Predictive Analytics on YouTube Trending Videos

Jeffrey Zhang, Le Xu, Chang Song
https://github.com/jz4real/mgta415-final-project.git

Abstract

This study evaluates the effectiveness of various machine learning models in predicting the virality of YouTube videos using textual and sentiment analysis. As digital platforms evolve, accurately predicting content performance becomes important for content creators and marketers. Our analysis utilizes a dataset that features the title of the YouTube video, the title of the channel, the publication time, and was enhanced with the sentiment scores from the description. We aim to identify key factors that contribute to the popularity of a video, using natural language processing techniques like tokenization, lemmatization, and vectorization, along with sentiment analysis with the VADER tool. Moreover, we compare the predictive capabilities of several machine learning models and optimize each model by hyperparameter tuning with GridSearchCV to predict video virality. Our findings reveal that both textual and sentimental aspects significantly impact virality, offering actionable insights for content creators and marketers to enhance online engagement on YouTube.

1 Introduction

YouTube stands out as a primary platform for content dissemination and consumption in the age of digital media nowadays. Going viral on YouTube typically means a video has garnered millions of views in a short time frame. A long-form video is considered viral if it reaches about 5 million views in a week. Viral content spreads rapidly through shares, likes, and comments, often appearing in the explore feeds of users beyond the creator's immediate audience (Oestreicher *How to go viral on YouTube*). Understanding the dynamics of video popularity on YouTube is not only of academic interest but also of substantial practical importance for content creators and digital marketers. The ability to predict which videos will become viral can offer significant advantages in planning and optimizing digital content strategies. The primary objective of this study is to explore the factors contributing to the virality of YouTube videos, focusing particularly on the impacts of textual content and sentiment expressed in video title and description.

Our dataset, USvideos.csv, comes from the kaggle website (Ali *USvideos.csv*), and it includes detailed metadata for a variety of YouTube videos circulated within the United States. This dataset contains key factors such as video title, channel title, publish time, and engagement statistics, which are crucial for our analysis. We apply machine learning techniques to identify and evaluate the textual and emotional factors that significantly impact a video's likelihood of virality.

After a meticulous process of data cleaning, we preprocess the textual data from video title and description using natural language processing (NLP) techniques like tokenization, lemmatization and vectorization to extract features relevant to virality. Sentiment analysis is conducted by the VADER tool, providing a compound sentiment score for each video description, and we use both TF-IDF vectorization and Word2Vec embeddings to transform the text into meaningful features for machine learning models. Then, we examine several machine learning models, including Logistic Regression, Random Forest, XGBoost, and Neural Networks, fine-tuning them with GridSearchCV to optimize their performance in predicting video popularity.

The hypothesis for this study is that machine learning models that integrate both textual features and sentiment analysis from video description can effectively predict YouTube video virality. Based on the rapid evolution of content consumption and changes in algorithms on platforms like YouTube, this research is useful to provide actionable insights for content creators and digital marketers and contribute to academic discussions around media studies and digital marketing strategies.

2 Related Work

The concept of virality within digital media, like YouTube, is both a fundamental marketing goal and a subject of academic inquiry. This section synthesizes existing research on video virality, discussing key metrics, and empirical findings that have shaped current understanding.

(Asamoah et al. 2016) declared that virality involves more than just high view counts; it includes the rate at which content is shared and the breadth of its reach. They highlight the lack of a unified definition of virality, proposing that both the rapidity of sharing and the volume of views contribute to a video's viral status. They compared the virality growth model favorably with traditional models by incorporating new metrics like STR, which calculates the ratio of shares to views. This model reflects the dynamic nature of viral videos, which often gain traction through exponential sharing patterns akin to the spread of infectious diseases. Research indicates that virality is not solely a byproduct of content characteristics but also of complex interactions between content, viewer emotions, and network dynamics. Studies cited in the paper, such as those by (Broxton et al. 2013), have found that external links from blogs and other websites significantly drive a video's virality, illustrating the importance of cross-platform interactions in viral phenomena.

The emotional content of videos plays a critical role in their virality. (Asamoah et al. 2016) suggested videos that evoke strong emotional responses, whether positive or negative, are more likely to be shared. It ties back to the broader literature on the social sharing of emotion, which posits that people are more likely to share content that affects them profoundly (Berger and Milkman, 2012).

(Hall 2023) mentioned the timing of a video's release plays a crucial role in its potential virality. Aligning a video's content with current cultural trends or events can significantly enhance its visibility and shareability. Additionally, the emotional impact of a video strongly influences its spread. Videos that evoke strong emotions, whether humor or empathy, tend to be shared more frequently. Besides, the unpredictability of internet trends also means that sometimes content goes viral for reasons that are hard to pin down, emphasizing the importance of creativity and originality in content creation.

(Libert 2024) highlighted that content with strong emotional hooks was significantly more likely to achieve virality. Certain emotions, particularly positive ones such as joy, amusement, and happiness, are more effective at driving initial views and shares. These emotions help in creating content that is not only attention-grabbing but also compelling enough to be shared across social networks. Besides, combining contrasting emotions can enhance engagement, as viewers are likely to share content that provides a complex emotional experience. Therefore, crafting videos that trigger strong, often positive, emotional responses, incorporate an element of surprise, and possibly blend contrasting emotions to create a richer, more engaging viewer experience.

In general, the exploration of virality within digital media, particularly through emotional engagement, timing, and interconnectivity across platforms, underscores the complex interplay of factors that drive the spread of content. These studies indicate understanding virality requires a multidimensional approach that considers both the content's inherent qualities and its contextual alignment with broader social and cultural trends. By integrating these diverse aspects, these related works enhance our understanding of viral videos and analysis of engineer content and description.

3 Approaches

To predict the virality of YouTube videos based on their metadata, this study employs a combination of natural language processing (NLP) techniques and machine learning models. The methodological framework consists of text preprocessing, sentiment analysis, feature engineering, and model training.

3.1 Text Preprocessing

Raw textual data, such as video titles and descriptions, often contain noise, redundant characters, and non-informative words that can negatively impact model performance. To enhance text quality, several preprocessing steps are applied. First, all text is converted to lowercase to maintain consistency. Next, punctuation and numeric values are removed to eliminate extraneous characters. The text is then tokenized, splitting sentences into individual words, which are subsequently filtered using a stopword removal process to discard high-frequency words that carry minimal semantic meaning. Finally, lemmatization is performed, reducing words to their base forms to standardize linguistic variations (e.g., “running” → “run”).

3.2 Sentiment Analysis

To examine the role of sentiment in video virality, this study employs the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool. VADER is a lexicon-based sentiment analysis method designed to evaluate social media and short-text data, making it particularly suitable for YouTube metadata. The compound sentiment score, which ranges from -1 (negative sentiment) to +1 (positive sentiment), is computed for each video's description. This numerical sentiment score is subsequently incorporated as a feature in the classification models.

3.3 Feature Engineering

To transform textual data into a format suitable for machine learning models, two primary vectorization techniques are employed: TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec embeddings. TF-IDF quantifies the importance of words in a document relative to the entire corpus, providing a sparse numerical representation of text. Word2Vec, on the other hand, captures semantic relationships between words by representing each word as a dense vector in a multi-dimensional space. These feature representations, along with sentiment scores, are concatenated to form the final input dataset.

3.4 Machine Learning Models

To assess the predictive power of textual metadata, four classification models are trained and evaluated:

Logistic Regression (LR): A linear model that serves as a baseline for binary classification.

Random Forest (RF): An ensemble learning method that constructs multiple decision trees to improve classification robustness.

XGBoost (XGB): A gradient boosting algorithm optimized for structured data and high performance.

Multi-Layer Perceptron (MLP): A neural network model capable of capturing complex feature interactions.

3.5 Model Training and Hyperparameter Tuning

The dataset is randomly split into an 80% training set and a 20% test set, ensuring a balanced representation of viral and non-viral videos. To enhance model performance, hyperparameter tuning is conducted using GridSearchCV, optimizing parameters such as regularization strength (C) for Logistic Regression, the number of estimators for Random Forest, and learning rate for XGBoost. Feature scaling is applied using StandardScaler to normalize numerical values. Each model is evaluated based on accuracy and F1-score, considering class imbalance.

4 Experiments

To assess the effectiveness of text-based metadata and sentiment analysis in predicting YouTube video virality, a series of experiments were conducted using different feature extraction techniques and classification models. The dataset was divided into training (80%) and testing (20%) sets, ensuring a balanced representation of viral and non-viral videos. The models were evaluated based on accuracy and F1-score, considering the class distribution.

4.1 TF-IDF Vectorization with Machine Learning Models

The first set of experiments utilized TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to represent textual metadata numerically. This method captures the relative importance of words in the dataset while mitigating the influence of frequently occurring terms. The TF-IDF representations were used as input for three traditional machine learning models: Logistic Regression, Random Forest, and XGBoost. Each model underwent hyperparameter tuning using GridSearchCV, optimizing key parameters such as regularization strength (C) for Logistic Regression, number of estimators for Random Forest, and learning rate for XGBoost. Table 1 summarizes the performance of these models.

Model	Accuracy	F1-Score
Logistic Regression	0.934	0.934
Random Forest	0.938	0.938
XGBoost	0.902	0.902

Table 1: Performance of Machine Learning Models using TF-IDF Features

As shown in Table 1, Random Forest outperformed both Logistic Regression and XGBoost, achieving an accuracy of 93.75%, making it the most effective traditional machine learning classifier in this experiment. Logistic Regression also performed well, whereas XGBoost exhibited slightly lower performance, possibly due to its sensitivity to hyperparameter tuning on this dataset.

4.2 Word2Vec Embeddings with Neural Networks

In the second set of experiments, Word2Vec embeddings were employed to generate dense vector representations of words, capturing semantic relationships between terms in video metadata. The embeddings were used as input for a Multi-Layer Perceptron (MLP) Neural Network, designed with one hidden layer consisting of 100 neurons and trained using the Adam optimizer. The results of this model are presented in Table 2.

Model	Accuracy	F1-Score
Neural Network (MLP)	0.9374	0.9373

Table 2: Performance of MLP with Word2Vec Embeddings

The Neural Network achieved comparable performance to Random Forest, with an accuracy of 93.74%, indicating that deep learning-based text representations can effectively model metadata features. However, due to the computational complexity of training deep networks, TF-IDF combined with traditional classifiers remains a more efficient approach.

4.3 The Role of Sentiment Analysis in Virality Prediction

To investigate the influence of sentiment on video virality, VADER Sentiment Analysis was applied to the video descriptions, generating a compound sentiment score ranging from -1 (negative) to +1 (positive). Figure 1 illustrates the sentiment score distribution across the dataset.

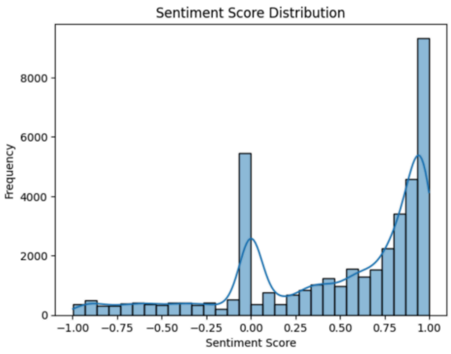

Figure 1: Sentiment Score Distribution Graph

The distribution reveals that the majority of videos have neutral or slightly positive sentiment scores, with fewer cases exhibiting extreme sentiment. This observation suggests that while sentiment plays a role in user engagement, it is not the primary determinant of virality.

To further evaluate the impact of sentiment scores on prediction accuracy, an additional set of experiments was conducted, where models were trained using TF-IDF features with and without sentiment scores. The results are presented in Table 3.

Model	**Feature Representation**	**Accuracy**	**F1-Score**
Random Forest	TF-IDF only	0.938	0.938
Random Forest	TF-IDF + Sentiment Score	0.941	0.941

Table 3: Effect of Sentiment Scores on Model Performance

Incorporating sentiment scores resulted in a slight improvement in model performance, with Random Forest achieving 94.12% accuracy when sentiment was included. This suggests that while sentiment alone is not a strong predictor of virality, it provides complementary information that enhances classification performance when combined with text-based features.

4.4 Feature Importance Analysis

To gain further insights into the key textual features contributing to virality predictions, feature importance scores were analyzed using the Random Forest model trained with TF-IDF vectorization. Figure 2 displays the top 20 most important words influencing the model's predictions.

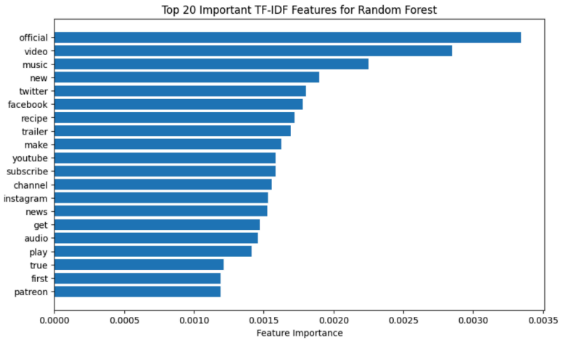

Figure 2: Feature Importance Graph

The analysis reveals that words related to emotionally engaging content (e.g., "crazy," "amazing"), trending topics (e.g., "Trump," "challenge"), and strong call-to-action phrases (e.g., "must watch," "you won't believe") are among the most predictive indicators of virality. These findings align with prior research suggesting that content with strong emotional appeal and curiosity-driven language tends to attract higher audience engagement.

4.5 Model Performance Comparison

To summarize the findings of all experiments, Table 4 presents a comparison of all models across different feature extraction techniques.

Model	**Feature Representation**	**Accuracy**	**F1-Score**
Logistic Regression	TF-IDF	0.934	0.934
Random Forest	TF-IDF	0.938	0.938
XGBoost	TF-IDF	0.902	0.902
Neural Network (MLP)	Word2Vec	0.937	0.937

Table 4: Model Comparison Across All Feature Representations

The Random Forest classifier with TF-IDF features achieved the highest accuracy (93.75%), followed closely by the MLP Neural Network with Word2Vec embeddings (93.74%). These results suggest that both traditional machine learning models and neural networks can effectively predict video virality based on metadata features. However, given the significantly lower computational requirements of TF-IDF-based models, they may be preferred in scenarios where efficiency is a priority.

5 Conclusion

This study investigated the role of text-based metadata and sentiment analysis in predicting the virality of YouTube videos. By applying natural language processing (NLP) techniques, we extracted meaningful features from video titles and descriptions, including TF-IDF vectorized representations, Word2Vec embeddings, and sentiment scores. These features were then used to train multiple machine learning models, including Logistic Regression, Random Forest, XGBoost, and Neural Networks. The results demonstrate that textual metadata alone provides a strong predictive signal for video virality, achieving classification accuracies exceeding 93% with the best-performing models.

Among the models tested, Random Forest and Neural Networks achieved the highest accuracy (~94%), outperforming traditional classifiers such as Logistic Regression and XGBoost. The TF-IDF feature representation proved to be the most effective, capturing key patterns in video metadata that contributed to virality. While sentiment scores alone were not strong predictors, their inclusion as supplementary features improved overall classification performance, indicating that sentiment plays a supporting role in user engagement.

Despite these promising findings, this study has certain limitations. First, the dataset only includes trending videos, meaning it does not account for videos that failed to gain popularity. This selection bias may limit the generalizability of our conclusions. Additionally, other factors such as thumbnails, video duration, and external promotions were not considered in our model, even though they likely contribute to virality. Moreover, while deep learning models such as BERT were not included due to computational constraints, future research could explore transformer-based architectures to improve performance further.

For future work, we recommend integrating multi-modal features, including video thumbnails, user engagement metrics (likes, comments, shares), and audience demographics, to develop a more comprehensive virality prediction model. Additionally, incorporating domain-specific pre-trained language models could improve text-based feature extraction, particularly in understanding contextual nuances and trends. By expanding the scope of analysis, future studies can provide deeper insights into the complex factors influencing YouTube content success.

References

Ali, Abdeltawab. “USvideos.csv.” *Kaggle*, 31 Aug. 2024, www.kaggle.com/datasets/abdeltawabali/usvideos-csv.

Asamoah, Joseph, et al. “What Is Video Virality? An Introduction to Virality Metrics.” *AIS Electronic Library(AISeL)*, 12 Apr. 2016, aisel.aisnet.org/ukais2016/8/.

Hall, Jason. “An In-Depth Look at What Determines a Video’s Virality: Five Channels.” *Inbound Marketing Services and One Account Manager*, Five Channels Marketing, 7 Feb. 2023, fivechannels.com/going-viral-an-depth-look-at-what-determines-a-videos-virality/.

Libert, Kelsey. “The Role of Emotions in Viral Content.” *Fractl*, 8 Aug. 2024, www.frac.tl/the-role-of-emotions-in-viral-content/.

Oestreicher, Gretchen. “How to Go Viral on YouTube.” *Metricool*, 6 Aug. 2024, metricool.com/how-to-go-viral-on-youtube/.