Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

Enhancing YouTube Video Popularity Prediction via Stage-Aware Multimodal Learning Across Cold-Start and Engagement Stages with Synthetic Description Generation

[1,] Hamsa Hameed Yousif [2,] Ayad Hameed Mousa
[1,2] Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq.

Article Info
Article history:
Received 24 -2-2026
Received in revised form 26-4-2026
Accepted 81-5-2026
Available online 30 -6 -2026
Keywords: YouTube Popularity Prediction, Multimodal Learning, Feature Fusion, DeBERTa, CLIP, Machine Learning, Social Media Analytics

Abstract

Most methods to predict a video's popularity on YouTube rely on post-publication factors, which results in target leakage and can only be applied after a video is published. This paper suggests a stage-aware multimodal framework that offers a clear separation between post-publication analyses and early-stage predictions. This framework includes features that are made of text, images, metadata, and user interactions. It captures characteristics of the content and user engagement. Text features are made using DeBERTa, images are made using CLIP, and metadata and user interaction features are added to extract more contextual and behavioral information in a given situation. Predicting early-stage popularity trends with only text features and STM is effective, and multimodal fusion is needed for improving overall performance of the model. The model that uses the text, image, metadata, and user interaction achieved an F1 score of 0.9827 and a maximum accuracy of 0.9848, which is a massive improvement from single-modality models. This study proposes a framework with a heterogeneous integration of features. The results show a positive improvement to avoid target leakage and help predict early-stage popularity trends

Corresponding Author E-mail: hamsa.h@s.uokerbala.edu.iq ,ayad.h@uokerbala.edu.iq
Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1. Introduction

1.1 Background and Motivation

YouTube has proven to be one of the most powerful video sharing platforms. Its rapid development has allowed them to establish channels to share content across entertainment, education, and news. With the competitive market opportunities becoming scarcer, predicting video popularity has proven to be an important field of research [1].

Views, likes, and comments have been the de-facto methods to gauge the popularity of the video. Along with user behavior, the recommendations also affect these variables. With that in mind, the problems of predicting views and engagement metrics are quite complex [2].

Furthermore, the fact that popularity prediction is time sensitive and depends on early engagement presents further challenges. This is then complicated by regional factors and differences in cultural practices [3], [4].

These factors are why the popularity of videos has led the research community to rely on machine-based methods.

The initial methods predominantly used classical machine learning methods. The more recent work has started to rely on multimodal methods, which are better suited for the prediction of popularity as they consider the visual, text and metadata associated with a video [5], [6].

Most of the work that has been done relies primarily on video engagement metrics, which can only be collected post-publication. Thus, these methods are limited in their ability to support early-stage prediction. In addition, such methodologies suffer from target leakage, as they rely on information that becomes available only after publication [7].

Most of the work that has been done, relies mostly on the videos' engagement metrics, which can only be collected post-publication. Thus, these methods are limited in their ability to predict views and engagement metrics. In addition to that, the methodologies suffer from data leakage in that they rely on data that has been collected post-publication [7].

Therefore, more work has to be done to provide realistic solutions, whereby the time of information can be temporally bounded and can be useful for predicting popularity even before the information is published.

1.2 Problem Definition and Research Gap

Although YouTube’s popularity prediction is receiving greater focus, many crucial limitations still exist in the present literature. One primary concern is the absence of an explicit definition of what a “popular” video is. Prior studies use different thresholds, evaluation criteria, or labeling. It undermines the comparability of studies and the ability to reproduce them**[1]** . Arguably, among the most significant shortcomings, is the reliance on interaction features such as likes, comments, and other engagement signals. While many of these features enable the ability to make accurate predictions, they rely on what would otherwise be an unreal expectation. This introduces the potential for target leakage when correlated with the prediction label [2].Moreover, several current approaches seem to neglect the importance of time with respect to the availability of and reliance on various features. Interaction features often

appear as a result of user engagement, whereas content-based features (e.g., textual and visual features) are available pre-publication. Ignoring these differences creates a mismatch between model design and real-world prediction requirements, particularly in early-stage prediction scenarios[3] .The absence of consideration for incomplete and underdeveloped textual features in current models also points to a shortcoming in the current studies. Several studies use underdeveloped features or descriptions, failing to fully analyze their impact with reliance on content-focused models as their primary method of modeling [4]. These gaps highlight the need for improved, organized frameworks that integrate several feature types, especially when addressing temporal limitations. This thesis seeks to invite consideration of a stage-based integrated framework that reconceptualizes prediction into the Pre-Interaction and Interaction stages of the model's framework. The Early Stage uses a combination of textual, visual, and metadata features, while the Interaction Stage models engagement features separately to prevent target leakage. Further, a deliberate design is implemented to improve textual completeness and analyze its effect on early prediction performance. This makes the design of early YouTube popularity prediction more consistent, understandable, and usable in practice. This design ensures a realistic and leakage-free prediction setting.

1.3 Research Objectives

This dissertation develops an accurate, rational framework for the prediction of YouTube video popularity, targeting the historically prominent, pernicious limitation of the lack of an awareness of the predictive potential of temporal features vis-à-vis the post-publication features of instructive value claimed by critics of the existing models. The framework explains the modality of the predictability of multi-modal features vis-à-vis the post-publication features through the manipulation of the video content and the user meta features interdependencies. Moreover, the early-stage prediction of populist potential is improved through the manipulation of the interactive predictability of the user meta content of the videos through the improvement of the systematic content generation.

In addressing the limitations presented above, the thesis aims at the following key contributions:
Multimodal, Stage-Aware Framework that separates Early from Interaction Stage, facilitating realistic prediction without target leakage.

Inclusion of differing modalities, integrating textual features (e.g., DeBERTa), visual features (e.g., CLIP), metadata features, and interaction feature in a unified and consistent framework.

A controlled approach to the description generation as a means to textually complete incomplete textual inputs to enable systematic assessment.

A consistent reproducible framework of defining video popularity as the 80th percentile of training set view counts.

A uniform Testing Framework using multiple classifiers (Logistic Regression, Random Forest, XGBoost, Linear SVM) and consistent evaluation metrics.

2. Literature Review

Previous attempts at predicting YouTube popularity predominantly acknowledged unitary feature representations and singular media forms at best.

The research [12] approaches popularity prediction with a multi-class classification format and utilizes the textual and metadata features from the Trending YouTube Video Statistics dataset. The authors assert that the best results (F1 = 0.736) occurred following feature selection and the use of XGBoost. While class imbalance is apparent, the study is underpinned by a structurally restrictive classification framework, and the overlapping class boundaries directly impact the prediction draw. Additionally, despite feature selection, the study is limited by the manifestation of overfitting.

The authors of [5] propose a model of Multimodal Deep Learning, wherein the model attends to and accommodates the weighted importance for each of the textual, visual, and metadata features. The model is successful (MAE ≈ 1.401) and realized the value of Multimodal Fusion. What is, however, the case, is that the aim of the model was a regression to positional values, as opposed to a classification, and, further, the attention-based architecture augmented significantly the computational load, and as a result, the model may not be efficient for real life purposes.

The study [13] achieves high accuracy (≈ 88%) by combining features from metadata and interactions augmented through fusion and feature engineering. Although this represents a notable achievement, the model, for reasons of practicality, particularly limits the prediction of popularity to events occurring post-publication, for example, the interactions of views and likes. Additionally, the industriousness sampling of fusion features limits the interpretability of the model and may result in information loss.

The research [14] has incorporated several classification techniques with the inclusion of metadata and interaction features, yielding Logistic Regression to culminate with the best performance of approximately 62.53%. The translation of the research results into various models; however, incipient accuracy reveals little capacity to predict the results. Thought the models are based on features that emerge after the publication, the models are based on various elements that obscure the raw potential of the model to generalize.

The research [15] has incorporated the use of word embeddings to create a measure of text features, alongside the use of traditional machine learning models which has led to some improved performance (accuracy ≈ 87%). Though such measure of text representation is an improvement to more simplistic methods, it still lacks a more dynamic understanding of the context amongst the words. Furthermore, the inability to capture the visual elements, alongside the metadata of the model are evident.

The research [16] has undertaken prediction based on images, applied through deep learning methods using thumbnail images, and has led to a measure of accuracy of (≈ 65.13%). Though the research has highlighted the potential use of visual features, it lacks other elements of prediction, no textual and metadata prediction elements of the model. Furthermore, the relatively small dataset size results in a high risk of overfitting the model to the dataset.

The research [17] has incorporated elements of a hybrid Model that combines text, metadata, and interaction features, yielding reasonable (≈ 85%) results. However, the use of the bag-of-words representation of text features leads to the model unable to capture more deeply the relationships between the words. Moreover, the separation of the regression and the classification design is inco-

herent and the lack of more than one construct of learning diminishes the design's performance.

In study [18], authors use interaction-based features via a binary classification method, achieving very high accuracy (≈ 99.74%). Despite these results being intriguing, the authors cited controlled dataset conditions and use of post-publication signals, resulting in target leakage, thus limiting the model's use of early prediction scenarios. Furthermore, these authors noted the exclusion of content-based features and the lessening of model robustness.

In study [19], strong results (≈ 87.77%) were found via the use of interaction and metadata features, utilizing statistical and machine learning techniques. Nevertheless, the use of regular machine learning models and the limited number of features constrained the models' ability to understand the less visible but complex nonlinear relationships. In addition, the desire to predict less was a result of insufficient complex integration of features coming from various modes.

The study [2] combines interaction and sentiment features using machine learning models and achieves an accuracy of approximately 86.5%. However, due to the domain-specific nature of the dataset (gaming), generalization is not possible, and the dependence on post-publication features compromises the early-stage prediction of the model, while the absence of uniform assessment introduces unpredictability in the results.

The study [20] centers on the development of a visual-based prediction framework, employing thumbnail features such as color, brightness, and structural composition. The framework achieves performance (MAPE ≈ 1.055%), indicating the usefulness of visual characteristics in predicting aspects of video performance. However, the study focuses solely on the use of visual features and disregards textual and metadata, which hampers the study's potential to comprehensively capture the semantic context of video content. In addition, the emphasis on regression-based framework limits design to a classification-based predictive framework. Furthermore, the study focuses on brand channel videos, which limits the versatility of the framework across various YouTube content.

The study [21] uses a text-based prediction framework by employing linguistic features from video titles. In spite of the use of several machine learning models, the prediction performance is at a level of $R2 \approx 0.199$. This signifies the weak prediction potential, which is attributed to the title-based features. This prediction model is deemed ineffective in real-world scenarios as a result of the absence of supportive visual, contextual, and semantic background.

The study [8] presents a multi feature-based classification framework that incorporates textual, metadata, and temporal elements. The framework achieves a moderate level of prediction performance (Accuracy ≈ 82.5%, $F1 \approx 0.75$).

The use of Naïve Bayes modeling assumes that all features are independent from one another. Unfortunately, this independence assumption limits modeling features that are related in complex ways, and also limits the use of time and interaction-based features in initial-stage predictions.

The study [22] offers a framework that draws upon the complementarity and synergy of the visual, textual, and meta-data. Predictions are shown to improve as models are fused. Unfortunately, this study is also limited as it uses continuous prediction metrics that lend themselves to a regression comparison, and it also increases the computational burden as the model processes all the video data, putting real-time systems out of consideration.

The study [23] builds upon all of the engagement and interaction predictors and adds a sentiment layer, which was measured

through the comments. The model showed strong performance (accuracy ≈ 84.3%, F1 ≈ 84.1%) among numerous validating metrics with insufficient confidence in predictive and engagement processes. However, the model does require raw comments which carry their own biases, and it does require post-launch interaction data. Thus, early-stage predictions are limited.

The study [6] offers a framework that draws upon the complementarity and synergy of the textual, visual, temporal, and behavioral layers. The model also showed strong predictive performance (SRC ≈ 0.7324), proving that positive predictive performance can be achieved with hierarchical fusion. Higher model complexity led to model deployment to be impractical due to a high predictive burden, and the outcome of the model remained ambiguous.

The study [24] offers a higher-order framework that draws upon complementarity and synergy of the visual, textual, and audio layers to describe more elaborate and advanced interactions between the different layers.

The performance puts MAPE ≈ 0.1842 and offers validation of more advanced fusion methods. The framework, however, depends on regression-based assessments. Also, it needs the presence of multiple modalities. This includes audio and user behavior data, which may not always be available. Further, the framework requires high computational resources, which makes it more infeasible to be deployed on a large scale.

3. Methodology

This section lays out the proposed pipeline framework for the automation of multimedia cross-referenced prediction of the popularity of YouTube videos. Preprocessing, feature extraction, and labeling, along with training and evaluation, form the core components of the pipeline. The Early Stage and Interaction Stage elements of the pipeline have been designed for pragmatic feature use.

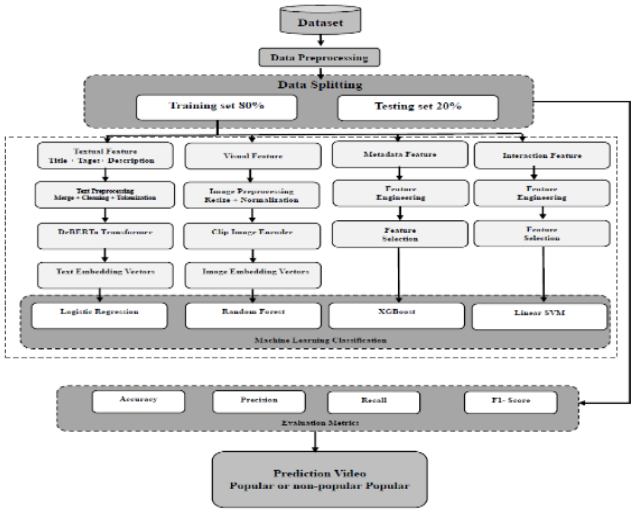

Figure 1: Stage-Aware Multimodal Framework Overall Architecture

During the Early Stage, gaps in feature collection are filled with the following: DeBERTa for text, CLIP for vision, and the inclusion of metadata. Feature collection for text can handle missing descriptions by creating text from the title/tags. During the Interaction Stage, only the interaction features of the metrics are available. View count is

purposely excluded to mitigate target leakage. Multimodal features that the framework encompasses are designed to integrate seamlessly regardless of time

3.1 Dataset Description

The (Trending YouTube Video Statistics). dataset, is used to perform the experiments. It contains an extensive historical collection of videos that have appeared on the trending lists for YouTube in several countries. This dataset contains a variety of content data, including the tags, videos, descriptions, and title metadata. Interaction data features are provided, including comments, likes, and views, along with the relating metadata features, including the category and time of publication. Data from different regions are combined to enable cross-country analysis, while an extra country identifier is added to each record to mark the original source of the record, and remove redundant data entries and standardize record features. Most online YouTube datasets involve trending videos. These datasets manifest the most common trending videos, and major data repositories like Kaggle curate the datasets for benchmarking prior research. Accordingly, the dataset for this study is benchmarked owing to utilization and standardization in most research, enabling reproducible experiments across research.

3.2 Data Preprocessing

The quality and consistency of data for model development were ensured through data preprocessing. The major countries merged records and removed duplicated, misrepresented, missing, or inconsistent values. After combining data from ten countries, duplicate datasets left a dataset of 363,372 records encompassing 17 attributes.

Several preprocessing steps were taken to remove the complexity of designing the model and ensure quality of data. The records were merged, and duplicates, formed from combining them, were eliminated. Secondly, the models' contributors were synthesized through the use of categorical model variables and standardization of numerical variables to fit the machine learning process. Additionally, the length of the title, description, and quantity of tags were quantified through feature engineering and served to standardize the attributes for quality representation of the model. One of the significant issues in the dataset is the lack of video descriptions, with 19,478 entries having an empty video description. A controlled approach was used to treat the empty video descriptions in one of two ways. The descriptions were either left as empty fields, or filled in with descriptions constructed from the corresponding video title and video tags. This design was enabled to assess the impact of completeness of text on the performance of predictions in this dataset. Once this was completed, an 80/20 train-test split was implemented on the dataset to allow for a fair comparison. A binary label was created through a predetermined threshold based on the 80th percentile of the view counts computed on the training dataset. Videos above the threshold were classified as popular, and all other videos were classified as non-popular.

3.3 Feature Extraction

Feature extraction is the process of transforming unprocessed data into a structured and quantitative format which machine learning models are capable of consuming. As proposed structured framework is multimodal, the feature extraction process is performed independently for each modality. Multimodal approaches in this framework aim to address the diversity of features and video content. Specifically, semantic textual features, visual features based on the thumbnail, and the structured attributes of the videos in the form of metadata are employed. This separation of

each modality guarantees that the comprehensive information for the audience in this case is represented for the purpose of prediction.

3.3.1 Textual Feature Extraction using DeBERTa

Textual features are the first of two content-based signaling that are available in the Early Stage in prediction based on pre-publication information. This sort of content is found in the semantic context of video content in the video title, tags, and video description. Textual features when used in early prediction of video popularity are valuable [5] .The value of textual features in video popularity prediction is largely based on the popularity prediction in early video popularity logical reasoning

In order to textual data conversion to the number description of data, one of the transformer-based language models (DeBERTa) is used to form the context in the form of 768-length. Models based in the transformer are often better at identification of semantic relations (Multimodal Deep Learning for Social Media Popularity Prediction with Attention Mechanism). One difficult aspect of the dataset is that the video description field contains missing values. Specifically, 19,478 records do not have descriptions. A two-staged controlled strategy is employed to resolve this problem. The first stage collects the missing descriptions as empty values, which are understood as unfinished textual inputs. The second stage, involving the generation of synthetic descriptions based on the accompanying video title and tags, ultimately brings improvement to the textual completeness.

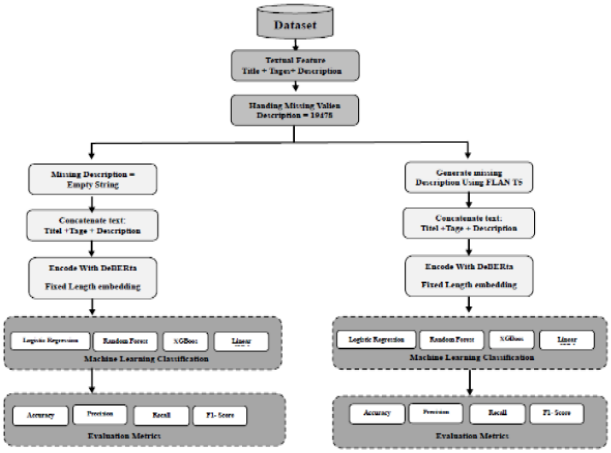

Figure 2: DeBERTa Text Feature Extraction Pipeline with Missing Description Handling

This development provides an opportunity to analyze the effect of textual completeness on Early-Stage prediction occurrence, thus operating within the bounds of the problems recognized by the authors of the work "Construction and Data Analysis of a New Media Content Popularity Prediction Model Based on Naive Bayes Algorithm," concerning incomplete feature representation.

Thereafter, the textual input is formed by bringing together the title, the tags, and the description, regardless of whether the missing description is original, empty, or has been generated. This integrated textual input is subsequently transformed into a fixed and formal contextual representation with the aid of the DeBERTa model to yield the embedding, which is in turn, supplemented to the input set for the supervised models.

To further proof the quality of the generated descriptions, we undertook a semantic similarity evaluation that involved the use of Sentence-BERT (SBERT). We utilized

cosine similarity to determine the degree of similarity of the generated description to the original textual context (the video title and the tags). From the similarity scores, the generated descriptions were placed under three categories of quality: High, Medium, and Low. High descriptions were those that semantic consistency with the original content, medium descriptions showed some level of alignment, and Low descriptions were those that showed a lack of semantic consistency. This quality of description aids the users of this generated dataset in supplementing text data descriptions to the original textual context in a systematic manner, thus improving the quality of the dataset they are using to train their model(s).To demonstrate the viability of the text generation method proposed, we present in Table 1, some examples of actual samples from the generated dataset to exhibit the interlinkage between the input parameters and the output generated.

Table 1: Representative Examples from the Generated Dataset (Original vs Generated Descriptions)

Row ID	Title	Tags	Original Description	Generated Description
594	[ENG SUB] BTS Woojin	bts woojin	NaN	BTS Behind Run Ep 40 39 ( ) and these tags: bts woojin
595	Drake - God's Plan	Drake	NaN	Drake - God's Plan (official music video) is Drake's official music video
597	ALIEN CRA Travel	japanese s	NaN	Japanese street food and sashimi are common in Japan
599	Drake - God's Plan (E-Mix)	[none]	NaN	Drake - God's Plan (E-Mix) is a song by rapper Drake
611	Real Madrid MATCH	[none]	NaN	Real Madrid vs Real Betis 5-3 - Real Madrid vs Real Betis

The table presents selected examples where the generated descriptions demonstrate strong semantic alignment with the video titles and tags across different content domains.

As shown in Table 1, the generated descriptions effectively capture the core semantic meaning of the video content. The model performs particularly well in structured domains such as music, sports, and entertainment, where clear contextual patterns are available. These results confirm the ability of the proposed approach to generate coherent and contextually relevant textual representations. To ensure a fair and unbiased evaluation, all preprocessing and embedding generation steps are performed using the training dataset only, preventing any form of data leakage. The resulting embeddings are evaluated under identical experimental conditions to ensure a controlled comparison between incomplete and enhanced textual representations.

The enhanced dataset, referred to as

“Enhanced Trending YouTube Video Statistics”, has been publicly released on Kaggle to ensure transparency, reproducibility, and accessibility for future research. This public release further supports the reliability

of the proposed methodology by enabling independent validation and reuse of the dataset. https://www.kaggle.com/datasets/hamsahameedyousif/generated-missing-descriptions-trending-youtube

3.3.2 Visual Feature Extraction using CLIP

Visual features are utilized to capture content-based signals available in the Early Stage through video thumbnails. Thumbnail images provide important visual cues that influence user attention and engagement prior to publication. To extract visual features, a pre-trained vision-language model (CLIP) is employed. The CLIP image encoder generates fixed-length embeddings of size 512 by leveraging large-scale image-text paired training, enabling the representation of high-level visual semantics aligned with human perception [5]. To ensure data quality, thumbnail URLs are validated prior to processing, and only valid images are encoded. A consistent extraction and filtering procedure is applied across all experiments to maintain fairness and comparability between different feature configurations.

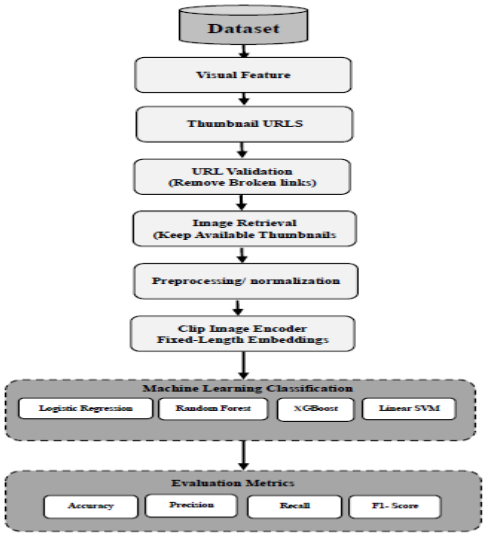

Figure 3: CLIP-Based Visual Feature Extraction Pipeline for Thumbnail Images

The resulting CLIP embeddings are used as input features for the supervised classification models, enabling the integration of visual information into the prediction framework.

3.4 Metadata Feature Engineering

Metadata features are incorporated to capture structural attributes associated with each video, complementing textual and visual representations. These features are selected based on their relevance to video characteristics and their availability in the dataset. In this study, the metadata feature set used in the experiments consists of the following attributes: category_id and categoryId as categorical features, comments_disabled, ratings_disabled, and video_error_or_removed as binary indicators, and title_length, description_length, and tags_count as numerical features.Categorical features are transformed into numerical representations using one-hot encoding, while binary and numerical features are scaled to ensure compatibility with the classification models. This structured representation enables the integration of metadata features within the multimodal framework under consistent experimental conditions.

3.5 Multimodal Feature Fusion (Text + Visual Features)

Multimodal feature fusion is applied in the Early Stage for the amalgamation of information in text and visual formats. In the experiment, text features are collected from the title, tags, and the description of the video, where DeBERTa is employed, providing fixed-length contextual embeddings of size 768. Visual features are obtained from the thumbnail of the video via CLIP, an image encoder, providing fixed-length visual embeddings of size 512.

After representation collection, DeBERTa and CLIP embeddings are synchronized in clips. When each video is assigned a text and thumbnail representation, these fused representations, obtained via the amalgamation of semantic text information and visual textual information, allow the model to capture richer content-based signals beyond the modalities.

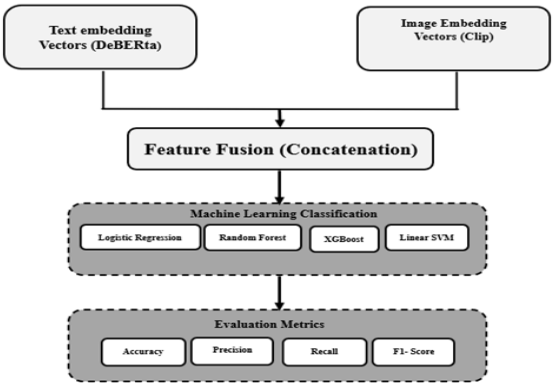

Figure 4: Multimodal Fusion of DeBERTa Textual Features and CLIP Visual Features in the Early Stage

The merged feature vectors obtained are used for plotting a supervised classification model under the same experimental conditions, which are applied to image and text single-modality experiments. Such a design is controlled and allows a comparison of text-modality, image-modality, and text-image fusion cases. It also allows an assessment of the Early-Stage prediction feature to see if it improves performance, along with the predictive use of textual and visual information, each functioning independently.

3.6 Multimodal Feature Fusion (Text + Visual + Metadata Features)

Multimodal feature fusion is applied in the Early Stage for amalgamation of information in text, visual, and metadata formats. Textual features are extracted using DeBERTa to create video title, tag, and description embeddings of dimension 768 using contextual fixed-length embeddings. The video thumbnail embeddings of dimension 512 using a fixed-length embedding of the CLIP image encoder are used to extract visual features. Features of metadata are constructed through the use of structured attributes slots that can either be categorical, binary, and/or numerical features that are transformed and scaled to a common scalar to create a metadata scalar feature vector. By extracting the features, representing text, video imagery, and metadata at the video level, visual, text, and metadata representations are all consistent. Each video is an entity and is represented in

an aggregate manner. Each video is represented as a single entity, and the representations of text are described semantically. The visual elements are represented from a video thumbnail, and the elements of metadata are described structurally

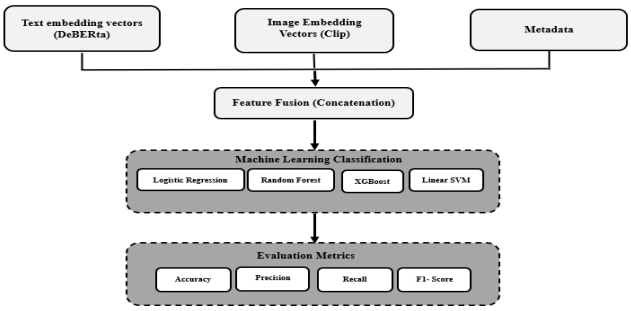

Figure 5: Multimodal Fusion of Textual (DeBERTa), Visual (CLIP), and Metadata Features in the Early Stage

This results in a more holistic representation of video content. The resultant multimodal feature vectors provide the input to the supervised classification models under consistent experimental circumstances. This allows for fair comparisons to be made between modalities. To feature fusion, enhances prediction from the various features more so than the individual features.

3.7 Interaction Stage

Interaction features are techniques that assist in measuring user response after content goes live. These capture additional nuances and engagement metrics of the audience with the video content. Interaction features, in this study, are formulated based on user engagement metrics in terms of people who liked, commented, and interacted, and the ratios of these engagements are derived from the different metrics of engagement.In this case, view count is expressly omitted from the interaction features to avoid unrealistic evaluations and target leakage by ensuring the model has no direct indicators of popularity to rely on, thereby preserving the fairness of the experimental setup.

Interaction features are outlined in a numerical representation that is structured, in the Early Stage of the study, and the same experimental conditions are employed for the content-based and interaction-based modeling approaches, fostering an equitable environment for comparison. While maintaining realistic boundaries, the outlined interaction features assist the model to simulate the dynamics of engagement that post-publication, thereby helping in the assessment of the interaction features.

3.8 Machine Learning Models and Evaluation

Multiple classification models in the proposed framework, such as Logistic Regression, Random Forest, XGBoost, and Linear SVM, incorporate diverse learning styles and non-linear modeling approaches to ensure an all-encompassing evaluation. All models are developed and evaluated using an even 80/20 train-test split with the same experimental conditions to make the results generalizable to new data. All models also employ the same feature representation and data processing techniques to ensure an even playing field. The models are deemed to have achieved a

balanced level of prediction qualitization once the results of the different models are evaluated by the same prediction level metrics. Also, confusion matrices are used to analyze results of a classification task and give great detail on˜ which predictions are correct and which are incorrect. Specifically, this type of evaluation structure offers a way for the author to clearly and reliably interpret the performance across various models and combinations of features and to make comparisons across various facets of this evaluation.

4. Experimental Results

This section describes the evaluation of the YouTube video popularity prediction model using the stage-aware architecture developed by the authors. Because of the importance of fairness and objectivity in evaluating the models and the different feature sets, all experiments were conducted using an 80/20 splitting of data into the training and testing sets, and all experiments were performed in a similar manner.

The authors assessed the performance of the models with four different classification methods, namely, RF, XGBoost, Logistic, and SVM. Different metrics of classification, including the accuracy, precision, recall, and F1 score, among others, were used to assess the performance of the models, thus providing detailed and objective evaluation of the efficiency of classification.

4.1 Text-Based Results

4.1.1 Text-Based Results using DeBERTa with Missing Descriptions

The authors assessed the performance of these models with the use of text features in two main settings: the original dataset (Trending YouTube Video Statistics) and the augmented dataset with the use of the descriptions. In order to clearly present the performance of the models in different scenarios, the authors present the results in two subsections. Table 2, with the section labeled "Text Based Results in Empty Descriptions" shows the performance of models using the DeBERT Based text features with descriptions nullified.

Table 2: Text-Based Performance (Empty Values)

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.9481	0.8456	0.9044	0.8740	0.9710
XGBoost	0.9307	0.8835	0.7507	0.8117	0.9577
Logistic Regression	0.7427	0.4244	0.8222	0.5598	0.8430
Linear SVM	0.7562	0.4402	0.8284	0.5749	0.8558

From Table 2, among all models, the RF model showed the best performance in terms of classification because it achieved an accuracy of 0.9481 and an AUC of 0.9710, indicating an excellent classifying procedure. In addition, XGBoost also showed excellent performance, while the Logistic Regression and Linear SVM models had the worst results.

This goes to show that contextual embeddings introduce complex non-linear rules that ensemble-based models capture well, while linear models leave much to be desired when it comes to the depth of the learned representations.

In order to better understand the behavior of the best performing model, we illustrate the Random Forest's ROC curve and confusion matrix in Figure 6.

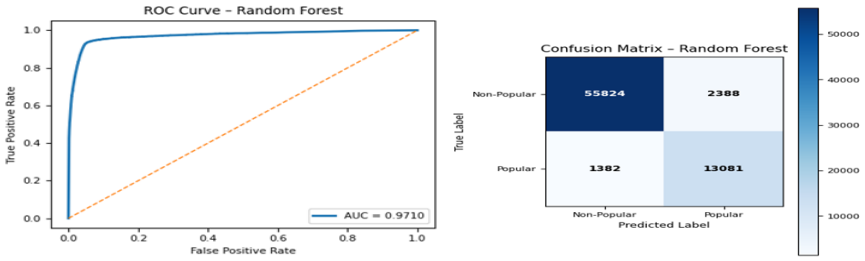

Figure 6: ROC Curve and Confusion Matrix for Random Forest

The ROC curve verifies that the Random Forest model has a strong ability of discrimination and the confusion matrix displays a better rate of correct classification with a much lower rate of misclassification. The Random Forest model outperformed linear models. Figure 7 shows all models with ROC Curves for each model.

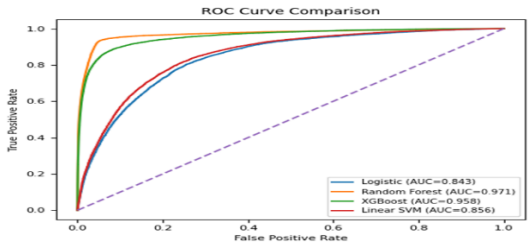

Figure 7: ROC Curve Comparison of Text-Based Models

In line with Figure 7 and in AUC terms, Random Forest outperformed all, with XGBoost being the second closest and both the Logistic Regression and Linear SVM being the lowest. It is obvious that tree-based models outperformed linear models in the use of the textual features. The assessment of the proposed model's effectiveness was achieved through the comparison of the Random Forest to the average of the YouTube popularity prediction models, using a classification-type machine learning model (the "YouTube Videos Prediction: Will This Video Be Popular?") to further understand the YouTube Video machine learning models that attained a high audience in YouTube based on the classification-type prediction.

This study lacked the use of high order contextual embeddings and employed ordinary textual features as opposed to the proposed one. This study claimed a Random Forest model accuracy of about 0.85, while the proposed model attained 0.94. The advancement has occurred through the usage of contextual embeddings (DeBERTa). They are able to provide a denser semantic representation, as opposed to older

techniques of model feature extraction, which allow more efficient modeling of text.

4.1.2 Text-Based Results using DeBERTa with Generated Descriptions

To build on the findings of texture feature enhancement, the dataset with generated descriptions was used to conduct additional trials.

Table 3: Text-Based Performance (Enhanced Dataset)				
Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9463	0.8624	0.8687	0.8655
XGBoost	0.9418	0.8738	0.8271	0.8498
Logistic Regression	0.8325	0.6471	0.3479	0.4525
Linear SVM	0.8271	0.7585	0.1924	0.3070

Table 3 shows that Random Forest and XGBoost exhibit excellent and consistent performance, achieving accuracies of 0.9463 and 0.9418, respectively. The results suggest that generated descriptions offer an improvement in the representation of text. When compared to the original dataset, the overall accuracy remains consistent, however certain models exhibit improved precision and F1-score, with evidence of improved predictive performance. Random Forest remains consistent and stable, suggesting it is robust to feature modifications, while linear models tend to exhibit a loss in recall, indicating an overall sensitivity to feature distribution. The results suggest that enriched
textual completeness contributes to more comprehensive semantic representations allowing the models, and more importantly the non-linear models, to better identify video popularity through the development of complex patterns.

4.2 Image-Based Results (CLIP)

The models were trained using the visual features extracted from the YouTube video thumbnails using the CLIP (Contrastive Language-Image Pre-training) model. An analysis was performed on these features to determine the usefulness of the visual features viewed in the context of video popularity prediction and to gain insights into the behavior of the models for each of the machine learning algorithms employed.

Table 4: Image-Based Performance (CLIP Features)				
Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.7935	0.1984	0.0086	0.0165
XGBoost	0.6436	0.2021	0.2616	0.2280
Logistic Regression	0.7972	0.2514	0.0039	0.0077
Linear SVM	0.5790	0.2009	0.3670	0.2597

Table 4 shows that most models performed better using textual rather than visual features. Logistic Regression (79.72%) and Random Forest (79.35%) models obtained the higher accuracy scores, but the models showed almost zero recall and F1-score, and hence, a certain majority class bias. It can be assumed that the models failed to recognize the popular videos by leveraging the visual features only. The models XGBoost and Linear SVM performed better than the models Logistic Regression and Random Forest in terms of capturing the patterns of minority class members. However, their total accuracy was not impressive, and their precision and recall scores would not be medically justified. The results indicate that the early-stage popularity prediction models failed to reach the accuracy level in terms of popularity prediction. This reflects the necessity of merging the visual features with the other types of features, in particular textual features.

These results were further strengthened by a benchmark taken with a recent study (A Multimodal Feature Fusion Model for User Interest Prediction) that applied CNN-based visual features, namely ResNet50, to the image-based prediction task. The best accuracy for the study was 0.3935 (image-only features), and the proposed approach achieved higher accuracy scores while using CLIP-based representations. This difference is due to CLIP being able to grasp the complex semantic relationships that exist between different types of visual content and the words used to describe them, leading to representations of features that are more detailed than those offered by classic CNN methods.

However, low recall values show that even though the image-only models are more accurate, they are still not successful at finding the most popular videos.

4.3 Metadata-Based Results

In this case, we are analyzing the features of the metadata and trying to understand their potential to predict the popularity of YouTube videos by themselves. Some of these features are categorical (like category identifiers), others are binary (like comments disabled, ratings disabled, and has been video error or removed), and some are numerical (like title, description, and number of tags). Here, we try to understand the potential of structured metadata to predict video popularity compared to content features.

Table 5: Metadata-Based Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9422	0.8195	0.9105	0.8626
XGBoost	0.8453	0.7177	0.3668	0.4855
Logistic Regression	0.8064	0.5415	0.1760	0.2657
Linear SVM	0.6808	0.3425	0.6572	0.4504

As seen in Table 5, performance using metadata features are mediocre in comparison to text-based results. Random Forest and XGBoost have comparatively better results, while Logistic Regression and Linear SVM provide poorer results. This suggests metadata features are useful in providing structural features about videos, but do not have the semantic depth required for modeling content-related patterns associated with video popularity. Metadata features are not as expressive as the textual features, as they do not express the semantic content of the video. Thus, the models are poor in determining sophisticated patterns about the content and popularity. In conclusion, results show metadata features are not enough to get high prediction performance. Yet, they can be useful as additional features in combination with the textual and visual features

4.4 Interaction-Based Results

This section explains the effects of interaction-based features in the context of predicting YouTube video popularity. Interaction-based features, which explain

user engagement and how well the video is likely to perform, provide excellent potential for predicting video popularity. The interaction features are made up of engagement signals (likes, comments) and tailored features such as log_likes, log_comments, like_dislike_ratio, and interaction_score. In order to conduct a fair and realistic assessment, view

count is removed from the feature set to prevent target leakage and are considered the only direct engagement features.

Table 6: Interaction-Based Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9482	0.8814	0.8549	0.8680
XGBoost	0.9250	0.8467	0.7608	0.8014
Logistic Regression	0.9098	0.8551	0.6584	0.7440
Linear SVM	0.9019	0.7071	0.8655	0.7783

Table 6 shows how Random Forest achieved the highest metrics involving interaction-based features, accuracy, and balanced precision and recall. XGBoost also performed well; however, Logistic Regression and Linear SVM performed somewhat lower, but competitive. Interaction-based features enhance prediction metrics more compared to the previous features. This shows intricate user engagement, as likes and comments, highly correlates with the popularity of videos. Interaction-based features, in contrast, emphasize the aftermath behavior of users, more so than the textual and visual features, and can be useful as an informative prediction metric. These features represent active engagement behavior, which is user behavior and can predict the popularity of the videos. To further validate these hypotheses, results from the three previous studies that consider engagement-based interaction as a major business component. Prior studies achieved accuracy, F-measures, and ROC-AUC in the engagement signals of likes, dislikes, and comments. For example, the (YouTube Trending Videos Prediction & Analysis) study that used SVM, KNN, and
Random Forest, which achieved values of 0.82, 0.85, and 0.88, respectively. Likewise, Interaction and Metadata features with the Application of a Decision Tree (YouTube Video Trending Analysis Based on ML) study achieved an accuracy of 0.877. In the absence of a view count, the proposed model achieved a 0.9482 accuracy. This result suggests high metrics achieved in the absence of view-based component reliance.

4.5 Multimodal Fusion Results

4.5.1 Multimodal Results (Text + Image)

n this study, the features of DeBERTa and CLIP were merged, DeBERTa provided the textual features and CLIP provided the visual features. Textual features were expressed as a 768-dimensional vector, and visual features as a 512-dimensional vector. The vectors from each video were concatenated,

producing a combined 1280-dimensional vector. This method aims to surpass the performance of the unimodal models, achieving better prediction performance by incorporating the rational and empirically supportive information from both text and images.

Table 7: Text and Image-Based Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9334	0.8148	0.8612	0.8373
XGBoost	0.9299	0.8865	0.7426	0.8082
Logistic Regression	0.8316	0.6530	0.3277	0.4364
Linear SVM	0.8401	0.6815	0.3689	0.4787

Table 7 illustrates the results of the multimodal models illustrating positive and consistent performance of the classifiers. Random Forest was able to achieve the strongest results, obtaining an overall performance of 0.9334 and an F1-score of 0.8373, closely followed by XGBoost. The results show a significant improvement when compared to the image-only models. The textual features provide the necessary and critical reasoning that predict the next pieces of information and provide the strongest results. When compared to models using text only, the results suggest a more careful adjustment of balance. It illustrates a more stable outcome of both precision and recall and better results. The results suggest a more careful adjustment of balance compared to models with text only.
The results illustrate the text features with the more dominant forms whereas the visual information features represent a more constitutive supportive element that predict the next pieces of information. The results illustrate that by combining the text and visual cues from the thumbnail, multimodal fusion results provide the models with a more extensive representation of the patterns. Compared to single-modality models, this combination offers better representation of video content and aids in generalization. Analyses show that, for prediction models, incorporating both text and images yield better results. Though better than text-only models, the improvement in image-only models is significant, but the multimodal approach offers in-depth insight in video characteristics.

4.5.2 Multimodal Results (Text + Image + Metadata)

By incorporating text, visuals, and metadata, the models capture content and context. Textual representations take the form of DeBERTa embeddings (768 dimensions), visual features are captured by CLIP (512 dimensions), and metadata input is provided in the form of a structured representation. All features are concatenated to create a unified representation for each video. Utilizing a structured contextual representation of metadata and complementary information from various types of features is likely to enhance prediction performance.

Table 8: Text, Image, and Metadata-Based Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9488	0.8495	0.9026	0.8752
XGBoost	0.8980	0.8486	0.5935	0.6985
Logistic Regression	0.8490	0.6815	0.4532	0.5444
Linear SVM	0.6642	0.3612	0.8944	0.5146

Given the information in table 8, Random Forest performs better on all prediction metrics (0.9488 in accuracy and 0.8752 in F1-score) than all other measures, including the previous multimodal models. The addition of metadata features shows a slight and consistent improvement on the Text + Image model, suggesting that structured contextual information adds value to prediction performance. The proposed model is further evaluated by comparing it to preceding models. The study (YouTube Case Study: Comparative Analysis of ML and ANN Models for View Prediction) reported a Random Forest accuracy of 0.91. Furthermore, the study (Machine Learning Enabled Models for YouTube Ranking) reported a Random Forest accuracy of 0. 784.In comparison, the proposed model in this work reported a Random Forest accuracy of 0.9488. This improvement can be attributed to the incorporation of multimodal features, especially the union of the textual, visual, and metadata representations.

4.5.3 Multimodal Results (Full Feature Integration)

In this case, the full multimodal framework looks at how the combination of textual, visual, metadata, and interaction features all folded into a single representation. Such a combination captures the semantic and behavioral aspects of popularity of a video, leveraging content and interaction features both prior and post-release, respectively. The demand of this arrangement is to push the boundaries of the predictive capabilities of the model, when all feature types are at a researcher's disposal. The results are summarized in Table 9. The column headings are defined as follows. Accuracy is defined as the number of correct predictions divided by total predictions. Precision is defined as the number of correct predictions divided by total predicted positives. Recall is defined as the number of predicted positives divided by total actual positives. Finally, the F1-Score is defined as the harmonic mean of precision and recall

Table 9: Full Multimodal Performance (Text + Image + Metadata + Interaction)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9848	0.9972	0.9685	0.9827
XGBoost	0.9783	0.9997	0.9517	0.9751
Logistic Regression	0.9734	0.9890	0.9510	0.9697
Linear SVM	0.9745	0.9928	0.9497	0.9708

From Table 9, we see the full multimodal model yielding the best performance across all configurations. The Random Forest model has the numerical edge on all models exhibiting performance across all configurations yielding almost perfect statistical correctness, and XGBoost was second best to the chain while Logistic Regression and Linear SVM were almost perfectly correct, respectively. It can be said that the integration of interaction features vastly improves the performance lightyears in the context of the earlier multimodal architectures due to the strong predictive power of engagement-based signals. The interaction features being added to the Text + Image + Metadata model is a key point of the model evidencing almost perfect correctness. Finally, it is fair to say that the above configuration is more numerous due to the post-based implementation which includes the interaction-based signals as almost perfect correctness predictor signals used. Which support the claim of the predictive power of engagement-based signals and interaction features. It is fair to state the focus of the above

implementation in is beneficial and Fuller in terms of research of popularity of videos due to the interaction features. Consequently, the results achieved here delineate a ceiling rather than a first-order prediction case of a nascent system. Herein, all results illustrate predictive performance of the 'best' Multimodal model, yet because of its exclusive reliance on post-engagement interaction attributes, its usefulness is confined to a case of Predictive Retrospective Analysis, rather than Real-time prediction of post-published videos. The results substantiate the significance of differentiating early-stage prediction from post-publishing predictions and illustrate the efficacy of the stage-aware, multimodal framework model to both prediction situations thus, fulfilling the core motivation of the framework.

5. Discussion

This part of the document assesses the proposed model's performance for various types of features and their roles in predicting YouTube video popularity. Different types of features such as textual, visual, metadata, interaction, and multimodal features have shown to have varying predictive performance in experiments. DeBERTa contributes to textual features and performs well in early-stage predictions. The strong performance of text-based models shows that the video's popularity is significantly affected by the semantic-rich, text-based informative content such as the video title, description, and tag that are available pre-publication. Contextual embedding improves the textual feature's representation, and the models capture text semantic relations more deeply than the legacy text features. Visual features shown to have limited predictability in thumbnails captured by CLIP when these features used in isolation. Accuracy of image-based models is at a decent level; however, the models show low F1-score and recall. This indicates class imbalance. This shows that popularity of video is captured better by other features than thumbnails alone. This shows that popularity of video is captured better by other features than thumbnails alone. Predictably, metadata features have shown to have moderate predictably. Category, title length, and video settings have shown to have depreciating depths of semantic relations predictive capability. This shows that these elements have limitations in capturing relations that are semantic. Compared to textual features, metadata features have limitations and provide less content representation. Features based on interaction achieve the highest predictive performance compared to other modalities. Popularity correlates with engagement, likes, and comments. Unfortunately, interaction-based features are only available after publication, which limits them for early-stage prediction. To test the model without the risk of target leakage, the required view count was omitted. Nevertheless, interaction features are informative and improve the model significantly. The combination of interaction, textual, visual, and metadata features lead to the highest accuracy and model performance and stability improvement. The results support the statement that integrating data from multiple sources helps the model capture the information that other forms don't capture. The results of all the experiments support the claim that interaction-based modalities achieve the highest model performance compared to textual features. Nonetheless, the use of these features depends on the data after publication, which limits the use of earlier predictions. In comparison, early prediction with textual features, especially when combined with other modalities, is achievable. These results support the claim that a stage-aware prediction framework is required that considers the available features from each prediction stage.

6. Conclusion and Future Work

6.1 Conclusion

This study has introduced a stage-aware multimodal framework for predicting YouTube video popularity while addressing critical gaps in existing methodologies related to feature accessibility and target leakage. By subdividing prediction into early and interaction stages, the framework fosters a structured design toward video popularity prediction. DeBERTa-processed textual features rendered the strongest outcomes in early-stage prediction. This was in contrast to the comparative effectiveness of visual and metadata features. For early-stage popularity prediction, visual features, in particular, were quite ineffective, while metadata features, due to their structural information, were of moderate value. Interaction-based features produced the best results in terms of video popularity prediction. However, their use in early-stage predictions was limited due to the post-publication constraints on their input. Early-stage prediction evaluation excluded view-based features due to target leakage, to allow post-publication metrics and measure. The multimodal framework test was able to show an increase in both the robustness and stability of the model due to the addition of textual, visual, and metadata features. This, in conjunction with interaction features, provided the best overall results. It is clear from this study that the combination of multimodal and stage-aware components positively influences predictive measures, while also allowing for realistic evaluation.

6.2 Future Work

While the framework presented is quite successful, there is plenty of room for further enhancement. For example, the incorporation of advanced models in multimodal deep learning, especially the attention-based architectures, can be predictive of an enhanced experience, as a result of enhancements in feature fusion and representation learning.

Second, the understanding and prediction of the popularity trend of videos could improve with the inclusion of traditional temporal modeling techniques. Third, prediction accuracy may improve with the inclusion of additional contextual features, such as user behavior, channel-level features, and recommendation signals. Finally, building the proposed framework into actual applications, such as content recommendation systems or creator support systems, will enhance understanding of the framework's scale, and how robust and effective it is in real-world applications. Finally, building the proposed framework into actual applications, such as content recommendation systems or creator support systems, will enhance understanding of the framework's scale, and how robust and effective it is in real-world applications.

7. Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research.

References

[1] S. Yang, D. Brossard, D. A. Scheufele, and M. A. Xenos, “The science of YouTube: What factors influence user engagement with online science videos?,” *PLoS One*, vol. 17, no. 5 May, May 2022, doi: 10.1371/journal.pone.0267697.

[2] “YOUTUBE TRENDING VIDEOS PREDICTION & ANALYSIS,” *International Research Journal of Modernization in Engineering Technology and Science*, May 2023, doi: 10.56726/irjmets38620.

[3] R. Y. Kharkar and F. Schoenberg, “Predicting and Characterizing Early Growth of YouTube Videos.”

[4] M. Limpijankit and J. Kender, "Detecting Cultural Differences in News Video Thumbnails via Computational Aesthetics," Nov. 2025, doi: 10.36190/2024.61.

[5] K. Xu, Z. Lin, J. Zhao, P. Shi, W. Deng, and H. Wang, "Multimodal Deep Learning for Social Media Popularity Prediction with Attention Mechanism," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 4580–4584. doi: 10.1145/3394171.3416274.

[6] L. Ye *et al.*, "HyperFusion: Hierarchical Multimodal Ensemble Learning for Social Media Popularity Prediction," Jul. 2025, [Online]. Available: http://arxiv.org/abs/2507.00926

[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 *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, Association for Computing Machinery, Inc, Apr. 2018, pp. 175–178. doi: 10.1145/3184558.3186972.

[8] Y. Ma, "Construction and Data Analysis of a New Media Content Popularity Prediction Model Based on Naive Bayes Algorithm," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 294–302. doi: 10.1016/j.procs.2025.04.207.

[9] M. Ziyada and P. Shamoi, "Video Popularity in Social Media: Impact of Emotions, Raw Features and Viewer Comments," Jul. 2024, doi: 10.1109/SCISISIS61014.2024.10759978.

[10] Y. Xu *et al.*, "SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity." [Online]. Available: https://github.com/zhuwei321/SMTPD

[11] D. S. Villegas, D. Preo¸tiucpreo¸tiucPietro, and N. Aletras, *Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks*. [Online Video]. Available: https://github.com/dan

[12] Y. Li, K. Eng, and L. Zhang, "YouTube Videos Prediction: Will this video be popular?"

[13] 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 (Switzerland)*, vol. 10, no. 23, Dec. 2021, doi: 10.3390/electronics10232962.

[14] "Predictive analysis of YouTube trending videos using Machine Learning."

[15] Z. Xu and M. Qian, "Predicting Popularity of Viral Content in Social Media through a Temporal-Spatial Cascade Convolutional Learning Framework," *Mathematics*, vol. 11, no. 14, Jul. 2023, doi: 10.3390/math11143059.

[16] Y. Chen, Y. Wang, and R. Tan, "Classifying YouTube Videos by Thumbnail."

[17] E. Ramalakshmi, A. B. S. Reddy, and S. G, "YouTube Data Analysis and Prediction of Views and Categories," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 6, pp. 568–573, Jun. 2022, doi: 10.22214/ijraset.2022.43636.

[18] M. S. Irshad, A. Anand, and M. Ram, "Trending or not? Predictive analysis for youtube videos," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 4, pp. 1568–1579, Apr. 2024, doi: 10.1007/s13198-023-02034-8.

[19] Z. Liu, "Youtube Video TrendingAnalysis Based on Machine Learning," 2023.

[20] H. Jang, S. H. Kim, J. S. Jeon, and J. Oh, "Visual Attributes of Thumbnails in Predicting Top YouTube Brand Channels: A Machine Learning Approach," in *Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6660–6662. doi: 10.1109/BigData55660.2022.10020875.

[21] C.-C. Wang, Y.-F. Lin, Y.-C. Hsieh, and Y.-H. Kao, "USING VIDEO TITLES TO PREDICT YOUTUBE AUDIENCE BEHAVIOR BASED ON MACHINE LEARNING," *J. Theor. Appl. Inf. Technol.*, vol. 103, no. 8, 2025, [Online]. Available: http://surveys.tw,

[22] H. Liu *et al.*, "Multi-Modal Video Feature Extraction for Popularity Prediction," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.01422

[23] A. Javed, N. Abid, M. Shoaib, M. F. Shahzad, F. Sabah, and R. Sarwar, "A Framework to Predict the Quality of a Video for Popularity on Social Media," *Engineering Reports*, vol. 7, no. 6, Jun. 2025, doi: 10.1002/eng2.70250.

[24] K. Xu *et al.*, "Higher-Order Vision-Language Fusion for Video Popularity Prediction," in *MM 2025 - Proceedings of the 33rd ACM International Conference on Multimedia, Co-Located with MM 2025*, Association for Computing Machinery, Inc, Oct. 2025, pp. 14086–14093. doi: 10.1145/3746027.3763762.