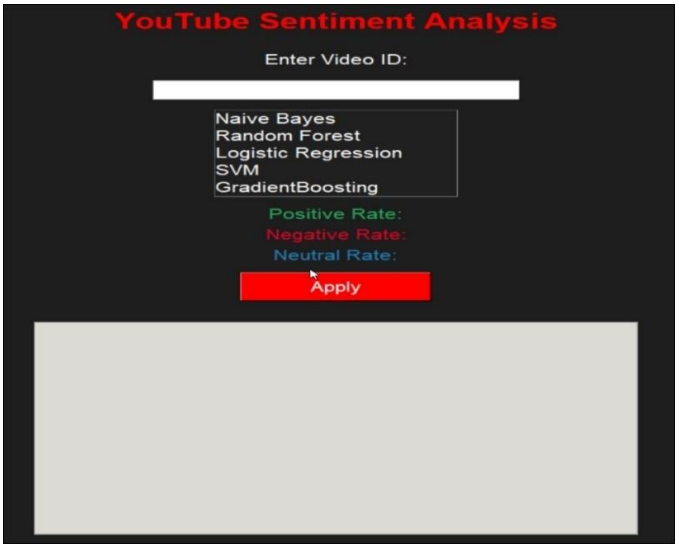

Fig 4: User Interface

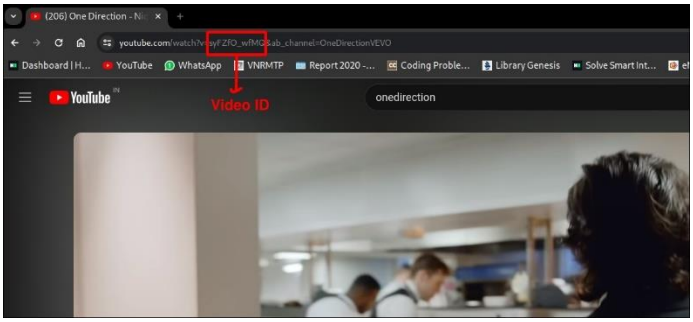

Fig 5: Collecting YouTube video ID

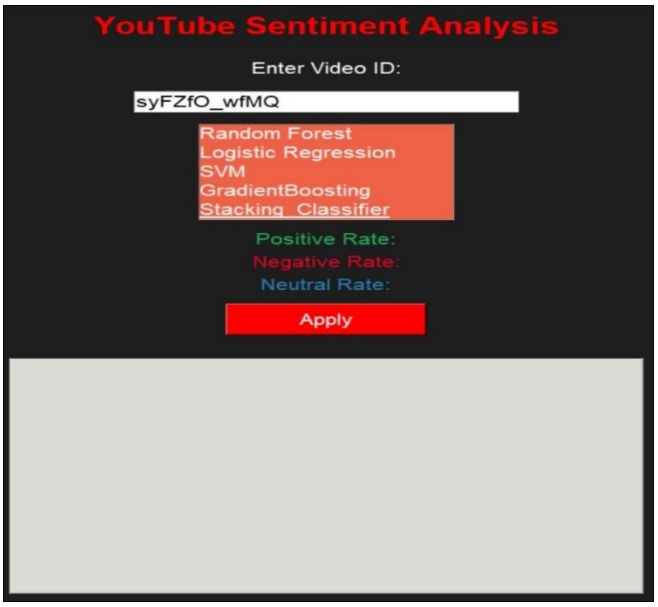

Fig 6: Entering ID and Selecting Algorithms

6.2: Importing Data

Imports the gathered YouTube comments, stored in youtube_comments.csv, into a Pandas DataFrame. This step enables efficient data manipulation and analysis within the Python environment.

	Comment
0	missing old days like seriously
1	2024 anyone?😢❤
2	I knew this song from my older brother when I ...
3	2024 anyone?
4	If u turn it down it's like asmr 😂
...	...
10092	We r only getting older 😔
10093	I am fall in love this song 🎧 so much
10094	Well, i'm gay now.
10095	Reunion please 😍
10096	Besssst

Fig 7: Extracted comments dataset.

6.3: Applying Data Visualization Techniques

Utilizes Matplotlib and Seaborn libraries to generate informative visualizations, including bar plots. These visuals depict the distribution of sentiments derived from the comments, offering a quick overview of sentiment proportions.

	Comment	Positive	Negative	Neutral	Compound	Sentiment
2	love india tamilnadu	0.677	0.0	0.323	0.6369	Positive
3	anyone 2024	0.000	0.0	1.000	0.0000	Neutral
4	love	1.000	0.0	0.000	0.6369	Positive
5	talantlive	0.000	0.0	1.000	0.0000	Neutral
6	anyone 2025	0.000	0.0	1.000	0.0000	Neutral
...	...	...	...	...	...	...
10088	bring tear eye remind golden old day one direc...	0.318	0.0	0.682	0.6369	Positive
10089	anyone 2023	0.000	0.0	1.000	0.0000	Neutral
10090	650 million view congratulation onedirection d...	0.433	0.0	0.567	0.6249	Positive
10091	song sound like 90 song	0.385	0.0	0.615	0.3612	Positive
10092	0041 goosebump	0.000	0.0	1.000	0.0000	Neutral

Fig 8: Sentiment values of comments.

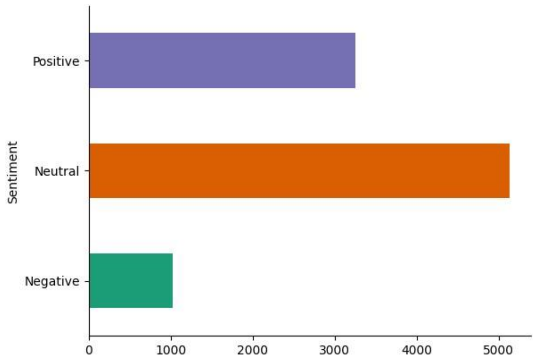

Fig 9: Graph categorizing comments.

6.4: Data Cleaning

Processes the comment data by eliminating noise elements such as punctuation, URLs, special characters, and emojis. Additionally, it addresses missing values (NaN) by removing them, ensuring cleaner and standardized text for subsequent analysis.

6.5: Data Preprocessing Techniques

Prepares the data for analysis by encoding sentiment labels, performing text preprocessing (e.g., stemming, lemmatization, dimensionality reduction), and balancing class distribution through resampling techniques. These steps contribute to improved model training and accuracy.

	Sentence	Sentiment
2	love india tamilnadu	2
3	anyone 2024	1
4	love	2
5	talantlive	1
6	anyone 2025	1
...	...	...
10088	bring tear eye remind golden old day one direc...	2
10089	anyone 2023	1
10090	650 million view congratulation onedirection d...	2
10091	song sound like 90 song	2
10092	0041 goosebump	1

Fig 10: Preprocessed comments dataset.

6.6: Splitting and Organizing the Data

Divides the dataset into training and testing subsets using the train_test_split method. While no explicit normalization techniques are implemented in this code segment, it sets the foundation for subsequent analysis.

6.7: Model Training Using Classifiers

Utilizes a variety of classifiers (Naive Bayes, SVM, Gradient Boosting, Logistic Regression, Random Forest) to train the sentiment analysis model. This approach evaluates and compares their performance in sentiment classification.

6.8: Analyzing the Results

Following model training, assesses and presents the performance metrics (accuracy, precision, recall, F1 score) for each classifier. This analysis provides valuable insights into the effectiveness of these classifiers in discerning sentiments from YouTube comments.

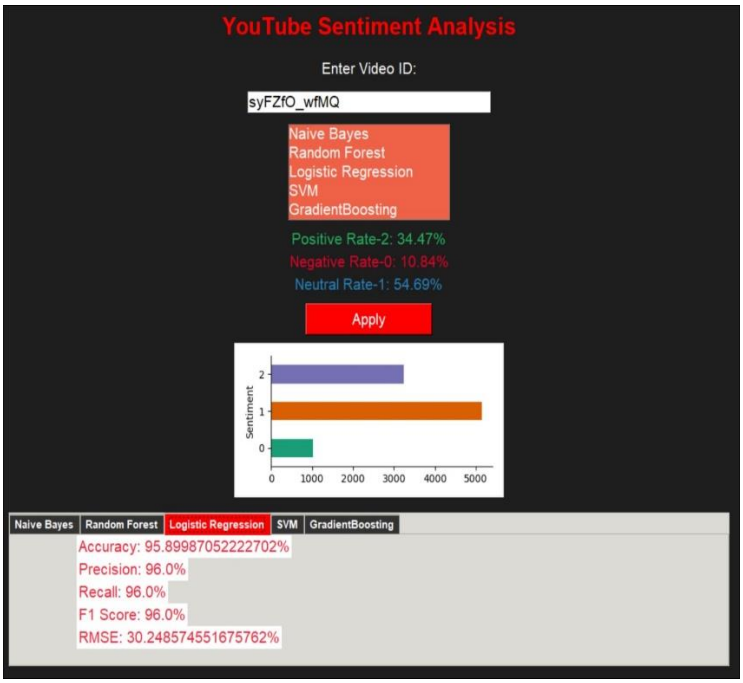

Fig 11: Obtained results in GUI

7. RESULTS:

The research paper provides an in-depth sentiment analysis of YouTube comments through various machine learning classifiers, including Naive Bayes, SVM, Random Forest, Logistic Regression, and Gradient Boosting. The study incorporates data preprocessing, feature engineering, and resampling methods. The outcomes highlight the classifiers' accuracy, precision, recall, and F1 score, offering significant insights into sentiment analysis applications.

S No.	Classifier Models	Accuracy	Precision	Recall	F1 Score	RMSE
1	Naive bayes	68.64%	74%	69%	68%	871.9
2	Support Vector Machine	97.77%	98%	98%	98%	237.7
3	Random Forest	97.66%	98%	98%	98%	179.9
4	Gradient Boosting	68.04%	78%	68%	67%	662.8
5	Logistic regression	95.89%	96%	96%	96%	302.4
6	Stacking Classifier	95.95%	96%	96%	96%	229.4

Table 1: Accuracies Obtained for each Model.

8. CONCLUSION:

In conclusion, our project advances content creation strategies by employing a robust sentiment analysis framework for YouTube comments. Utilizing machine learning classifiers, including Naive Bayes, SVM, Logistic Regression, Random Forest, Gradient Boosting, and Stacking, the result shows that SVM(97.7%) performs better than the other classifiers. Following SVM, Random Forest(97.6%) performs well. In the case of the macro average, the performance of SVM and

Random Forest classifiers is same(98%) while computing F1 score. Notably Random Forest(179.9) obtained the lowest RMSE value. After Random Forest, Stacking Classifier with 229.4. The user-friendly interface facilitates input of video IDs and algorithm selection. The modular code structure allows adaptability for future sentiment analysis research. Our project bridges the gap between creators and audiences, improving the efficiency and insightfulness of feedback analysis in the dynamic world of online content..

9. FUTURE SCOPE:

This project's future scope include content moderation, targeting the identification and removal of spam, hate speech, and harmful content. We aspire to predict trending videos, deliver personalized recommendations to users based on preferences, provide insights for content creators to comprehend audience responses. Furthermore, the techniques developed have the potential to extend beyond YouTube, offering benefits to creators on diverse social media platforms where comments are prevalent. This broader impact aims to contribute significantly to the evolving landscape of digital content creation.

10. REFERENCES:

[1] S. Aiyar, N. Shetty, N-gram assisted "YouTube spam comment detection", 2018 *International Conference on Computational Intelligence and Data Science*; ICCIDS 2018.

[2] Meeyoung Cha et al.; "Comparing and Combining Sentiment Analysis Methods"; *IEEE*; 2020

[3] A. Madden, I. Ruthven, D. McMenemy, "A classification scheme for content analyses of youtube video comments"; *Journal of documentation* ; 2013.

[4] Ritika Singh, Ayushka Tiwari**; "**YouTube comments sentiment analysis"; *International Journal of Scientific Research in Engineering and Management (IJSREM)*; May 2021.

[5] Alberto Tulio C. Alberto, Johannes V. Lochter, Tiago A. Almeida ; "TubeSpam: Comment Spam Filtering on YouTube" ; 2015 *IEEE 14th International Conference on Machine Learning and Applications.*

[6] Eleanor Mulholland , Paul Mc Kevitt1, Tom Lunney and Karl-Michael Schneider; "Analysing Emotional Sentiment in People's YouTube Channel Comments"; School of Creative Arts and Technologies, Ulster University; 2017.

[7] Abbi Nizar Muhammad et al; "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier", 2019; Indonesia.

[8] Rawan Fahad Alhujaili and Wael M.S. Yafooz; "Sentiment Analysis for YouTube Videos with User Comments : Review" ; *Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021) IEEE Xplore Part Number: CFP21OAB-ART; ISBN*: *978-1-7281-9537-7*; 2021.

[9] Rhitabrat Pokharel and Dixit Bhatta; "Classifying YouTube Comments Based on Sentiment and Type of Sentence"; *Creative Commons Attribution-NonCommercialNoDerivs 4.0 International (CC-BY-NC-ND4.0) license*; 2021.

[10] HAYOUNG OH ; "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model" ; *IEEE*; 2021.

[11] Sainath Pichad, Sunit Kamble et al.; "Analysing Sentiments for YouTube Comments using Machine Learning"; *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*; 2023.

[12] Sanjeepan Sivapiran et al. ; "Sentiment Analysis in Dravidian Code-Mixed YouTube Comments and Posts" ; *ISSN 1613-0073 CEUR Workshop Proceedings* ;2021.

[13] Apoorv Agarwal B et al.; "Sentiment Analysis of Twitter Data" ; *Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38,Portland, Oregon,c 2011 Association for Computational Linguistics.*

[14] Abhilasha Sancheti et al.; *"A review on sentiment analysis techniques and their applications in social media platforms."*; International Journal of Computer Applications, 2018.

[15] Abdullah Alamoodi et al. "*Sentiment Analysis and Its Applications in Fighting COVID-19 and Infectious Diseases: A Systematic Review*"; Elsevier; 2020.