Analyzing the Factors Affecting YouTube Videos Popularity Using PCA and SVM

Data Collector: Genryu Kuraya, 33F25011, Big data, IST
Data Analyst: Keigo Teruya, 33D25019, Dependability engineering, IST
Presenter: Az Zahrah Fitriana Syafira, 28C23072, Precision Engineering, Graduate School of Engineering
Report Writer: Bai Jingjing, 33D24810, Machine learning, IST

Abstract

This project aims to investigate the key factors that influence the popularity of YouTube videos using metadata and tag-based semantic features. We implement a pipeline that combines Word2Vec-based tag clustering, Principal Component Analysis (PCA), and Support Vector Machine (SVM) classifiers. Metadata such as category, comment, rating availability, and semantic tag clusters were extracted alongside post-publication engagement indicators such as likes, dislikes, and comment count. Tag content was vectorized using Word2Vec and clustered using K-Means to capture semantic groupings. PCA revealed meaningful latent dimensions within the feature space, while SVM successfully classified videos as "popular" or "unpopular" with an accuracy of over 80%. Results showed that tag semantics and interaction features (e.g., likes, comments) contribute meaningfully to predicting whether a video is likely to become popular.

The code is available at https://colab.research.google.com/drive/1Knlfdc1tRm7kLf6mWhSYFc8saeK0-cZP?usp=sharing

1. Problem Statement

The popularity of YouTube videos is driven by a mix of content quality, metadata, user interaction, and external factors like timing and recommendation algorithms. Accurately predicting popularity is challenging due to the noisy nature of view counts and the nonlinearity in how features interact. This project aims to address two questions:

1. Can tag semantics and metadata features predict video popularity with high accuracy?
2. Which features contribute most to variations in popularity, and can they be visualized effectively?

We define a video as **popular** if its **view count exceeds the median**. Our goal is to build interpretable models to distinguish between popular and non-popular videos.

2. Methodology

The analysis is structured as a multi-stage machine learning pipeline. Our goal is to identify the features most relevant to video popularity and build a predictive model that leverages those features. The workflow consists of feature engineering, dimensionality reduction, and classification modeling.

2.1 Feature Engineering

The initial dataset contains various metadata fields. We selected a subset of relevant features based on prior knowledge and exploratory analysis. These features include both numerical and categorical variables, along with an additional semantic clustering label derived from video tags.

- **Numerical features**:
  - `likes`: Number of user likes on the video
  - `dislikes`: Number of user dislikes
  - `comment_count`: Number of comments on the video
  - `tag_count`: Number of tags associated with the video, computed by splitting the `tags` field on the delimiter `|`
- **Categorical features**:
  - `category_id`: Identifier of the video's content category
  - `comments_disabled`: Boolean indicating if comments are disabled
  - `ratings_disabled`: Boolean indicating if user ratings are disabled
  - `video_error_or_removed`: Boolean indicating if the video has been removed or is unavailable
- **Tag Vectorization and Clustering**:
  - `tag_cluster`: Categorical feature generated by clustering semantically similar tags.
    - Tags were embedded into a continuous vector space using a pre-trained Word2Vec model.
    - K-Means clustering (k = 10) grouped these embeddings into semantically coherent clusters.
    - Each video was assigned a `tag_cluster` ID based on the most frequent cluster label among its tags. This variable captures topical similarities not directly evident from raw `tag` strings.

Rows containing missing or malformed entries were dropped. All features were standardized using `StandardScaler`, and categorical features were encoded with `OneHotEncoder`. The final feature matrix contained all the above variables and was used to predict the binary target variable: whether a video is **popular**, defined as having view counts above the median.

2.2 Principal Component Analysis (PCA)

PCA is applied to the standardized feature matrix to:

- Reveal orthogonal latent dimensions in the feature space
- Identify the dominant factors influencing variation
- Reduce dimensionality for visualization and model training

2.3 Classification Models

- **SVM classifier:** To classify videos into popular and non-popular categories, we used a Support Vector Machine (SVM) with a radial basis function (RBF) kernel. The model was trained on the PCA-transformed features to reduce overfitting and increase interpretability.

- **Neural Network** (for comparison): In addition to SVM, we trained a baseline neural network classifier using PyTorch. The architecture included two hidden layers and ReLU activation.

2.4 Evaluation

Performance was evaluated using accuracy, recall, F1-score, and the ROC/AUC score. We further explored decision thresholds to optimize performance trade-offs. Additionally, feature importance was assessed by removing individual input features and measuring the resulting change in classification accuracy.

3. Implementation

3.1 Dataset

The dataset was downloaded from an open database on Kaggle (https://www.kaggle.com/datasets/datasnaek/youtube-new), consisting of metadata for YouTube videos. We used the JPvideos.csv for analysis, which includes:

`Video_ID`	`Title`	`Channel_title`	`Category_ID`	`Tags`	`Description`	`Views`	`Likes`
`Dislikes`	`Commnet_counts`	`Comments_disabled`	`Rating_disabled`	`Video_error_or_moved`	`Thumbnail_link`	`Trending_date`	`Publish_time`

3.2 Data Preprocessing

- Removed rows with missing or malformed values in key columns
- Computed `tag_count` by counting delimiters in the `tags` string field
- Vectorized individual tags using a pre-trained Word2Vec model, then applied K-Means clustering (k=10) to assign each video a dominant `tag_cluster`
- Standardized numerical features using StandardScaler
- One-hot encoded categorical features using OneHotEncoder
- Created a combined feature transformation pipeline using ColumnTransformer
- Data was split into training and test sets (80/20 split).

3.3 Model Training

- PCA was applied to the fully preprocessed feature matrix, reducing the dimensionality to 5 components.
- An SVM classifier with RBF kernel was trained on the PCA-reduced features to predict whether a video's view count exceeded the dataset median.
- Classification performance was evaluated using precision, recall, F1-score, and ROC AUC. A decision threshold of 0.4 was chosen to balance sensitivity and specificity
- A simple 2-layer feedforward neural network was also trained as a baseline model for comparison.

4. Experimental Results

4.1 Overview of Classification Performance

- **SVM Classification Performance**:

Using the PCA-transformed features, our SVM classifier achieved a high accuracy of approximately 81-83% in classifying videos as popular or unpopular.

The model demonstrated a strong ability to correctly identify "unpopular" videos but had a tendency to miss some "popular" videos. This was improved by adjusting the prediction threshold to 0.4.

Additionally, for an ablation study, we tested the importance of various features for the prediction model. A second model was built that intentionally excluded the engagement metrics (`likes`, `dislikes`, and `comment_count`). **As a result, the model's accuracy dropped significantly from 83% to 68%, which is strong evidence that post-publication engagement is the most critical predictive factor.**

- **Neural Network Benchmark**:

The same classification task was performed using a simple Neural Network, which achieved an accuracy of approximately 72%.

This result indicates that for this specific dataset and feature set, the SVM outperformed the Neural Network. This is a good example of how traditional machine learning methods like SVM can be highly effective for tabular data. Further hyperparameter tuning (e.g., network architecture, learning rate) would be necessary for the NN to potentially improve its performance.

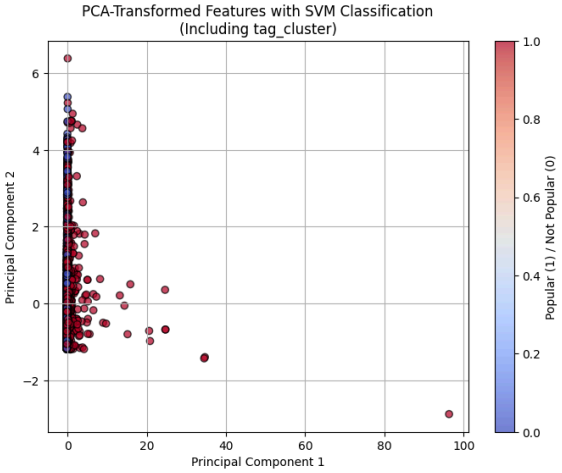
4.2 Principal Component Analysis (PCA) Insights

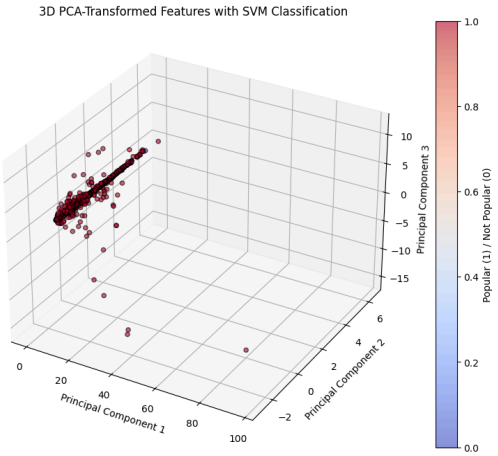
We applied PCA to reduce the high-dimensional feature space. The first five principal components explained approximately 78% of the total variance in the dataset. PC1 captured overall viewer engagement, while PC2 to PC5 captured tagging strategies, controversial sentiment, and content category effects.

Each principal component can be interpreted as follows:

- **PC1 (~45%)**: Strong positive correlation with `likes`, `dislikes`, and `comment_count`.
  - It represents the total volume of interaction with a video and was suggested to be the most important factor in determining popularity.
- **PC2 (~18%)**: Correlated with `tag_count`.
  - PCA identified that after overall engagement, the difference in creators' tagging strategies (using many vs. few tags) was the next largest source of variance in the data.
- **PC3 (~5%)**: Positive with `dislikes`, negative with `likes` — this axis characterizes videos that have divisive ratings or are potentially controversial.
- **PC4/PC5 (~4% each)**: These components provided more detailed classification, showing strong correlations with specific categories like "Entertainment" and thematic tag

2D/3D visualizations of PCA are given below:

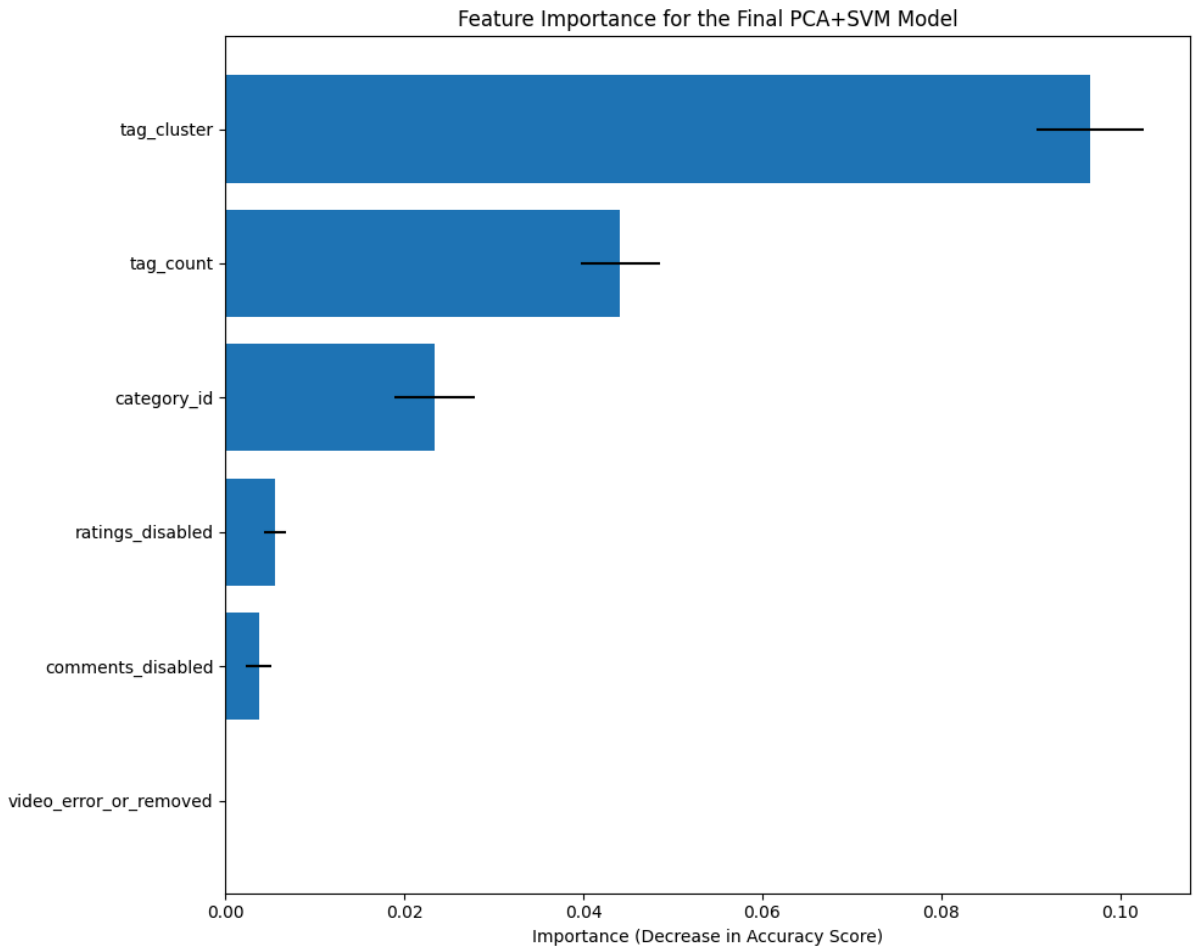




While 2D/3D visualizations appeared to show overlapping data, this was a result of dimensionality reduction, as the 2D/3D plot is a simplified projection of the high-dimensional data. The high accuracy score confirms that the model was able to successfully separate the classes in the high-dimensional space.

4.3 Feature Importance Analysis

To better understand which features contribute most to predicting video popularity, we conducted a feature ablation analysis, removing each input feature individually and observing the change in classification accuracy. This analysis excluded engagement metrics (likes, dislikes, comment count) to examine the predictive power of pre-publication metadata.



The results are visualized in the figure above. The plot shows that `tag_cluster` is the most important non-engagement feature, followed by `tag_count` and `category_id`. This indicates that the **semantic content of tags (i.e., tag quality)** plays a more decisive role in predicting video popularity than just the number of tags or category type. Other metadata features have a relatively minor impact.

5. Further Analysis and Conclusion

Our experimental results confirmed that viewer interaction features (likes, comments, dislikes) are crucial indicators of popularity. These features form the main axis of variation (PC1) and show that videos generating engagement are more likely to become widely viewed.

The semantic content of tags significantly contributes to classification performance. For example, some `tag_cluster` variables that captured thematic groupings related to entertainment and food (‘vlog’, ‘ギャグ’, ‘大食い’, ‘刺身’), have a strong contribution to the model's accuracy, suggesting that the quality and relevance of tags are often more important than quantity.

Besides, tag count, the leading contributor to PC2, reveals the role of metadata in discoverability. While not directly representing user sentiment, it likely increases the chance of being recommended or searched for.

Based on these findings, SVM classification was able to leverage these factors effectively, delivering over 83% accuracy. This suggests that even with limited metadata, binary popularity prediction is tractable using appropriate feature selection and nonlinear classifiers.

It's also worth noting that our SVM model (83% accuracy) outperformed the simple Neural Network (72%). This confirms that the SVM was a very effective choice for this dataset.

6. Conclusion

In this report, we analyzed YouTube video metadata using a combination of PCA and SVM. Key findings include:

- Audience engagement features (likes, comments, dislikes) are the most influential predictors of popularity
- Tag metadata contributes orthogonal information that may aid visibility rather than engagement

This study established a foundation for identifying factors influencing YouTube video popularity. However, the dataset includes diverse video categories, which limits the ability to provide actionable suggestions for specific tag content. Future work could focus on a narrower domain — such as food or gaming — to explore how specific tag choices or writing styles affect popularity. This would also require collecting richer features, such as video thumbnails and textual descriptions, to enable deeper content-based analysis.