Analyzing the Factors Affecting YouTube Viewership Using PCA and SVM

Nigar Alizada
Software Engineering Laboratory
The University of Osaka
Suita, Osaka, Japan
nigar.alizade00@gmail.com

Kojima Yutaka
Big Data Engineering Laboratory
The University of Osaka
Suita, Osaka, Japan
yutayuta540@gmail.com

Yamasaki Aoto
System Engineering Laboratory
The University of Osaka
Suita, Osaka, Japan
aoto2002@gmail.com

Nalishuwa Chama Joshua
Mobile Computing Laboratory
The University of Osaka
Suita, Osaka, Japan
c-nalishuwa@mc.net.ist.osaka-u.ac.jp

***Abstract*—In an era where numerous streaming platforms, particularly YouTube, have emerged, there has been a significant increase in content creation. As the demand for views arise there is a need to analyse factors that influence viewership. This study investigates key factors that influence the number of views on YouTube videos. We analyze video metadata, such as comments, likes, descriptions, view count, etc., using Principal Component Analysis (PCA) and Support Vector Machines (SVM). The aim is to identify the most impactful features and develop predictive models for viewership trends. By examining data from ten distinct YouTube channels, we provide insights that can assist content creators in optimizing their videos for better reach.**

***Index Terms*—YouTube, Principal Component Analysis, Support Vector Machine, Viewership, Social Media Analytics**

I. Introduction

With over 500 hours of video content uploaded every minute, YouTube has become one of the largest and most influential content-sharing platforms globally. As content creators and businesses increasingly rely on YouTube for visibility and monetization, understanding what drives video viewership has become both a strategic and analytical challenge. While some videos achieve viral success, others receive minimal attention despite similar content quality, making it essential to investigate the factors that influence viewership.

In this paper, we address this gap by analyzing multiple metadata attributes, including thumbnails, tags, descriptions, view count, and others, using Principal Component Analysis (PCA) and Support Vector Machines (SVM).

PCA is a widely used dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while retaining the most significant variance in the dataset [1]. This process is especially beneficial in machine learning tasks where high dimensionality may lead to overfitting or increased computational cost. SVM, on the other hand, are supervised learning models that construct hyperplanes to effectively classify data into distinct classes with maximum margin [2]. The combination of PCA for feature reduction and SVM for classification has proven effective in various applications, including image recognition and bioinformatics [3]. PCA reduces redundancy and noise, while SVM leverages the most informative components to enhance classification accuracy.

With these insights, we can directly support content creators, digital marketers, and platform engineers by providing a clearer understanding of which content features most effectively capture viewer attention. Potential applications include automated content optimization tools, recommendation system enhancement, and more targeted digital marketing strategies.

The remainder of this paper is structured as follows. Section II reviews related work in YouTube analytics and machine learning methods for social media content analysis. Section III details our methodology, including data collection, data preprocessing, PCA, and SVM modeling. Section IV presents our results and analysis. Section V discusses the implications of our findings and suggestions for future research directions. Finally, Section VI concludes this paper with a summary.

II. Related Work

Referencing Jang et al. [4], prior work has demonstrated that visual and textual metadata influence video engagement. Halim et al. [5] analyzed content-agnostic features influencing video popularity on YouTube across seven regions. Using machine learning classifiers, including SVM, they identified key metadata such as titles, descriptions, video duration, and view count as significant predictors. However, their work does not emphasize dimensionality reduction techniques like PCA, nor does it focus on optimizing SVM performance using a reduced feature set. Our study extends this by integrating PCA explicitly to enhance SVM prediction accuracy and interpretability, providing clearer insights into which metadata attributes are most critical.

Chen and Chang [6] developed a model for early prediction of YouTube video popularity using a large set of features and applied PCA for dimensionality reduction before feeding data into an SVM classifier. Their study focuses on early lifecycle prediction, emphasizing uploader information and early view trajectory. Our work differs by concentrating on static metadata attributes (e.g., comments, likes, descriptions, view count) rather than time-dependent metrics, making it applicable even before video publication.

Kong et al. [7] explored viral video prediction using feature analysis, PCA, and SVM regression. While their focus was primarily on early view counts and engagement metrics to predict eventual popularity, their approach emphasizes regression over classification. Our work shifts focus to classify high, medium and low viewership based on static metadata, which is more actionable for content creators aiming to optimize prepublication strategies.

Jeon et al. [8] proposed a hybrid machine learning framework combining PCA with various classifiers, including SVM, to predict the popularity of newly released content across streaming platforms. Their research emphasizes volatility and the use of ensemble methods. By contrast, our study narrows the scope specifically to YouTube and prioritizes a simpler, more interpretable PCA + SVM pipeline without ensemble methods, aiming for practical usability in academic and industry contexts.

Nisa et al. [9] benchmarked several machine learning models for YouTube video popularity prediction, noting that SVM's performance improves with PCA-preprocessed features. While they ultimately favor XGBoost, their work reinforces the relevance of PCA + SVM as a viable baseline. Our contribution emphasizes reproducibility and accessibility by focusing exclusively on PCA-enhanced SVM, alongside a curated dataset from ten YouTube channels across distinct categories to validate the generalizability of our findings.

III. Methodology

A. Data Collection

Using the YouTube Data API v3, metadata was collected from 10 YouTube channels across various categories. The following video-level features were collected: title, thumbnail URL, description length, tags, published date and time, view count, like count, comment count, duration, and category ID. Each video was also linked with its channel-level data: title, subscriber count, channel age, and total number of uploads. The data for each video was stored in CSV format for further processing.

B. Data Preprocessing and Feature Engineering

After collecting metadata from each YouTube channel, the dataset underwent two stages of preprocessing: an initial sampling phase, followed by structured feature engineering to prepare the data for machine learning.

1) Initial Sampling: After retrieving all videos from every channel via the API, we combined them into a single dataset and then partitioned the full collection into three equal groups—top 33.3%, middle 33.3%, and bottom 33.3%—based on view counts to ensure balanced representation across performance levels. This step resulted in a total of 5479 videos. The rationale was to capture the full spectrum of video performance—high-performing, average (middle-performing), and low-performing examples.

2) Type Conversion and Missing Value Handling: All numerical fields such as view count, like count, comment count, and subscriber count were explicitly converted to numeric types. Missing values in engagement metrics were handled by replacing them with zeros, ensuring no disruptions during model training.

3) Time Parsing and Temporal Features: The video's published date, originally stored in ISO 8601 format, was parsed into multiple temporal features, including:

- **Upload Hour** (0–23)
- **Day of the Week** (1–7)
- **Month of Upload**

These features were used to identify patterns in audience engagement related to upload timing.

4) Derived Metrics and Normalized Features: Several new features were engineered to normalize and contextualize the raw metrics:

- **Popularity Score**:

  $$\text{Popularity Score} = \text{View Count}$$

  This metric accounts for channel size and enables fairer comparison across large and small creators.
- **Like Rate** and **Comment Rate**:

  $$\text{Like Rate} = \frac{\text{Likes}}{\text{View Count}}, \quad \text{Comment Rate} = \frac{\text{Comments}}{\text{View Count}}$$

  These engagement metrics reflect viewer interaction quality, not just volume.
- **Description Length**: The number of characters in the video description, used as a proxy for content richness.
- **Tag Count**: The number of tags assigned to each video, serving as a proxy for metadata richness.

These preprocessing steps resulted in a clean, structured dataset suitable for dimensionality reduction via PCA and classification using SVM.

C. Dimensionality Reduction and Modeling

1) Principal Component Analysis (PCA): To reduce the dimensionality of the dataset and identify the most informative features, we applied Principal Component Analysis (PCA) to a curated set of 13 numeric features. These included: Duration (seconds), Description Length, Tags Count, Title Length, Channel Subscriber Count, Channel Age (days), Total Number of Channel Uploads, Published year, Published Month, Day of the Week, Published Hour, Like Rate, and Comment Rate.

The PCA was performed with a dynamic component threshold to capture at least 95% of the variance. This resulted in 11 principal components, which together explained 97% of the total variance in the data. These components served as input to both classification and regression models.

2) SVM Classification: We used a Support Vector Machine (SVM) classifier to predict a video's popularity category. The target variable was derived by dividing videos into three quantile-based categories based on their view counts:

$$\text{Popularity Category} = \begin{cases} \text{Low} & \text{if in bottom } 33.3\% \\ \text{Medium} & \text{if in middle } 33.3\% \\ \text{High} & \text{if in top } 33.3\% \end{cases}$$

The SVM classifier used a radial basis function (RBF) kernel and was trained on 80% of the data, with the remaining 20% used for testing. Model performance was evaluated using accuracy and a classification report detailing precision, recall, and F1-score for each class.

3) SVM Regression: In addition to classification, we used SVM regression to predict a continuous *popularity score*. This score was defined as the log-transformed view count, capturing the skewed nature of YouTube view distributions:

$$\text{Log Popularity Score} = \log(1 + \text{View Count})$$

An SVM regressor with an RBF kernel was trained and tested using an 80/20 split. Model performance was evaluated using mean squared error (MSE) and the coefficient of determination (R^2).

4) Model Visualization and Insights: We visualized the PCA results through explained variance plots and scatter plots in the principal component space. For SVM classification, a confusion matrix and precision/recall bar charts were generated. In the regression task, scatter plots of predicted vs. actual scores and residual histograms provided insight into model performance.

In total, this PCA + SVM pipeline enabled both interpretability of important video features and practical predictive modeling of video popularity on YouTube.

IV. Results and Analysis

A. Dataset Summary

A total of 5,479 videos were analyzed across ten YouTube channels. Key descriptive statistics are summarized below:

- **Average View Count**: 20.8 million
- **Maximum View Count**: 1.58 billion (from MrBeast)
- **Average Popularity Score**: 0.2997
- **Maximum Popularity Score**: 13.2135
- **Average Video Duration**: 13.0 minutes

B. Principal Component Analysis (PCA)

PCA revealed that 11 components explained 97% of the variance in the feature space. 11 components were chosen so as to maintain a cumulative variance of over 95%. The first principal component (PC1) was found to be most influenced by:

- Published Year
- Channel Age
- Like Rate
- Description Length

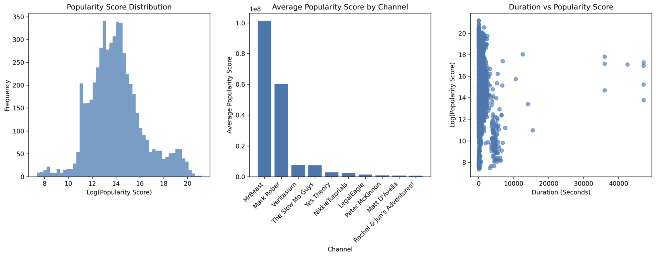

Fig. 1. Popularity Distribution Scores

- Subscriber Count

These findings suggest that PC1 interprets a combination of temporal factors and engagement quality.

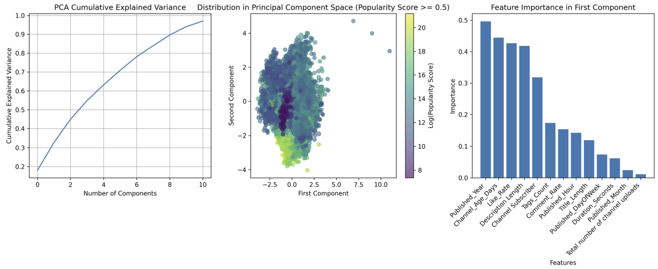

Fig. 2. Cumulative Explained Variance by Principal Components

C. SVM Performance

1) Classification: The SVM classifier achieved an accuracy of **70.2%** when predicting video popularity categories (Low, Medium, High). Performance was higher in distinguishing high and low popularity videos, while medium popularity proved more ambiguous.

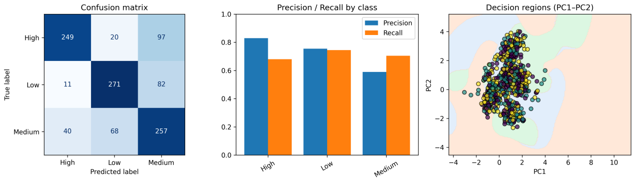

Fig. 3. SVM Classification Results: Confusion Matrix and Precision/Recall

2) Regression: For predicting continuous popularity scores, the SVM regressor achieved an R^2 score of **0.507**, indicating moderate predictive power.

D. Top Performing Channels

Table I shows the top three channels ranked by average popularity score:

V. Discussion

Our findings reveal that temporal factors, such as publishing year and channel age, are key drivers of YouTube video popularity. This observation is consistent with prior research showing that meta-level features like upload timing

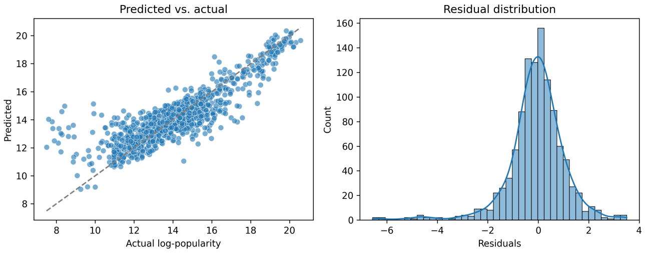

Fig. 4. SVM Regression: Predicted vs. Actual and Residuals

TABLE I
TOP 3 CHANNELS BY AVERAGE POPULARITY SCORE

Rank	Channel	Avg. Score
1	Mark Rober	0.8678
2	The Slow Mo Guys	0.4934
3	Rachel & Jun's Adventures!	0.4905

and channel maturity significantly influence viewership growth and longevity [10].

Engagement-quality metrics, particularly like rate and comment rate, also emerged as more predictive than raw view counts. This supports the conclusion from Wu et al. that engagement indicators provide a more reliable measure of user interest and video impact than view metrics alone [11].

Additionally, our classification results suggest that high and low popularity videos are more easily separable, while medium popularity videos are harder to categorize accurately. This reflects patterns noted in popularity modeling studies, where extreme cases tend to have more distinct feature profiles than those in the mid-range [10].

To build upon this work, future research could incorporate deep learning for capturing complex patterns, time-series modeling for trend analysis, and natural language processing (NLP) to leverage unstructured text such as titles and descriptions. These enhancements could improve predictive accuracy and offer richer interpretability of content success factors.

VI. CONCLUSION

This study demonstrates that temporal factors, particularly publishing year and channel age, play a significant role in determining a video's popularity on YouTube. Channels with a longer presence and more recent content tend to attract greater audience engagement.

Engagement-based metrics, such as like rate and comment rate, proved to be more predictive of success than raw view counts alone. This suggests that how audiences interact with content is a stronger signal of its impact than visibility alone.

Lastly, the results indicate that consistent, high-quality content production across diverse content genres contributes meaningfully to sustained popularity. These insights can inform content strategy for creators and serve as the basis for more advanced predictive systems in future research.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Dr. Suyong Eum for assigning this engaging and insightful project. This task not only allowed us to apply our knowledge of data analysis and machine learning but also allowed us to collaborate effectively as a team. Nigar Alizada, served as the Project Lead and Presenter, managing timelines, coordinating tasks, and overseeing the overall direction of the project. Kojima Yutaka, contributed as both Data Collector and Analyst, playing a key role in gathering YouTube metadata and executing PCA and SVM-based analysis. Yamasaki Aoto also worked as Data Collector and Analyst, supporting dataset preparation and model development throughout the project. And Chama Nalishuwa as the report writer.

REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[4] H. E. Jang, S. H. Kim, J. S. Jeon, and J. H. Oh, "Visual attributes of thumbnails in predicting youtube brand channel views in the marketing digitalization era," *IEEE Transactions on Computational Social Systems*, vol. 11, July 2023.

[5] Z. Halim, S. Hussain, and R. H. Ali, "Identifying content unaware features influencing popularity of videos on youtube: A study based on seven regions," *Expert Systems with Applications*, vol. 206, p. 117836, 2022.

[6] Y.-L. Chen and C.-L. Chang, "Early prediction of the future popularity of uploaded videos," *Expert Systems with Applications*, vol. 133, pp. 59–74, 2019.

[7] Q. Kong, M.-A. Rizoiu, S. Wu, and L. Xie, "Will this video go viral? explaining and predicting the popularity of youtube videos," in *Proceedings of the WWW 2018 Companion*, 2018, pp. 175–178.

[8] H. Jeon, W. Seo, E. Park, and S. Choi, "Hybrid machine learning approach for popularity prediction of newly released contents of online video streaming services," *Technological Forecasting and Social Change*, vol. 158, p. 120303, 2020.

[9] M. U. N. Nisa, D. Mahmood, G. Ahmed, S. Khan, M. A. Mohammed, and R. Damaševičius, "Optimizing prediction of youtube video popularity using xgboost," *Electronics*, vol. 10, no. 23, p. 2962, 2021.

[10] W. Hoiles, A. Aprem, and V. Krishnamurthy, "Engagement dynamics and sensitivity analysis of youtube videos," *arXiv*, 2016, arXiv:1611.00687.

[11] S. Wu, M.-A. Rizoiu, and L. Xie, "Beyond views: Measuring and predicting engagement in online videos," *arXiv*, 2017, arXiv:1709.02541.