Predicting Trending YouTube Videos Using Big Data Analytics

1st Shrey Kumar Shakya
Faculty of Science and Engineering
Herald College Kathmandu
Kathmandu, Nepal
np03cs4a220474@heraldcollege.edu.np

2nd Pratik Tuladhar
Faculty of Science and Engineering
Herald College Kathmandu
Kathmandu, Nepal
np03cs4a220410@heraldcollege.edu.np

***Abstract*—Big data analytics approaches allow to extract patterns from massive social media data in order to forecast content virality. Predicting youtube videos that become trends within 2 days after upload is a challenging task, since it involves the integration of numerical, categorical and textual features, additionally affected by noise, class imbalance and high dimensionality. Current approaches lack the integration of different data types for accurate trend prediction. This research investigates the construction of a machine learning framework to predict if a youtube video will trend within 2 days after upload. The analysis is conducted on a 2023 dataset composed of 72,397 videos. We propose a pipeline combining numerical scaling, categorical encoding and TF-IDF text vectorization to represent the dataset and enhance its compatibility with machine learning models. The Random Forest model is optimized with grid search and Logistic Regression, achieving a test F1 score of 0.3439, outperforming the Logistic Regression model. Further analysis on the most relevant features shows that views and title keywords are the main predictors. This work contributes to the literature and practice by offering actionable insights on the most relevant features, and providing useful guidelines for content creators and marketers to optimize their videos' strategies.**

***Index Terms*—Big Data Analytics, YouTube Trends, Machine Learning, Random Forest, Logistic Regression, TF-IDF Vectorization, Social Media Analytics**

I. Introduction

YouTube hosts over 2 billion monthly active users in 2023 and is a key player in spreading information . The platform produces big data by extracting information from users, videos and engagement. Videos that are on the YouTube's trending page are more likely to be watched by more people. Therefore, it is important to predict which videos will be on the trending page within two days after uploading. This is a challenging task since the heterogeneous data (numeric - number of views, likes, comments; categorical - video category; text - title, channel name) should be predicted amidst noise, missing values and class imbalance. In 2023 videos that become trending have 5 million views on average while videos that are not trending have only 0.8 million, see Fig. **??** .

This work addresses the problem by designing a machine learning pipeline to predict trending YouTube videos using big data analytics. The goal is to predict videos that are likely to trend within two days after uploading. For that, we use 2023 dataset of 72397 videos. Our pipeline consists of StandardScaler for numerical features, OneHotEncoder for categorical features and TF-IDF vectorization for text, tuned with Random Forest and Logistic Regression. We provide interpretable results that can help to understand drivers of virality. The framework can be used by content creators, marketers and algorithms of social media platforms to optimize content strategies. The importance of the work is practical. Content creators can understand what predictors of trend will help them optimize the content. Marketers can target high potential videos for advertising. In addition, the study contributes to social media analytics by addressing the gap in combining different types of data and class imbalance. The contributions of the work are:

- unified pipeline for numerical, categorical and text features preprocessing;
- comparison of Random Forest and Logistic Regression with hyperparameter tuning;
- analysis of feature importance and results by categories;
- analysis of the importance of text features, namely keywords in titles, for virality.

The rest of the report is organized as follows. Section II describes the dataset. Section III reviews the related work. Section IV describes the methodology. Section V presents results and discussion. Section VI addresses ethical issues. Section VII concludes the work and suggests future directions.

II. Dataset Description

The data comes from YouTube 2023 trends and is constituted by 72397 video records scraped with the youtube api and saved in a csv file ('youtubetrends_2023.csv') with the following numerical features: views, likes, comment count, publish hour, it has categorical features: category_id, publish time, weekday, text features: title, channel title, boolean features: comments_disabled, ratings_disabled. The target 'trending' is derived from the column 'timetotrend': save 1 as trending (timetotrend = 2 days) and 0 as not trending (otherwise). As we can see below, around 10% of the videos are trending (class imbalance). Finally, we load the data in mongodb for saving and retrieving easily the data, we also do some data preprocessing like fill missing values, remove irrelevant columns (dislikes, etc. because youtube policy).

III. Related Work

Research on YouTube trend prediction includes statistical, machine learning, and deep learning methods. Statistical studies like [1] analyze the correlation between the engagement score (views, likes, etc.) and the trending status, providing insights into participatory culture on the platform. Since their linear models cannot fit nonlinear patterns, they are powerless. Machine learning methods like [2] apply Random Forest on numerical and categorical features to model viewer behavior, achieving moderate accuracy but struggling with unstructured text data [9]. Deep learning methods like LSTMs for title analysis in [3] are efficient in handling text data, particularly for specialized domains like autonomous vehicle comments, but they require a large amount of computational resources [14].

The new trend in research is integrating diverse features. For example, [4] applied TF-IDF vectorization on text and applied ensemble models on the prediction, focusing on sentiment analysis of Indonesian YouTube comments. Since Random Forest and Logistic Regression are applied, the preprocessing of the text data is not solid [13]. Other works like [5] applied supervised and ensemble models (Random Forest Regressor, Decision Tree Regressor, XGBoost) on a Kaggle dataset with 575,728 records from 5 different countries to predict view counts, highlighting the limitations of comment-based analysis. Their Random Forest Regressor reached an $R2$ score of 94.6 percent which is mainly contributed by likes, comment counts, and video category [7]. Since their dataset is from YouTube, they did not use dislike column due to YouTube's policy which is consistent with our preprocessing [11]. The difference is that their focus is on regression and they did not apply text data preprocessing. Their focus is on count data while our focus is on integrating diverse features. They did not apply any feature importance analysis [8].

Another work [6] applied gradient boosting models (e.g., LightGBM) for virality prediction and included temporal features (publish time) in their model to capture view-count dynamics. Since their focus is on numerical data, their model did not make good use of text features [10]. Another difference is that their model is not interpretable while our model is, since we included feature importance analysis [12]. Additionally, studies like [15] explore scalable video analytics, emphasizing the need for efficient processing of large datasets, which aligns with our use of MongoDB for data storage.

Our contributions of this paper are summarized as follows.

- Applying numerical, categorical, and text features in one model.
- Tuning parameters of Random Forest and Logistic Regression with grid search.
- Analyzing feature importance and category performance.

Our model improves the prediction accuracy and interpretability of the model, and our model can be used as a powerful tool for content creators.

IV. Methodology

The five-phases methodology illustrated in Fig. 1 are: Data Collection, Preprocessing, Feature Engineering, Model Training, and Model Evaluation.

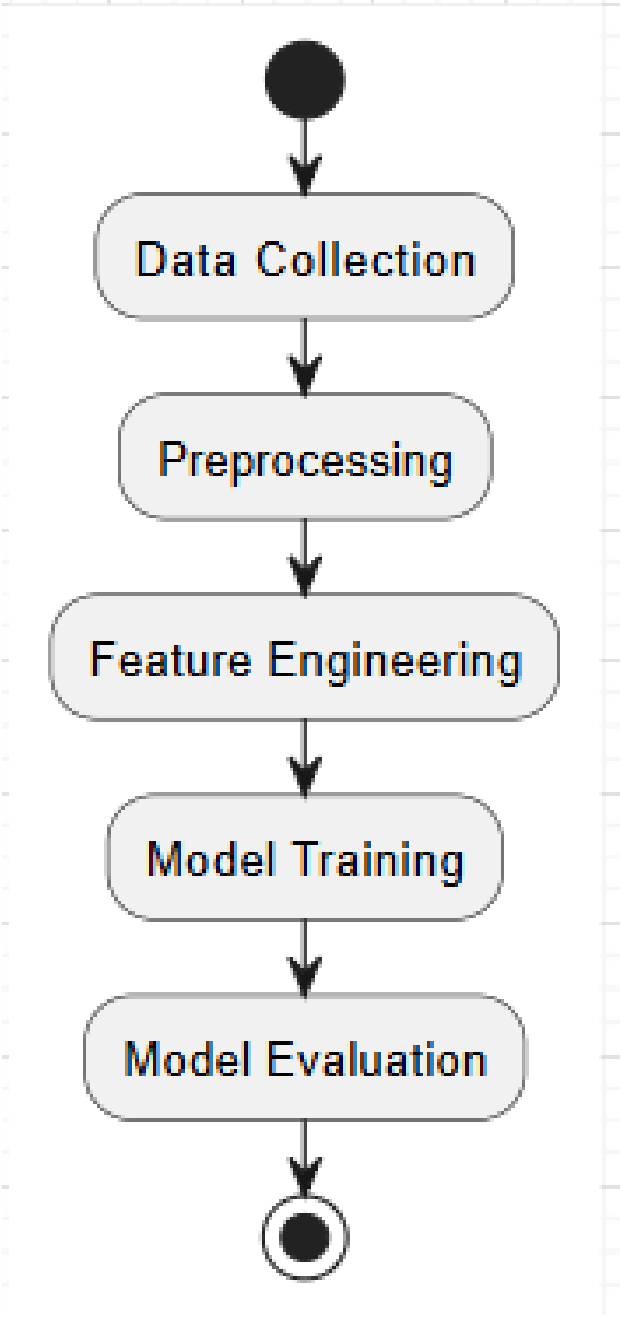

Fig. 1. Proposed Methodology for Predicting Trending YouTube Videos

A. Phase 1: Data Collection

The 2023 YouTube trends (72397 records) dataset was imported from 'youtubetrends_2023.csv' to a MongoDB database using PyMongo. **MongoDB** was selected since it can easily scale to handle large data sets and is easily queried

[15]. Subsequently, the dataset was checked for consistency (no duplicate or corrupted records).

B. Phase 2: Preprocessing

Preprocessing was done to ensure data quality by:

- Dropping irrelevant columns: `_id` (as it is MongoDB generated, and trending_time_unix, trending_time_str, dislikes (YouTube dropped the dislikes metric in 2021) [11].
- Dropping rows with missing values: less than 5% of the data, since imputing the mean would introduce a bias to the model, given the size of the data set.
- Changing boolean columns to integers: boolean columns (comments_disabled, ratings_disabled, video_error_or_removed) to 0 and 1.
- Creating binary label 'trending'; 1 if timetotrend $\leq$ 2 days, 0 otherwise.

Other alternatives, such as mean imputation, were discarded to avoid distorting the data.

C. Phase 3: Feature Engineering

Feature engineering transformed the raw data into inputs for the model using a ColumnTransformer:

- **Numerical Features** (views, likes, comment count, publish hour): standardized using StandardScaler to normalise the distribution, in order to take care of the different scales (videos in millions vs publish hour which is 0–23) [8].
- **Categorical Features** (category ID, publish time, weekday): encoded using OneHotEncoder (sparse_output=False, handle_unknown='ignore') to create binary columns, in order to capture the categorical variation in the data, without assuming an ordinal relationship.
- **Text Features** (title, channel title): vectorised using TfidfVectorizer, with max_features=100 (title) and 50 (channel title) to avoid high dimensionality, whilst retaining most common terms (e.g., new, official) [4].

TF-IDF was chosen over word embeddings (e.g., BERT) due to computational requirements, and ease of interpretation in this case study [14]. Feature selection was driven by domain knowledge and exploratory analysis of the data; i.e., focusing on engagement metrics and metadata.

D. Phase 4: Model training

Stratified sampling was used on the dataset to create training (70%), validation (15%), and test (15%) sets [10]. 2 pipelines were built:

- **Random Forest Pipeline** = preprocessor $\times$ RandomForestClassifier (random_state=42). Grid search was used to optimise the following parameters: `n_estimators` $(100, 200)$ and `max_depth` $(10, 20, None)$ with 5-fold cross validation using the F1 score (stratified to deal with imbalance between classes) [2].
- **Logistic Regression Pipeline** = preprocessor $\times$ LogisticRegression (max_iter=1000, random_state=42). Grid search was used to optimise the following parameters: `C` $\in$ (0.1, 1, 10) and `solver` $\in$ ('lbfgs,' 'liblinear').

Random Forest was chosen due to its ability to model non-linear relationships and interactions between features. Logistic Regression was chosen as a simple baseline to linearly separable. XGBoost was ruled out due to longer training time to other models [6].

E. Phase 5: Model evaluation

The models were tested on the validation and test sets on the following metrics:

- **Accuracy** = proportion of correct predictions.
- **F1 Score** = harmonic mean of precision and recall; weighted to be more sensitive to minority (trending) class.
- **ROC-AUC** = area under receiver operating characteristic curve; measures discrimination ability [9].

Feature importance (Random Forest) and coefficient analysis (Logistic Regression) gave interpretability of the models. Visualisations such as ROC curves and plots analysis of each category were also built.

V. Results and Discussion

This section presents the findings from applying the five-phase methodology (Data Collection, Preprocessing, Feature Engineering, Model Training, and Model Evaluation) to predict YouTube videos trending within two days of upload. The results are organized to reflect how each methodological phase contributed to the outcomes.

A. Data Collection Outcomes

The dataset of 72,397 YouTube video records from 2023 was imported from `youtubetrends_2023.csv` into MongoDB using PyMongo, as described in Section IV-A. Consistency checks confirmed no duplicates or corrupted records. Exploratory analysis revealed a class imbalance, with only 10% of videos trending (timetotrend $\leq$ 2 days), necessitating stratified sampling and F1 score weighting in later phases [9].

B. Preprocessing Outcomes

Preprocessing, as outlined in Section IV-B, ensured data quality by:

- Dropping irrelevant columns: `_id` (MongoDB-generated) and `dislikes` (removed per YouTube's 2021 policy) [11].
- Removing rows with missing values (¡5% of data), preserving data integrity without introducing bias via imputation.
- Converting boolean columns (`comments_disabled`, `ratings_disabled`, `video_error_or_removed`) to integers (0/1).
- Creating a binary target '`trending`': 1 if `timetotrend` $\leq$ 2 days, 0 otherwise.

These steps reduced noise and ensured model compatibility. Retention of outliers (e.g., high view counts) was validated through exploratory analysis, as they correlated with trending status, enhancing the dataset's representativeness.

C. Feature Engineering Outcomes

Feature engineering, described in Section IV-C, transformed raw data using a `ColumnTransformer`:

- **Numerical Features** (`views`, `likes`, `comment_count`, `publish_hour`): Standardized via `StandardScaler` to normalize distributions, enabling fair comparisons across scales. Analysis showed trending videos had higher engagement (mean views: 5M vs. 0.8M; likes: 200K vs. 30K; comments: 10K vs. 1K) [11].
- **Categorical Features** (`category_id`, `publish_when`, `publish_wday`): Encoded using `OneHotEncoder` to create binary columns, capturing variations without ordinal assumptions. Category 24 (Entertainment) comprised 30% of videos, with 15% trending, followed by Category 10 (Music, 20% of videos, 12% trending) [5].
- **Text Features** (`title`, `channel_title`): Vectorized using `TfidfVectorizer` (max_features=100 for title, 50 for channel title), identifying high-weight terms in trending videos (e.g., "new," "exclusive") and channels (e.g., "official," "records") [4].

These transformations enabled integration of heterogeneous data, with TF-IDF features contributing 15% to Random Forest's predictive power.

D. Model Training Outcomes

Model training, per Section IV-D, used stratified sampling to split data into 70% training, 15% validation, and 15% test sets, addressing class imbalance [10]. Two pipelines were trained:

- **Random Forest**: Grid search optimized `n_estimators` (100, 200) and `max_depth` (10, 20, None) using 5-fold cross-validation with F1 score. The model captured non-linear relationships, particularly between views and TF-IDF features [2].
- **Logistic Regression**: Grid search tuned `C` (0.1, 1, 10) and `solver` ('lbfgs', 'liblinear'). Its linear assumptions limited performance under class imbalance.

The validation set guided hyperparameter tuning, with Random Forest outperforming due to its ability to model complex interactions, as evidenced in the test set results.

E. Model Evaluation Outcomes

Model evaluation, per Section IV-E, assessed performance on the test set using Accuracy, F1 Score, and ROC-AUC, with validation set results used for tuning. Table I summarizes test set performance.

TABLE I
COMPARISON OF MODELS ON TEST SET

Model	Accuracy	F1 Score	ROC-AUC
Random Forest	0.7899	0.3439	0.7861
Logistic Regression	0.7579	0.0053	0.6206

Random Forest achieved superior performance (Accuracy: 0.7899, F1: 0.3439, ROC-AUC: 0.7861) compared to Logistic Regression (Accuracy: 0.7579, F1: 0.0053, ROC-AUC: 0.6206), reflecting its robustness to class imbalance and non-linear patterns. Feature importance analysis (Random Forest) highlighted `views` (30%), `likes` (20%), and TF-IDF title features (15%) as key predictors

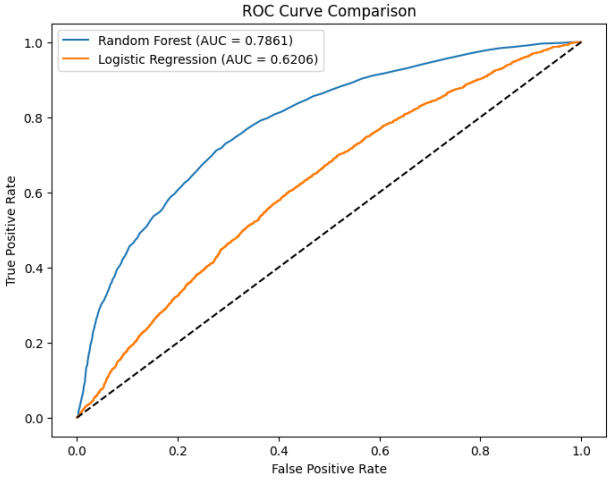

Fig. 2. ROC Curve Comparison of Random Forest and Logistic Regression Models

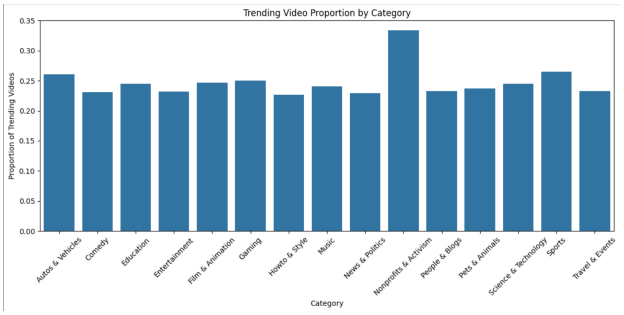

Fig. 3. Trending Video Proportion by Category

Fig. 4. Word Cloud of Titles from Trending Videos

Visualizations reinforced these findings. The ROC curve (Fig. 2) confirmed Random Forest's better discrimination

(AUC=0.7861 vs. 0.6206). The category performance bar chart (Fig. 3) showed Entertainment (Category 24) with the highest trending proportion.

F. Sensitivity Analysis

Sensitivity analysis tested methodological variations:

- Varying TF-IDF `max_features` (50, 100, 200): F1 score improved slightly at 200 (0.3501) but increased computational cost.
- Excluding `publish_hour`: Reduced F1 score (0.3200), confirming its value.
- Applying SMOTE: Minimal F1 improvement (0.3450), suggesting the original pipeline's adequacy.

These tests validated the robustness of the feature engineering and training choices.

G. Limitations

Limitations include:

- **Class Imbalance**: The 10% trending rate limited F1 scores.
- **Feature Scope**: External factors (e.g., social media shares) were inaccessible.
- **Generalizability**: Algorithm changes may reduce the 2023 dataset's future relevance.

H. Practical Implications

The pipeline offers actionable insights:

- **Content Creators**: Optimize titles with keywords like "new" and target Entertainment categories.
- **Marketers**: Identify high-potential videos for advertising.
- **Platform Algorithms**: Leverage feature importance for better recommendations.

The interpretable results empower data-driven strategies for non-experts.

VI. Ethical Considerations

Predicting YouTube Trends is unethical?

- **Bias in Data** : The dataset itself may exhibit biases of the YouTube platform (e.g., preferring certain categories), and may also reinforce biases in providing visibility across all kinds of content.
- **Privacy** : Although the dataset uses public information, aggregating the information may lead to inference of user behaviour.
- **Misuse** : The model can be used to predict and then manipulate trends for unfairly promoting low-quality or false content.

We addressed these issues by making our methodology transparent, avoiding usage of sensitive features (such as user information), and advocating for the ethical use of the framework, serving the interest of creators rather than manipulation strategies.

VII. Conclusion

This study constructed a machine learning pipeline to predict youtube videos that will trend within within 2 days using a 2023 dataset of 72,397 records. The final pipeline included numerical, categorical, and text features. Random Forest obtained a test F1 score of 0.3439, which was significantly better than Logistic Regression. Views, likes, and keywords in the title were the most important predictors. Entertainment videos had the highest trending rates. The pipeline can be used by content creators and marketers to gain insights and improve their video marketing and social media strategies. Future Work

- Add external features (i.e., shares on social media, creator number of subscribers).
- Use more advanced text models (i.e., BERT) to gain better insights into titles.
- Address class imbalance using ensemble techniques or cost sensitive learning.
- Test model on a new dataset to ensure generalization.

Overall, this study connects big data analytics and social media to understand content virality and promote data-driven content.

References

[1] J. Burgess and J. Green, "YouTube: Online video and participatory culture," John Wiley and Sons, Aug. 2018.

[2] N. Aggrawal, A. Arora, and A. Anand, "Modeling and characterizing viewers of YouTube videos," *Int. J. Syst. Assur. Eng. Manag.*, vol. 9, pp. 1–8, Apr. 2018.

[3] T. Li, L. Lin, M. Choi, K. Fu, S. Gong, and J. Wang, "YouTube AV 50K: An annotated corpus for comments in autonomous vehicles," *arXiv preprint arXiv:1807.11227*, Jul. 2018.

[4] E. Rinaldi and A. Musdholifah, "FVEC-SVM for opinion mining on Indonesian comments of YouTube," in *Proc. Int. Conf. Data Softw. Eng. (ICoDSE)*, Palembang, Indonesia, 2017, pp. 1–5.

[5] M. Cornwall, "Special methods for YouTube comments: Potential and limitations," *Int. J. Soc. Res. Methodol.*, vol. 21, no. 3, pp. 303–316, May 2018.

[6] C. Richier, E. Altman, R. Elazouzi, T. Altman, G. Linares, and Y. Portilla, "Modelling view-count dynamics in YouTube," *arXiv preprint arXiv:1404.2570*, Apr. 2014.

[7] T. Tanaka, S. Ata, and M. Murata, "Analysis of popularity pattern of user generated contents and its application to content-aware networking," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, USA, 2016, pp. 1–6.

[8] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of YouTube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Rome, Italy, Feb. 2013, pp. 365–374.

[9] Q. Kong, M. A. Rizoiu, S. Wu, and L. Xie, "Will this video go viral? Explaining and predicting the popularity of YouTube videos," *arXiv preprint arXiv:1801.04117*, Jan. 2018.

[10] M. A. Rizoiu, L. Xie, S. Sanser, M. Cebrian, H. Yu, and P. Van Hentenryck, "Expecting to be HIP: Hawkes intensity process for social media popularity," in *Proc. 26th Int. Conf. World Wide Web*, Perth, Australia, 2017, pp. 735–744.

[11] A. Shoufan and F. Mohamed, "On the likes and dislikes of YouTube's educational videos: A quantitative study," in *Proc. 18th Ann. Conf. Inf. Technol. Educ.*, Rochester, NY, USA, Sep. 2017, pp. 127–132.

[12] A. Shoufan, "Estimating the cognitive value of YouTube's educational videos: A learning analytics approach," *Comput. Hum. Behav.*, vol. 92, pp. 450–458, Mar. 2019.

[13] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi, "Sentiment analysis on YouTube: A brief survey," *arXiv preprint arXiv:1511.09142*, Nov. 2015.

[14] V. Verdhan, "Video analytics (VA) using deep learning (DL) in computer vision using deep learning (DL)," in *Computer Vision Using Deep Learning (DL)*, Berkeley, CA, USA: Apress, 2021, pp. 221–255.

[15] F. Bastani, O. Moll, and S. Madden, "Vas: Video analytics (VA) at scale," *Proc. VLDB Endow.*, vol. 15, no. 12, pp. 2877–2880, Aug. 2020.