

Automated YouTube Video Success Analyzer Using AI

Ms. Praveena K¹ Mohammed Sidhan NM², Aswin P³, Sreesha M⁴, Vishnu S⁵,

¹Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, India

²Professor and Dean, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, India

Abstract — YouTube is one of the most popular platforms for sharing video content, with millions of videos uploaded daily. Predicting which videos will become viral is a significant challenge. This project develops a machine learning-based system to predict the popularity of a YouTube video before it is published by analyzing factors such as thumbnail, hashtags, description, and video quality. Convolutional Neural Networks (CNN) are used to evaluate thumbnail attractiveness, while Natural Language Processing (NLP) analyses textual elements like hashtags and descriptions. The system also considers video quality attributes such as resolution and duration. Using tools like TensorFlow, Scikit-learn, OpenCV, and the YouTube Data API, the model predicts engagement potential and provides suggestions to improve video reach. This proactive approach helps content creators optimize their videos before publishing and demonstrates practical applications of AI in social media analytics.

Keyword - Machine Learning, CNN, NLP, Video Popularity Prediction, Recommendation System, Recommendation System, Multimodal Learning, Social Media Analytics, Content Optimization, Computer Vision, Text Analysis, Feature Extraction, Trend Analysis, Predictive Modelling, Engagement Prediction, Transfer Learning, Sentiment Analysis, Click-Through Rate (CTR)

I. INTRODUCTION

YouTube is one of the most widely used platforms for sharing video content, with billions of active users and millions of videos uploaded daily. However, only a small percentage of videos achieve viral success, making it difficult for content creators to predict performance before publishing. This creates a need for intelligent systems that can analyze content and provide predictive insights in advance.

Machine Learning (ML) and Artificial Intelligence (AI) have shown significant potential in analyzing patterns and predicting outcomes in large datasets. In the context of YouTube, factors such as thumbnail design, hashtags, video descriptions, and video quality play a critical role in influencing user engagement. These features directly impact metrics like click-through rate and audience retention.

This project proposes a machine learning-based system that predicts the popularity of unpublished YouTube videos by analyzing both visual and textual features. Convolutional Neural Networks (CNN) are used for thumbnail analysis, while Natural Language Processing (NLP) techniques are applied to textual elements. The system also includes a recommendation engine that provides actionable suggestions to improve content performance before publishing. Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

II. RELATED WORK

1. Traditional Approaches for Video Popularity Prediction

Early research in video popularity prediction primarily focused on statistical and regression-based methods. These approaches utilized metrics such as views, likes, comments, and sharing patterns to estimate the success of online videos. Linear regression, decision trees, and support vector machines were commonly applied to analyze relationships between user engagement and video performance. These methods were effective for analyzing already published content but lacked the ability to predict outcomes before video release.

However, these traditional models had significant limitations, as they relied heavily on post-publication data. They were unable to capture content-specific features such as thumbnail attractiveness or textual relevance. As a result, they provided limited insights for content creators who wanted to optimize their videos before uploading. This limitation led researchers to explore more advanced approaches using deep learning and multimodal analysis.

2. Deep Learning Techniques in Image Analysis

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have become widely used for image-based analysis. CNN models such as ResNet and MobileNet are capable of extracting complex visual features from images, including shapes, textures, and object patterns. These models have been successfully applied in tasks like object detection, image classification, and visual attractiveness prediction.

In the context of YouTube videos, thumbnails play a crucial role in attracting viewers. Research has shown that visually appealing thumbnails significantly improve click-through rates. CNN-based models can automatically analyze thumbnail features such as brightness, contrast, and layout, making them highly effective for predicting video engagement. This capability forms a key component of modern video popularity prediction systems.

3. Natural Language Processing for Text Analysis

Natural Language Processing (NLP) techniques are widely used to analyze textual content such as titles, descriptions, and hashtags. Methods like TF-IDF and BERT embeddings help in extracting meaningful features from text,

including keyword relevance and semantic context. These techniques allow models to understand the importance of specific words and phrases in relation to user search behavior.

In video platforms like YouTube, textual metadata plays a significant role in discoverability. Optimized titles and descriptions improve search rankings and audience reach. NLP-based models can analyze trends, sentiment, and keyword frequency to predict how well a video will perform. This makes NLP an essential component in multimodal prediction systems.

4. Multimodal Learning Approaches

Recent studies have focused on combining multiple data sources, such as images, text, and audio, to improve prediction accuracy. Multimodal learning integrates features from different domains to provide a more comprehensive understanding of content. By combining visual features from thumbnails and textual features from captions, these models achieve better performance than single-modality approaches.

Multimodal systems are particularly useful for platforms like YouTube, where content performance depends on multiple factors. For example, a video with a strong thumbnail but weak title may not perform well. By analyzing both aspects simultaneously, multimodal models can capture complex relationships between different features, leading to more accurate predictions.

5. Audio and Video Processing Techniques

Audio and video analysis have also gained importance in recent research. Tools like Librosa are used for audio signal processing, enabling the extraction of features such as pitch, tone, and intensity. Similarly, frameworks like Whisper are used for speech-to-text conversion, allowing models to analyze spoken content within videos.

Video processing libraries such as MoviePy are used to extract frames and metadata from videos. These features help in understanding content quality and structure. Integrating audio and video analysis provides additional insights into content performance, further improving prediction models. However, many existing systems do not fully utilize these features, leaving room for improvement.

6. Limitations of Existing Systems and Research Gap

Despite significant advancements, existing systems still have several limitations. Many models rely on post-upload engagement metrics, making them unsuitable for predicting video success before publication. Additionally, some systems focus only on a single modality, such as text or image, without considering the combined impact of multiple features.

Another major limitation is the lack of actionable recommendations. While many models can predict popularity, they do not provide guidance on how to improve content. This creates a gap between prediction and practical

application. The proposed system addresses these limitations by integrating CNN, NLP, and multimodal analysis while also providing improvement suggestions for content creators.

III. PROPOSED METHODOLOGY

I. SYSTEM OVERVIEW



Fig. 1. workflow diagram of the YouTube Video Virality Prediction

The proposed system is designed to predict the virality potential of YouTube videos before publishing by analysing multiple content features such as thumbnails, captions, and video data. Unlike traditional systems that rely on post-upload engagement metrics, this approach focuses on pre-upload analysis, enabling creators to optimize their content in advance. The system takes input from the user in the form of a thumbnail image, video captions, and optionally the video file, and processes these inputs using deep learning techniques.

The system follows a multimodal architecture, combining visual, textual, and audio features for improved prediction accuracy. A Convolutional Neural Network (CNN) is used to extract visual features from thumbnails, while Natural Language Processing (NLP) techniques analyze captions and textual content. Additionally, tools such as Whisper and Librosa are used to process audio data, and MoviePy is used for video frame extraction. These features are integrated to form a comprehensive representation of the video content.

The processed data is then passed through a prediction model implemented using TensorFlow and PyTorch frameworks. The model is trained iteratively over 15–25 epochs, allowing it to learn patterns and improve performance over time. The final output is a percentage score indicating the likelihood of the video becoming viral, along with suggestions to improve content quality and engagement.

II. Module Description

1) Data Input and Preprocessing Module

This module is responsible for collecting and preparing input data from the user. The system accepts inputs such as thumbnail images, video captions, and video files. These inputs are essential for analysing different aspects of the content, including visual appeal, textual relevance, and overall quality. The module ensures that all inputs are in the correct format before processing.

Preprocessing is performed using tools like OpenCV and Pillow for image handling, and NumPy for numerical operations. Images are resized, normalized, and enhanced to improve feature extraction. Text data is cleaned by removing stop words and irrelevant symbols, ensuring better performance in NLP analysis. This module plays a critical role in preparing high-quality input data for the deep learning model.

2) Image Analysis Module (CNN-Based)

The image analysis module focuses on extracting visual features from the thumbnail using Convolutional Neural Networks (CNN). Pre-trained models such as MobileNet or ResNet are used to identify patterns such as brightness, contrast, color distribution, and object presence. These features are crucial in determining how attractive a thumbnail is to viewers.

The CNN processes the input image through multiple hidden layers, extracting hierarchical features that represent visual characteristics. These features are then converted into numerical vectors and passed to the prediction model. This module significantly improves the accuracy of the system by incorporating visual cues that influence user engagement.

3) Text Analysis Module (NLP-Based)

The text analysis module processes captions, titles, and hashtags using Natural Language Processing techniques. Methods such as TF-IDF and BERT embeddings are used to extract meaningful features from text data. This includes keyword importance, sentiment analysis, and relevance to trending topics.

The module helps the system understand the context and quality of the textual content. By analyzing how well the captions align with current trends and audience preferences, the model can predict the effectiveness of the content. This module plays a vital role in improving discoverability and engagement prediction.

4) Audio and Video Processing Module

This module handles the processing of video and audio data using tools such as MoviePy, Librosa, and Whisper. MoviePy is used to extract frames and metadata from video files, while Librosa analyzes audio signals to extract features such as pitch and intensity. Whisper is used for speech-to-text conversion, enabling the system to analyze spoken content within the video.

By incorporating audio and video analysis, the system gains additional insights into content quality and structure. These features complement visual and textual data, making the prediction more accurate. This module enhances the system's ability to perform comprehensive content analysis.

III. Use-Case Illustration

The primary use case of the proposed system involves content creators who want to evaluate their videos before publishing. A user uploads a thumbnail image, enters captions or hashtags, and optionally provides the video file. The system processes these inputs and predicts the potential success of the video in terms of engagement percentage.

For example, a creator preparing a YouTube video can use the system to analyze whether their thumbnail is visually appealing or if their captions are optimized for search visibility. Based on the prediction results, the system provides suggestions such as improving brightness, adding trending keywords, or modifying video length.

This use case demonstrates how the system can act as a decision-support tool for creators. By providing early feedback and recommendations, it helps users improve their content strategy, increase engagement, and enhance their chances of success in a competitive digital environment.

IV. DESIGN ANALYSIS AND DISCUSSION

The design of the proposed system focuses on creating a multimodal deep learning framework capable of analyzing different aspects of YouTube video content before publication. The system integrates image, text, audio, and video processing techniques to provide a comprehensive evaluation of video performance. This section discusses the design choices, model behavior, system efficiency, and overall performance in detail. The combination of CNN, NLP, and supporting tools ensures that the system captures both visual and contextual information, leading to improved prediction accuracy and usability.

1. Multimodal System Design

The proposed system adopts a multimodal approach, combining visual, textual, and audio features to enhance prediction performance. Unlike traditional systems that rely on a single data source, this design integrates multiple inputs such as thumbnails, captions, and video content. This allows the system to capture a broader range of features that influence video virality, resulting in more accurate predictions.

By combining different modalities, the system can overcome the limitations of individual components. For example, even if the thumbnail is weak, strong textual features may compensate for it. This complementary behavior improves the robustness of the system and ensures that predictions are based on a holistic understanding of the content.

2. CNN-Based Visual Feature Analysis

The use of Convolutional Neural Networks (CNN) plays a crucial role in extracting visual features from thumbnails. The model analyzes attributes such as brightness, contrast, color composition, and object presence. These features are essential for determining the attractiveness of thumbnails, which directly affects user engagement and click-through rates.

The design incorporates pre-trained models such as MobileNet or ResNet to improve efficiency and accuracy. These models reduce the need for large datasets and training time while maintaining high performance. The CNN-based approach enables automatic feature extraction without manual intervention, making the system scalable and reliable.

3. NLP-Based Textual Feature Analysis

The textual analysis component uses Natural Language Processing (NLP) techniques to evaluate captions, titles, and hashtags. Methods such as TF-IDF and BERT embeddings are used to extract meaningful features, including keyword relevance, sentiment, and contextual meaning. This helps the system understand how well the content aligns with user interests and trending topics.

The design ensures that textual features complement visual features in the prediction process. For instance, a well-optimized title and relevant hashtags can significantly improve discoverability, even if the visual quality is average. This integration of NLP enhances the system's ability to provide accurate and meaningful predictions.

4. Training Strategy and Model Optimization

The model is trained iteratively over 15–25 epochs, allowing it to learn patterns progressively and improve performance. During each iteration, the system evaluates performance using metrics such as accuracy, precision, recall, and F1-score. This iterative approach helps in fine-tuning the model and reducing prediction errors.

Hyperparameter tuning and feature selection are also considered in the design to optimize model performance. By adjusting parameters such as learning rate and batch size, the system achieves better convergence and generalization. This ensures that the model performs well on both training and unseen data.

V. CONCLUSION AND FUTURE WORKS

The proposed system presents an effective approach for predicting the popularity of YouTube videos before publication using a combination of Machine Learning and Deep Learning techniques. By analyzing key factors such as thumbnails, captions, and video content, the system provides a comprehensive evaluation of a video's engagement potential. The integration of Convolutional Neural Networks (CNN) for image analysis and Natural Language Processing (NLP) for text analysis ensures that both visual and contextual features are effectively captured. This multimodal approach improves prediction accuracy compared to traditional single-feature systems.

The use of additional tools such as OpenCV, Pillow, Librosa, Whisper, and MoviePy enhances the system's ability to process diverse data types, including images, text, audio, and video. Training the model over multiple epochs (15–25 epochs) allows it to learn patterns efficiently and improve performance iteratively. The system not only predicts the likelihood of a video becoming viral but also provides actionable suggestions, enabling content creators to optimize their videos before uploading. This makes the system practical and valuable in real-world applications.

Despite its advantages, the system has certain limitations. The prediction accuracy depends heavily on the quality and size of the dataset used for training. Additionally, external factors such as audience behavior, timing, and trending topics can influence video performance, which may not always be captured by the model. The use of simplified models may also limit generalization across different content categories or niches.

Future work can focus on enhancing the system by integrating real-time trend analysis using live data from YouTube and other platforms. Advanced deep learning techniques, such as multimodal transformers, can be incorporated to improve feature integration and prediction accuracy. Expanding the dataset to include additional engagement metrics like comments, shares, and watch time can further strengthen the model. Moreover, improving the recommendation engine to provide more personalized and adaptive suggestions can make the system more effective and user-friendly.

VI. REFERENCES

- [1] D. V. Lindberg and H. K. H. Lee, "Optimization under constraints by applying an asymmetric entropy measure," *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.
- [2] B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

- [3] I. Boglaev, "A numerical method for solving nonlinear integro-differential equations of Fredholm type," *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.
- [4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICLR)*, pp. 1–14, 2015.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [7] H. T. Cheng, L. Koc, J. Harmsen, et al., "Wide & Deep Learning for Recommender Systems," *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10, 2017.
- [8] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
- [9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL- HLT*, pp. 4171–4186, 2019.
- [10] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.
- [11] A. Howard et al., "Searching for MobileNetV3," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.
- [12] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations (ICLR)*, pp. 1–15, 2020.
- [13] J. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1877–1901, 2020.
- [14] A. Radford et al., "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," *OpenAI Technical Report*, pp. 1–15, 2022.