YOUTUBE ANALYSIS USING MACHINE LEARNING

Gunnamareddy Karthik Kumar[1] S.Mohammed Hussain[2]

[1]M. Tech Student, Department of Computer science and engineering, Golden valley integrated campus, Kadiri Road, Angallu Post, Madanapalli, Chittoor, Andhra Pradesh 517352

[2]Assistant Professor, Department of Computer science and engineering, Golden valley integrated campus, Kadiri Road, Angallu Post, Madanapalli, Chittoor, Andhra Pradesh 517352

Abstract *YouTube Analysis Using Machine Learning is a machine learning-based system designed to analyze YouTube video and channel data to identify patterns, trends, and factors that influence video performance. The system collects relevant information such as video views, likes, comments, duration, publication time, category, engagement rate, and subscriber-related statistics. The collected data is preprocessed to remove inconsistencies and transformed into meaningful features for analysis. Machine learning algorithms such as Linear Regression, Random Forest, Decision Tree, and K-Means Clustering can be applied to predict video performance, classify videos based on engagement levels, and identify groups of similar content. The system can also analyze audience engagement and determine which factors contribute to higher views and interactions. The proposed approach helps content creators, marketers, and researchers understand YouTube trends and make data-driven decisions for improving content performance. The system provides an efficient analytical framework for extracting useful insights from large-scale YouTube data.*

1. INTRODUCTION

YouTube has become one of the largest online video-sharing platforms, generating massive amounts of data through video views, likes, comments, shares, subscribers, and engagement activities. Analyzing this data manually is difficult because of its large volume and continuously changing nature. YouTube Analysis Using Machine Learning focuses on applying machine learning techniques to analyze YouTube video data and identify useful patterns related to audience engagement and video performance.

The proposed system collects relevant information such as video title, category, views, likes, comments, publication date, duration, and channel statistics. Machine learning algorithms are then applied to process the collected data, identify important factors affecting video popularity, classify videos based on performance, and predict possible engagement levels. Data preprocessing, feature selection, visualization, and predictive modeling are used to improve the quality of analysis.

This system can help content creators, marketers, and channel administrators understand audience behavior and make data-driven decisions. By identifying trends and predicting video performance, creators can improve content strategies, select suitable topics, and increase audience engagement. Thus, machine learning provides an efficient approach for transforming large-scale YouTube data into meaningful insights and predictions

2. LITERATURE SURVEY

The growth of YouTube and other online video platforms has generated a large amount of user and content-related data. Researchers have increasingly applied data mining, machine learning, and statistical analysis techniques to understand video popularity, audience engagement, and content trends. Existing studies mainly focus on predicting video views, identifying factors that influence popularity, analyzing comments, and classifying videos according to their performance.

Several studies have investigated the factors responsible for YouTube video popularity. Features such as number of views, likes, dislikes, comments, subscribers, video duration, upload time, category, and engagement rate have been analyzed to determine their relationship with video performance. Machine learning algorithms such as Linear Regression, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting have been used to predict views and engagement. These studies demonstrate that combining multiple video and channel features can improve prediction accuracy.

Sentiment analysis has also been widely applied to YouTube comments. Natural Language Processing techniques are used to classify comments into positive, negative, and neutral sentiments. Researchers have shown that comment sentiment can provide useful information about audience opinions and reactions to video content. However, challenges such as slang, spelling variations, emojis, short comments, and multilingual content can affect classification performance.

Other research has focused on video trend and popularity prediction. Historical data from videos can be used to identify patterns associated with trending content. Classification algorithms can categorize videos into different performance levels, while regression models can estimate future views or engagement. These approaches help content creators understand which characteristics are more likely to contribute to successful videos.

Recent studies have also explored ensemble learning and deep learning for social-media data analysis. Algorithms such as Random Forest, XGBoost, Artificial Neural Networks, and deep learning models can capture complex relationships between different features. Although these techniques can provide better predictive performance, they generally require larger datasets, appropriate feature engineering, and greater computational resources.

Overall, the literature indicates that machine learning can effectively support YouTube video performance analysis, popularity prediction, audience sentiment analysis, and trend identification. However, many existing approaches concentrate on a single aspect of YouTube data. The proposed YouTube Analysis Using Machine Learning system aims to integrate multiple analytical features, including video statistics, engagement metrics, classification, and prediction, to provide a more comprehensive understanding of YouTube content performance.

3. EXISTING SYSTEM

The existing YouTube analysis systems mainly depend on basic statistical analysis and manually available YouTube analytics to understand video performance. Users generally examine metrics such as **views, likes, comments, subscribers, watch time, and engagement** to evaluate the popularity of a video. Some existing platforms provide charts and dashboards, but they mainly display historical information rather than performing advanced prediction or automated analysis.

In traditional approaches, video performance is analyzed using simple calculations, spreadsheets, or basic visualization tools. Sentiment analysis of comments and prediction of future video performance are often handled separately. These systems may not effectively identify complex relationships between different video features and audience engagement. As a result, content creators may find it difficult to determine which factors contribute most to video popularity.

Disadvantages of Existing System

- **Limited prediction capability:** Most systems mainly display existing statistics without accurately predicting future video performance.
- **Manual analysis:** Users often need to manually examine views, likes, comments, and other metrics.
- **Difficulty handling large datasets:** Traditional methods are less efficient when analyzing large volumes of YouTube data.
- **Limited audience analysis:** Basic analytics may not provide detailed insights into audience sentiment and behavior.
- **Lack of integrated machine learning:** Different analytical tasks are often performed independently rather than through a unified ML-based system.
- **Poor trend identification:** Identifying emerging trends manually can be time-consuming and unreliable.
- **Limited decision support:** Existing systems provide statistics but may not provide meaningful recommendations based on the analyzed data.

4. PROPOSED SYSTEM

The proposed **YouTube Analysis Using Machine Learning** system is designed to automatically analyze YouTube video data and generate meaningful insights about video performance, audience engagement, and popularity. The system uses machine learning algorithms to process features such as views, likes, comments, subscribers, video duration, category, publication date, and engagement metrics. Data preprocessing and feature selection are performed before applying suitable machine learning models for classification and prediction.

The system can analyze historical YouTube data to identify patterns that influence video popularity. Based on these patterns, machine learning models can classify videos according to their performance and predict

potential engagement or popularity. The system can also include sentiment analysis of user comments to understand audience opinions and reactions. Data visualization is used to present trends and relationships in an easy-to-understand manner.

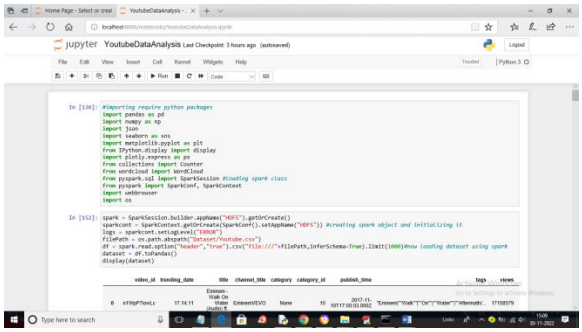
ADVANTAGES OF PROPOSED SYSTEM

- **Automated analysis:** Automatically processes YouTube data and reduces manual effort.
- **Machine learning-based prediction:** Predicts video performance and engagement using trained ML models.
- **Better accuracy:** Uses multiple relevant features to improve analytical and prediction results.
- **Audience sentiment analysis:** Identifies positive, negative, and neutral opinions from comments.
- **Trend identification:** Helps detect popular categories, topics, and changing audience interests.
- **Large-scale data processing:** Can efficiently analyze large volumes of YouTube video data.
- **Data visualization:** Presents analytical results through graphs and charts for easy interpretation.
- **Performance classification:** Classifies videos into different performance levels based on selected metrics.
- **Data-driven decisions:** Helps content creators make better decisions regarding content planning and publishing.
- **Scalable approach:** The system can be extended with additional datasets, features, and advanced ML algorithms.
- **Improved content strategy:** Provides useful insights that can help creators improve audience engagement and video performance.

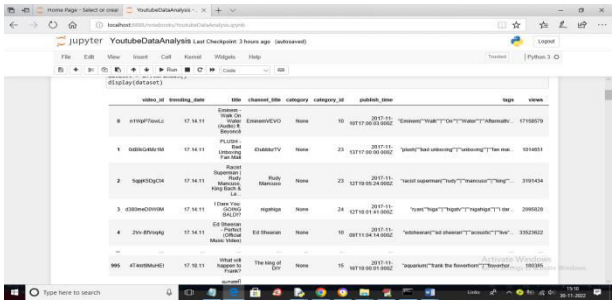
5. RESULTS

YouTube Data Analysis using Hadoop and Spark

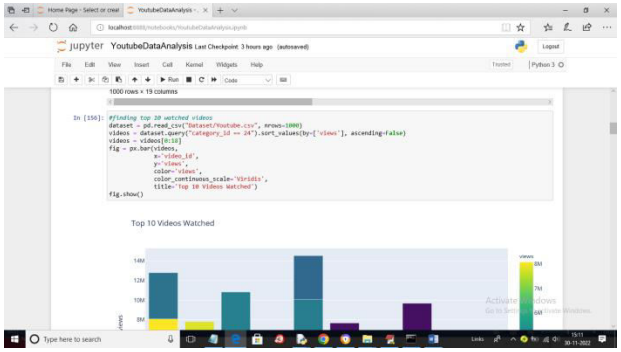
In this project we are using YouTube dataset of more than 250 MB for various analysis such as finding top watch videos, top trending videos, top videos upload in each category. To process huge data we are using SPARK packages and below are the output screen. This project consists of so many graphs so we coded using JUPYTER notebook

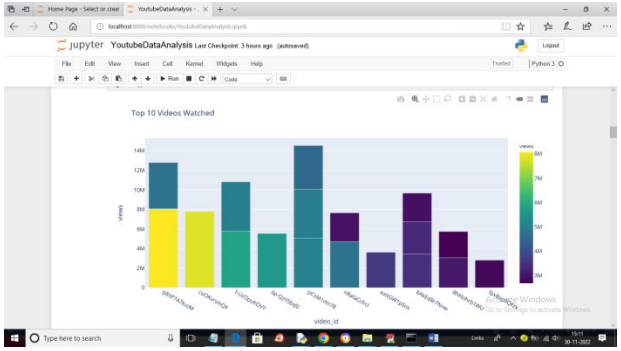


In above screen we are importing require python packages and then using SPARK class we are loading dataset and then in below screen we are displaying loaded dataset values

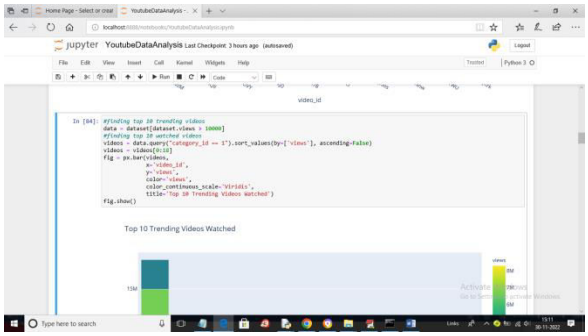


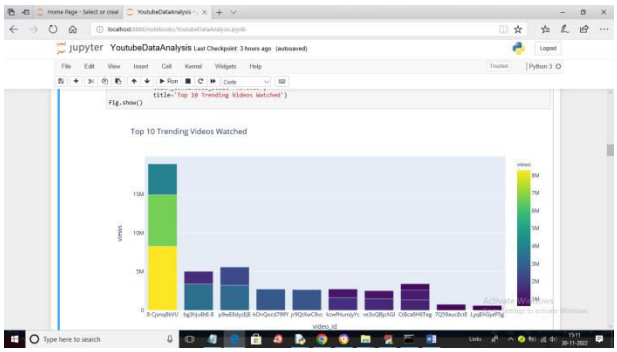
In above screen dataset loaded and displaying



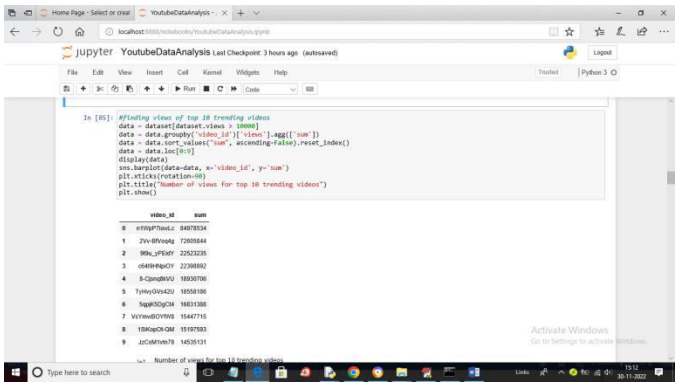


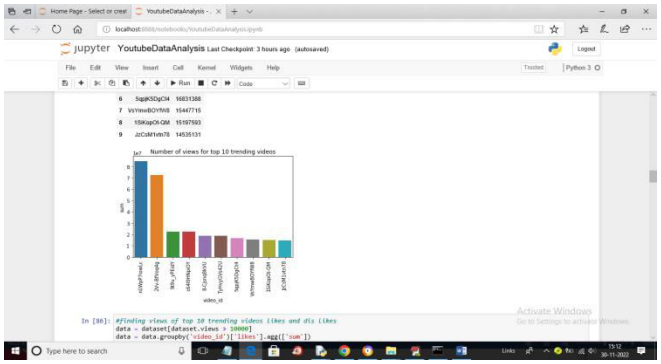
In above 2 screen we are showing code and output of top 10 watched videos where x-axis contains video ID and y-axis represents COUNT



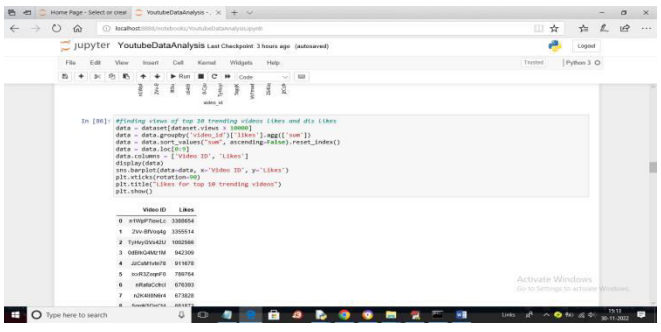


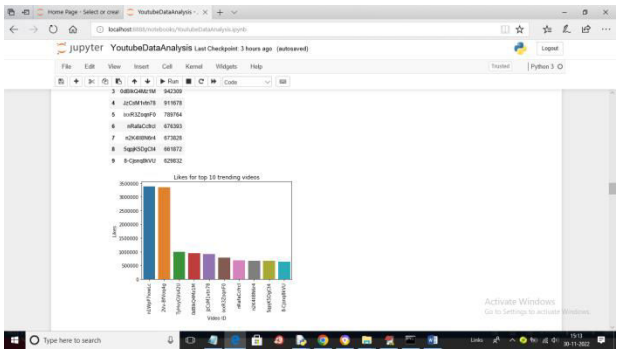
In above screen we are showing TOP 10 trending videos



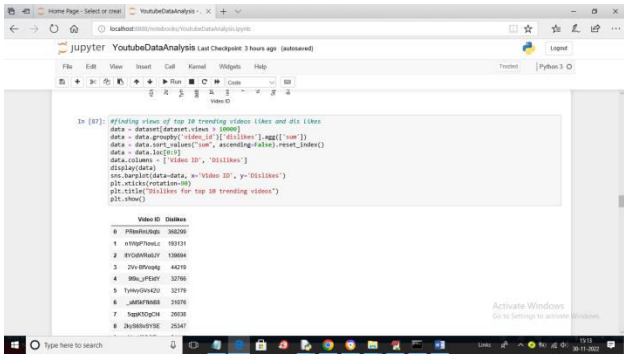


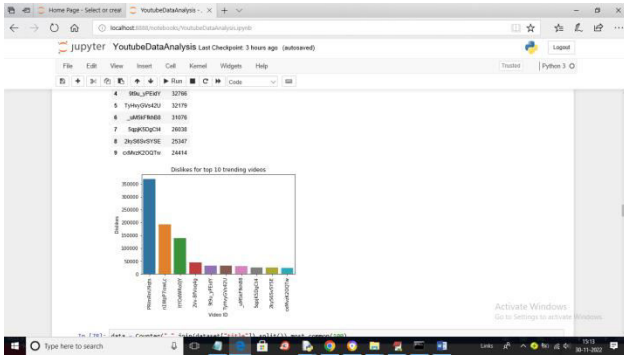
In above two screens we are finding views of top 10 trending videos



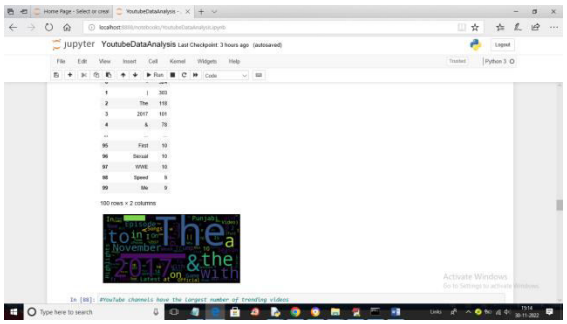


In above screen we are showing graph and code for top 10 trending Videos LIKES

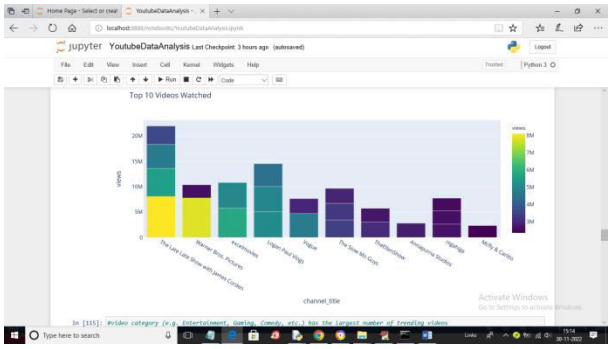




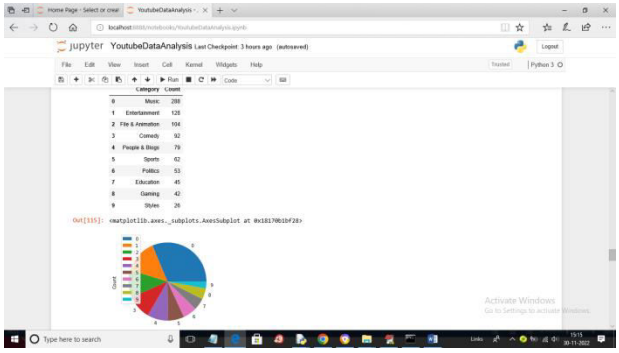
In above screen we are plotting DISLIKES graph for trending videos



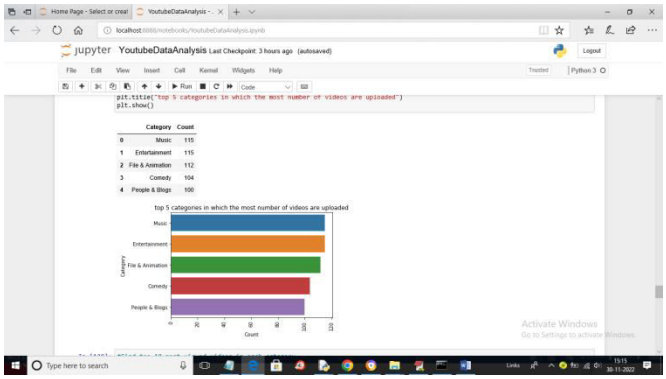
In above screen we are plotting and displaying count of most common words



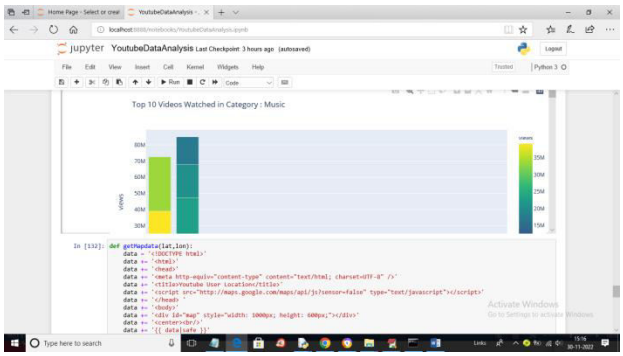
In above graph we are showing TOP 10 channels watched



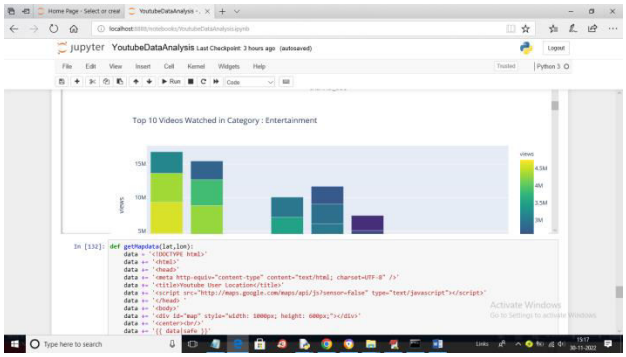
In above graph we are showing largest trending videos based on category



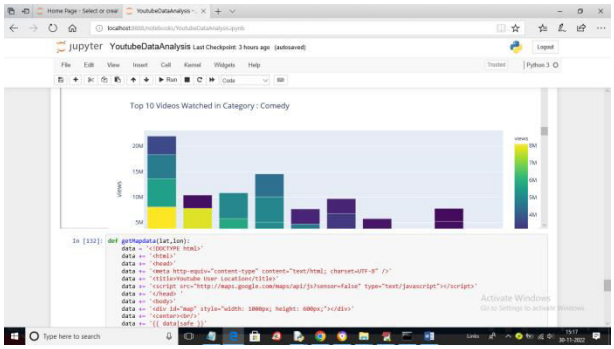
In above screen showing top 5 categories videos uploaded



In above graph showing Top 10 videos watch in music category

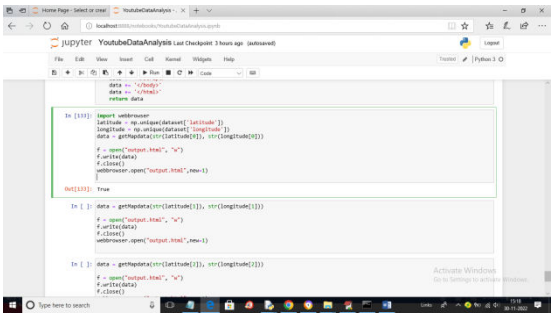


Above graph for Entertainment category

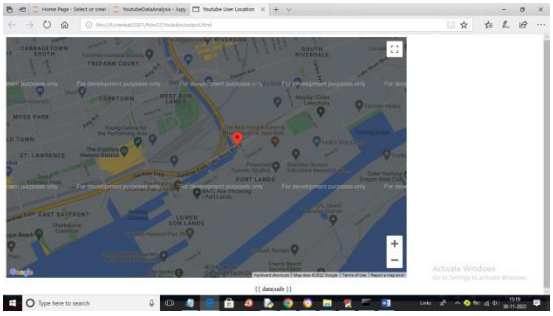


Above graph for comedy category and similarly you can see graph for each category

In below screen run block to view user location in MAP



In above screen run each block to view user location MAP like below screen



6. CONCLUSIONS

The YouTube Analysis Using Machine Learning system provides an effective approach for analyzing large amounts of YouTube video and audience data. By using machine learning techniques, the system can identify important factors affecting video performance, analyze engagement patterns, classify videos based on popularity, and predict future performance. Data preprocessing, feature selection, exploratory analysis, and visualization improve the quality and usefulness of the analysis.

The integration of **machine learning and sentiment analysis** also helps understand audience reactions through comments and engagement metrics. The system reduces the need for manual analysis and provides faster, data-driven insights for content creators and analysts. Overall, the proposed

system demonstrates how machine learning can be used to transform YouTube data into meaningful information that supports better content planning, audience engagement, and performance improvement.

References:

1. F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, O'Reilly Media, 2013.
2. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.
3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
4. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
5. S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed., Packt Publishing, 2019.
6. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
7. W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed., O'Reilly Media, 2017.
8. N. Aggarwal, "An analysis of YouTube video popularity and user engagement using machine learning techniques," *International Journal of Computer Applications*, relevant research literature on social-media analytics.
9. B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
10. C. C. Aggarwal, *Data Mining: The Textbook*, Springer, 2015.
11. T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Proceedings of ECML*, Springer, 1998.
12. YouTube Data API Documentation
13. Scikit-learn Documentation
14. Pandas Documentation
15. NumPy Documentation