BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

Modeling YouTube Channel Success Using Multidimensional Metrics and Predictive Analytics: A Preliminary Study

Wawang Adi Darma[1*], Iyus Rusmansyah[2], Deny Ariestiandy[3]

Department of Sytem Information[1,2,3]
Institut Citra Buana Indonesia[1,2,3]
adidarma2k15@gmail.com[1*], sastrapangeran@yahoo.co.id[2], deny@cbi.ac.id[3]

Abstract. YouTube channel success is commonly evaluated in a partial manner using single metrics such as view counts or subscriber numbers, even though success is inherently multidimensional. This preliminary study (i) constructs a multidimensional composite index that integrates four dimensions, namely popularity, engagement, consistency, and content quality, and (ii) examines, with a leakage-free design, how much of a channel's category-relative audience standing can be predicted from behavioural metrics alone. Data were collected through the YouTube Data API for 105 Indonesian-focused channels in five content categories; 95 channels with at least five collected videos (7,852 videos) were analysed. The index is a weighted additive aggregation of min-max normalised dimension scores, and its sensitivity to the weights was examined. The target is whether a channel's subscriber count is at or above the median of its category, and all subscriber-derived variables were excluded from the predictors. Four classifiers were evaluated with 5-fold cross-validation repeated 20 times. With behavioural metrics only, the best model reached a cross-validated AUC of 0.70 (pooled bootstrap 95% CI 0.60-0.81); adding median views per video gave 0.75 (95% CI 0.64-0.84), whereas a deliberately leaky reference including subscriber-derived features reached 0.91 (best model). Label-permutation tests rejected chance performance ($p = 0.010$). Because of the small sample, the findings are exploratory and are intended as a methodological basis for a larger follow-up study.

Keywords: composite index; ensemble learning; multi-criteria decision analysis; predictive analytics; YouTube channel

1. INTRODUCTION

YouTube has grown into a dominant digital content platform with more than two billion monthly active users, creating a highly competitive ecosystem for content creators [1]. In this environment, measuring and predicting the success level of a channel has become critical for creators, marketers, and platform managers alike. Conventionally, channel success is frequently assessed partially through single metrics such as subscriber counts or cumulative views [2], [3]. This approach, however, risks producing an incomplete understanding and unsustainable strategic recommendations, because it disregards other aspects such as audience engagement, upload consistency, and the quality of the content produced [4], [5].

Although prior work has made valuable contributions to the analysis of YouTube content performance, most studies remain limited to the analysis of individual videos or to a single evaluation dimension, and have not systematically integrated multiple dimensions of success into a unified framework that treats the channel as a complete entity. This gap motivates the present study.

Based on this observation, three research questions are formulated: (1) how can a holistic multidimensional composite index be formulated to represent YouTube channel success; (2) how sensitive is that index to the choice of dimension weights, and how do the dimensions relate to each other and to audience size; and (3) how much of a channel's category-relative audience standing can be predicted from behavioural metrics once all variables that define that standing are excluded. Accordingly, this study aims to design a multidimensional composite index integrating popularity, engagement, consistency, and content quality metrics, to assess its sensitivity, and to estimate, with a leakage-free validation design, the predictability of audience standing in a small sample.

This paper reports a preliminary study within a larger planned research programme, and its contribution is correspondingly modest and methodological. It shifts the unit of analysis from the individual video to the channel as a whole; it specifies a transparent four-dimension index inspired by multi-criteria decision analysis [6]; and it documents a leakage-free evaluation design for predicting

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

channel standing, quantifying how much leakage inflates apparent performance. With 95 channels, the results are exploratory estimates with wide uncertainty and not validated predictive claims.

2. RELATED WORK

Research on YouTube content performance has evolved along several methodological strands. A first strand focuses on predicting the popularity of individual videos using machine learning algorithms such as ensemble methods based on metadata and thumbnail features [1], [2]. A second strand performs longitudinal comparative analysis of video performance within a single channel [4]. A third strand advances sentiment analysis and topic detection from YouTube comments, achieving high accuracy using hybrid frameworks and big data processing [8], [9]. While these studies provide robust techniques, their unit of analysis remains the individual video rather than the channel.

Ensemble learning is a common choice for tabular prediction. Darsiti [12] compared LightGBM, XGBoost, and Random Forest for a regression task on audio features and reported that boosting based models outperformed bagging based alternatives, which informs the candidate algorithms in the present study; in very small samples, however, differences among algorithms are often smaller than the sampling error, as shown in Section 4.4.

Within the Indonesian social media context, Noerhadi et al. [11] applied sentiment analysis to Indonesian-language Twitter data and confirmed the robustness of machine learning classifiers in handling high-dimensional text from a low-resource language, which motivates the future use of comment sentiment as a content quality measure; such data were not collected in the present study. From a methodological standpoint, Widiyanto [13] emphasised the importance of strict separation between training and unseen test data when evaluating models on imbalanced datasets, in line with the general literature on data leakage [15], [16] and on cross-validation with small samples [17]; these principles underlie the evaluation design in Section 3.

The concept of multidimensional evaluation and composite index construction is already mature in other fields, such as the evaluation of synthetic data generators [7] and decision support systems employing methods such as the Characteristic Objects Method (COMET) [6]. Work on creator ranking has relied mainly on network position, for example in influencer recommendation [10], rather than on a multidimensional assessment of channel performance. Nevertheless, the adaptation of this systematic framework to assess the success of social media channels remains very limited, creating a gap between the real-world complexity of the success phenomenon and the simplified metrics currently in use.

3. RESEARCH METHOD

This study employs a quantitative computational approach with an integrated method spanning data processing, multidimensional index construction, factor analysis, and predictive modeling. The overall research flow is presented in Fig. 1.

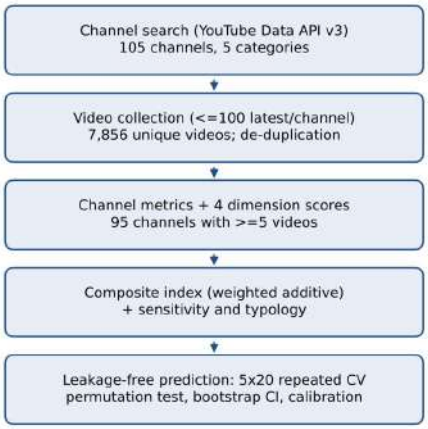


Fig. 1. Research flow diagram.

3.1. Research Object and Data Collection

The research objects are Indonesian-focused YouTube channels drawn from five content categories (Education, Entertainment, Lifestyle, Technology, and Music). Channels were identified by purposive keyword search with the YouTube Data API v3 (five search queries per category, up to 20 results per query, region code ID, Indonesian relevance language; language was not verified manually) and retained if they had at least 10,000 subscribers and at least 30 published videos. This yielded 105 unique channels (Education 12, Entertainment 16, Lifestyle 33, Technology 22, Music 22). Channel statistics and the 100 most recent videos of each channel (view, like, and comment counts, duration, and publication date) were collected on 19 January 2026. The collection files contained duplicated records (15,178 rows for 7,856 unique videos), which were removed by channel and video identifier. Ten channels were excluded (seven without video data and three with fewer than five videos), leaving 95 channels and 7,852 videos for analysis. Because the 100 most recent videos were collected, the observation window depends on upload frequency (median 10.2 months, interquartile range 4.1-18.6; 34 channels have windows shorter than six months), which limits the consistency dimension. Sampling is purposive, so the sample does not represent the population of Indonesian channels.

3.2. Metric Extraction and Dimension Scores

1) Popularity: measured through the median number of views per video (log-transformed) [1], [2]. Subscriber count was collected but is used neither in the index nor as a

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

predictor, because it defines the prediction target (Section 3.4).

2) Engagement: measured through the like rate (likes/views), the comment rate (comments/views), and the mean per-video engagement rate ((likes + comments)/views), combined as E = 0.4·LR + 0.4·CR + 0.2·ER. Public dislike and share counts are not available in the collected data and are not used.

3) Consistency: measured through upload frequency f (videos per month over the observation window, capped at 8) and the coefficient of variation (CV, %) of views across videos: C = 0.6·min(f/8, 1) + 0.4·max(0, 1 − CV/200).

4) Content Quality (proxy): comment sentiment and viewer retention were not collected. Quality is therefore approximated by audience appreciation, Q = (0.6·LR + 0.4·CR)/2. This proxy is almost collinear with the engagement score (Spearman ρ = 0.98), so the quality dimension adds little independent information (Section 4.3). All metric definitions are listed in Appendix A (Table A1).

3.3. Construction of the Multidimensional Composite Index

The composite index is a weighted additive multi-criteria aggregation (weighted sum model), complemented by a profile typology inspired by the Characteristic Objects Method (COMET) [6]. The procedure has four steps and uses only the four dimension scores, never subscriber information.

Step 1 (normalisation). Each raw dimension value r_{id} of channel i on dimension d ∈ {P, E, C, Q} (popularity, engagement, consistency, quality) is min-max normalised over the analysed channels: $x_{id} = (r_{id} - \min_d) / (\max_d - \min_d)$, so that $x_{id} \in [0, 1]$, with $r_{iP} = \log(1 + \text{median views per video})$. (1)

Step 2 (weights). The weight vector $w = (w_P, w_E, w_C, w_Q) = (0.25, 0.30, 0.25, 0.20)$, $\Sigma_d w_d = 1$, was assigned by the authors a priori, giving the largest weight to engagement as direct audience interaction and the smallest to the quality proxy. The weights were neither elicited from external experts nor estimated from the data; their influence is examined in Step 4.

Step 3 (index). The composite index of channel i is $I_i = 100 \cdot \Sigma_d w_d x_{id}$, which lies on a 0-100 scale. (2)

Step 4 (sensitivity and typology). Sensitivity was assessed by (a) equal weights, (b) dropping the quality dimension and renormalising the remaining weights, and (c) 1,000 random weight vectors drawn from a Dirichlet distribution centred on w (concentration 20), reporting Spearman rank correlations with the baseline index and the overlap of the ten highest-ranked channels. For descriptive typing, four characteristic objects were defined on the normalised dimensions, viral (1.0, 0.7, 0.6, 0.5), community (0.6, 1.0, 0.8, 0.9), consistent (0.7, 0.6, 1.0, 0.7), and balanced (0.8, 0.8, 0.8, 0.8), and each channel was assigned to the object o_k with the highest cosine similarity between $w \circ x_i$ and $w \circ o_k$. Cosine similarity depends on the shape of a profile rather than its level, so it is used only for typology and not as a score (Section 4.2).

3.4. Predictive Model Development

Target definition. The binary target equals 1 if a channel's subscriber count is at or above the median subscriber count of its content category (computed over the 95 analysed channels) and 0 otherwise (48 positive and 47 negative channels). It is a category-relative audience-standing criterion and is distinct from the composite index. The threshold is part of the outcome definition rather than a learned parameter and was therefore fixed once; all data-dependent steps of the predictive pipeline (scaling and model fitting) are fitted inside the training folds. Audience size is only one facet of success (Section 4.6).

Leakage audit and feature sets. Because the target is a deterministic function of subscriber count, any predictor that contains or can be reconstructed from subscriber count leaks the target [15], [16]. A first version of this analysis used log-subscribers, log-videos, and log-subscribers-per-video as its only predictors and was discarded for that reason. In the present design, subscriber count, its transformations, and ratios involving it are excluded from all predictors. Four feature sets were compared. Set A (behaviour only): like rate, comment rate, mean engagement rate, log upload frequency, view CV, log mean video duration, and share of videos of 60 s or less. Set B: set A plus log median views per video, an audience-reach measure that is related to subscriber count but not derived from it. Set C: the four dimension scores. Set X, a deliberately leaky reference: set B plus log-subscribers and log-subscribers-per-video, reported only to quantify the effect of leakage. The audit also included single-feature AUC screening and label-permutation tests (Section 4.4).

Validation. With 95 channels, a 20% holdout set would contain about 19 channels, and AUC estimates from such a set are highly unstable (Section 4.4). Following guidance on small-sample validation [17], performance was therefore estimated with stratified 5-fold cross-validation repeated 20 times (100 train-test splits), stratified by category and target, with scaling and model fitting inside each fold. Hyperparameters were fixed a priori and not tuned, so nested cross-validation, which addresses selection bias from tuning [14], was not required. Confidence intervals come from a bootstrap (2,000 resamples of channels) of the repeat-averaged out-of-fold probabilities, and chance-level performance was tested with 100 label permutations [18].

Modeling used ensemble algorithms: L2-regularised logistic regression (C = 1) as a linear baseline, Random Forest (200 trees, minimum leaf size 3) [21], XGBoost (100 trees, depth 2, learning rate 0.05) [22], and LightGBM (100 trees, 4 leaves, learning rate 0.05, minimum 5 samples per leaf) [23], implemented with scikit-learn [24]. Discrimination was measured by accuracy, precision, recall, F1-score (threshold 0.5), and AUC-ROC. Calibration was

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

measured by the Expected Calibration Error, ECE = $\Sigma_{b=1}^{5}$ (n_b/n) $|acc_b - conf_b|$ [19], computed with five equal-frequency bins of about 19 channels each (equal-width bins would be nearly empty at n = 95), where acc_b is the observed share of positives and $conf_b$ the mean predicted probability in bin b, and by the Brier score [20]. ECE lies between 0 and 1, and lower is better.

Reproducibility. The analysis script (Python, random seed 42) and the channel-level metrics of the 95 analysed channels are provided as supplementary material.

4. RESULTS AND DISCUSSION

4.1. Dataset Characteristics and Success Dimension Profiles

The analysed dataset comprises 95 channels (Education 10, Entertainment 14, Lifestyle 33, Technology 20, and Music 18; 105 were identified) with 7,852 videos (median 100 per channel; 67 channels reached the 100-video cap). Table 1 shows descriptive statistics as medians because the distributions are strongly right-skewed (K = thousands; ER = mean engagement rate; Index = composite index). Median subscriber counts range from 94.7 thousand (Education) to 2.47 million (Entertainment), and the four dimension scores and the index differ across categories (Kruskal-Wallis $p < 0.01$ for each). Category sizes are small (10 to 33 channels), so category differences are descriptive and should not be generalised.

TABLE 1. Descriptive statistics by content category (medians; n = analysed channels)

Category (n)	Subs. (K)	Views/ video (K)	ER (%)	Index
Education (n = 10)	94.7	1.0	2.31	26.2
Entertainment (n = 14)	2,470.0	247.7	2.16	38.8
Lifestyle (n = 33)	98.2	21.2	3.27	29.6
Technology (n = 20)	378.0	19.4	3.10	31.2
Music (n = 18)	129.4	2.6	1.55	23.0
All (n = 95)	211.0	16.5	2.66	29.2

4.2. Composite Index and Sensitivity

The composite index (Eq. 2) ranged from 12.0 to 62.3, with a mean of 31.2 (SD 10.3) on the 0-100 scale; its distribution departs from normality (Shapiro-Wilk p = 0.005). Mean dimension scores were 47.6 for popularity, 10.7 for engagement, 55.9 for consistency, and 10.4 for the quality proxy (Fig. 2). The index was insensitive to the exact weights: its Spearman correlation with an equal-weight index was 1.000, and 0.97 when the quality dimension was dropped. Across 1,000 random weight vectors the mean correlation was 0.95 (minimum 0.72), and on average 8.2 of the ten highest-ranked channels remained in the top ten (minimum 4). The index correlates with subscriber count at $\rho = 0.55$ (popularity 0.64, consistency 0.40, engagement 0.06, quality 0.11), so it is related to, but not redundant with, audience size. Typology assigned 65 channels to the consistent profile, 26 to viral, and 4 to community. A similarity-based score built on the same cosine measure was only weakly related to the additive index ($\rho = 0.31$) and placed 5 of its ten highest-ranked channels in the lower half of the additive index, which confirms that it describes profile shape and is unsuitable as a score.

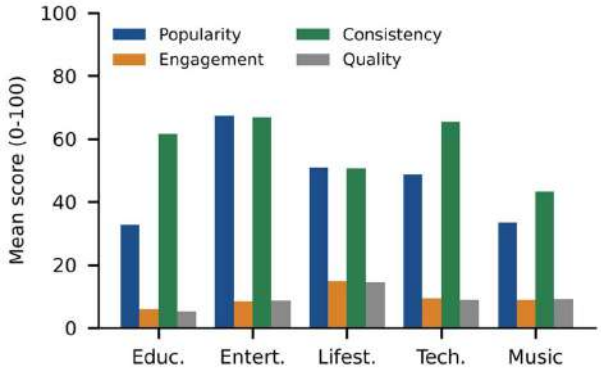


Fig. 2. Mean score of the four success dimensions by content category (0-100 scale; Educ. = Education, Entert. = Entertainment, Lifest. = Lifestyle, Tech. = Technology).

4.3. Relationships among Dimensions

Spearman correlations among the four dimension scores were weak (absolute values up to 0.17) with one exception: engagement and quality were almost perfectly correlated ($\rho = 0.98$), as expected because the quality proxy is built from the same like and comment ratios. In this dataset the quality dimension therefore does not carry independent information, and the claim of four independent dimensions is not supported; dropping it changes the ranking of channels only slightly ($\rho = 0.97$). Replacing the proxy by sentiment- or retention-based measures is a priority for the larger study.

4.4. Predictive Modeling Results and Model Evaluation

Table 2 reports cross-validated performance for feature sets A and B (5-fold cross-validation repeated 20 times; 100 train-test splits). With behavioural metrics only (set A), the best model, Random Forest, reached a mean fold AUC of 0.70 (SD 0.12 across folds), an accuracy of 64.9%, and an F1-score of 0.66. Adding median views per video (set B) raised the best AUC to 0.75 (SD 0.11), the accuracy to 69.7%, and the F1-score to 0.70. The four classifiers differed by less than 0.04 AUC, much less than the fold-level standard deviation, so no algorithm can be declared superior, and the linear model was competitive, which is typical for small samples. A model using the category alone gave an AUC of 0.44, i.e. chance level, as expected for a category-relative target.

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

TABLE 2. Cross-validated performance (5 x 20 repeated stratified cv; means, AUC as mean ± sd across folds)

Model (Set)	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Logistic reg. (A)	62.1	0.62	0.70	0.65	0.69 ± 0.12
Random Forest (A)	64.9	0.65	0.69	0.66	0.70 ± 0.12
XGBoost (A)	63.4	0.63	0.66	0.64	0.69 ± 0.12
LightGBM (A)	62.6	0.63	0.65	0.63	0.70 ± 0.11
Logistic reg. (B)	64.9	0.66	0.68	0.66	0.73 ± 0.11
Random Forest (B)	69.7	0.71	0.71	0.70	0.75 ± 0.11
XGBoost (B)	68.1	0.69	0.70	0.69	0.71 ± 0.11
LightGBM (B)	66.8	0.68	0.68	0.67	0.73 ± 0.11

Table 3 summarises the best model of each feature set with pooled out-of-fold AUCs and bootstrap 95% confidence intervals: 0.71 (0.60-0.81) for set A, 0.75 (0.64-0.84) for set B, 0.76 (0.65-0.85) for the dimension scores (set C), and 0.91 (0.84-0.97) for the leaky reference (set X). Label permutation gave a null AUC of about 0.5 and p = 0.010 for sets A, B, and C (the smallest attainable p with 100 permutations is 0.0099), so the behavioural models perform better than chance. The intervals are wide (about ±0.1), reflecting n = 95. No separate holdout set was used because it would be uninformative: 200 random 80/20 splits (n_test = 19) of the set B Random Forest gave test AUCs with mean 0.76, SD 0.09, and a 95% range of 0.57-0.92. The ROC curves are shown in Fig. 3.

TABLE 3. Best model per feature set: pooled out-of-fold AUC, bootstrap 95% CI, and label-permutation test

Set	Model	CV AUC	Pooled AUC [95% CI]	Null AUC	p
A	Random Forest	0.70 ± 0.12	0.71 [0.60, 0.81]	0.49	0.010
B	Random Forest	0.75 ± 0.11	0.75 [0.64, 0.84]	0.50	0.010
C	XGBoost	0.75 ± 0.09	0.76 [0.65, 0.85]	0.49	0.010
X (leaky)	Logistic reg.	0.91 ± 0.06	0.91 [0.84, 0.97]	0.51	0.010

Effect of leakage. Including subscriber-derived features raised the AUC of the Random Forest from 0.75 (set B) to 0.89 (set X), and log-subscribers alone achieved a single-feature AUC of 0.91, against 0.63 for the strongest purely behavioural feature (comment rate, which is higher in channels below the category median). A model that includes subscriber-derived variables therefore largely recovers the labeling rule, and the AUC of about 0.9 in such a design, as in the first version of this analysis, reflects leakage and not predictive skill [15], [16].

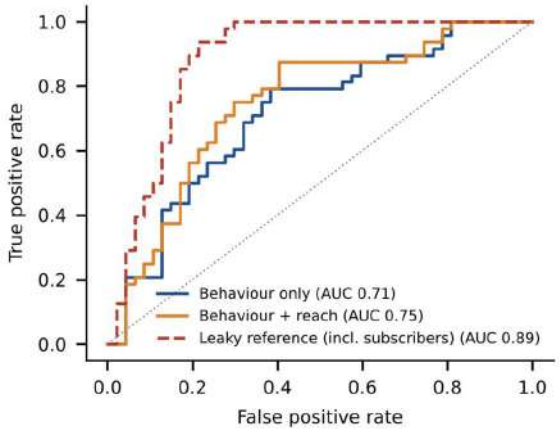


Fig. 3. Out-of-fold ROC curves (repeat-averaged probabilities, Random Forest) for feature sets A and B and for the leaky reference (X).

Calibration was assessed on the repeat-averaged out-of-fold probabilities of the best model of sets A and B (Fig. 4). The ECE was 0.080 for set A and 0.072 for set B, with bootstrap 95% intervals of 0.048-0.177 and 0.054-0.190, and the Brier scores were 0.215 and 0.205 against 0.250 for a constant prediction of the class share. Because ECE is the weighted mean absolute gap between predicted probability and observed frequency, these values correspond to an average gap of roughly 7-8 percentage points. With only 19 channels per bin the calibration curve is noisy, and the interval shows that considerably worse calibration cannot be excluded. We therefore make no claim that the predicted probabilities are well calibrated or free of over-confidence; they should be used as rank scores until a larger sample allows post-hoc calibration on separate data.

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

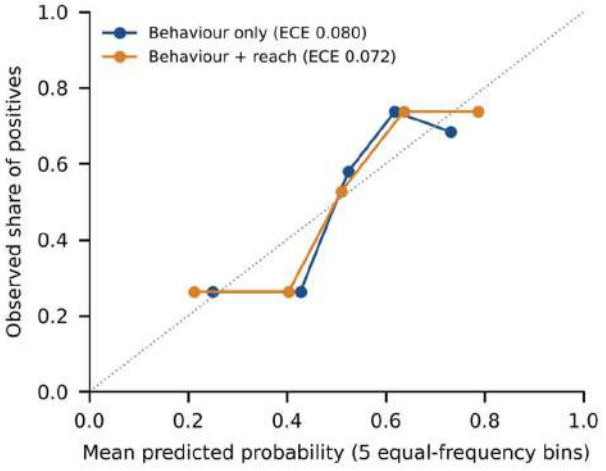


Fig. 4. Calibration curve (five equal-frequency bins, repeat-averaged out-of-fold probabilities; ECE in legend).

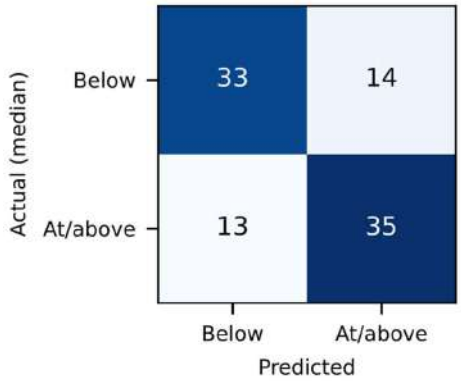


Fig. 5. Confusion matrix of the best model of set B (Random Forest) at threshold 0.5, from repeat-averaged out-of-fold probabilities (n = 95).

Feature importance was measured as the drop in held-out AUC when the values of one feature are randomly permuted in the test fold (10 permutations in each of 3 x 5 folds) for the set B Random Forest. The results are presented in Fig. 6.

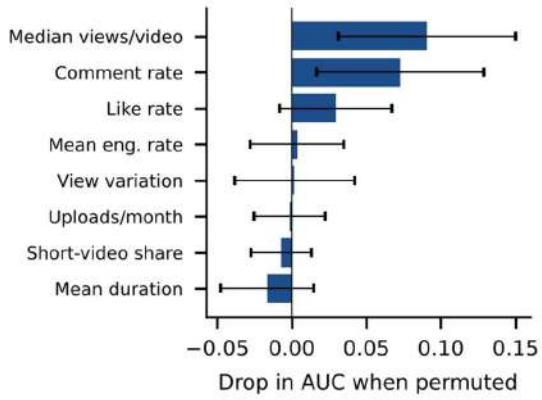


Fig. 6. Permutation importance (drop in held-out AUC, mean ± SD across folds) for the set B Random Forest.

As shown in Fig. 6, median views per video had the largest importance (0.091), followed by comment rate (0.073) and like rate (0.030); the remaining features were close to zero, and only the first two exceed one standard deviation across folds. With behavioural metrics only (set A), comment rate (0.118), like rate (0.060), and view variation (0.050) were the leading features. The encoded category was not used as a feature because the target is already category-relative.

4.5. Direction of Associations

Channels at or above their category's median subscriber count had higher like rates (median 2.06% vs 1.36%) but lower comment rates (0.08% vs 0.18%), lower view variation (CV 131% vs 179%), and fewer uploads per month (8.0 vs 10.6). The comment-rate pattern is plausible if larger audiences comment less per view, but these are associations in a small purposive sample and not causal effects, and the opposite signs of the like-rate and comment-rate coefficients in the logistic model should not be over-interpreted at n = 95. SHAP-based interpretation was not used in this preliminary study.

4.6. Discussion and Limitations

This preliminary study supports three conclusions. First, once leakage is removed, the performance of the models (cross-validated AUC 0.70-0.75) is much lower than the roughly 0.9 obtained when subscriber-derived features are included, and the difference quantifies the leakage in a design that predicts a subscriber-based label from subscriber-based features [15], [16]. Second, behavioural metrics, mainly comment and like rates, and audience reach carry detectable but modest information about category-relative audience standing (permutation p = 0.010), in line with video-level studies in which engagement signals relate to popularity [1], [2]. Third, the composite index is stable under plausible changes of the weights, but in this dataset it reflects popularity, consistency, and an engagement-quality pair that is almost collinear. Because the weights were assigned by the authors and the quality dimension is a proxy, the index should be regarded as a transparent descriptive construct and not as a validated measure of success.

The sample is small (95 channels, 10 to 33 per category) and purposive, so the results do not generalise to the population of channels. Only the 100 most recent videos were collected, and the observation window is shorter than six months for 34 channels. The target is an audience-size proxy and the data are a cross-sectional snapshot, so the models describe association and do not forecast success. The weights are not empirically validated, the quality dimension is a proxy that duplicates engagement, and sentiment and retention data are missing. No external dataset was used, the probabilities are not demonstrably calibrated, and the category cells are too small for reliable category-specific modelling.

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

Future work. The larger study will enlarge and randomise the sample, define the target from an independently measured outcome such as growth in a later period, collect comment sentiment and retention data, elicit weights from external experts and apply the full COMET procedure, calibrate probabilities on separate data, and extend the framework to other platforms such as TikTok and Instagram Reels.

5. CONCLUSION

This preliminary study constructed a multidimensional composite index of YouTube channel success from four dimensions, namely popularity, engagement, consistency, and content quality (a proxy), for 95 Indonesian-focused channels, using a transparent weighted additive aggregation [6] whose ranking is stable under plausible changes of the weights. The quality proxy is almost collinear with engagement (ρ = 0.98). A leakage audit showed that predicting a subscriber-based label from subscriber-derived features yields an AUC near 0.91 that reflects the labeling rule. With these variables excluded, behavioural metrics predicted category-relative audience standing with a cross-validated AUC of 0.70 (95% CI 0.60-0.81), rising to 0.75 (95% CI 0.64-0.84) when median views per video were added, better than chance (permutation p = 0.010) but with wide uncertainty. These findings are exploratory. The planned larger study will use a larger randomised sample, an independently measured outcome, sentiment and retention data, expert-elicited weights, and external validation.

APPENDIX A. METRIC INVENTORY

Table A1 lists all metrics used in the study, their definitions, and their role in the composite index and in the predictive feature sets (A, B, C, X; Section 3.4).

TABLE A1. Metric inventory and mapping to dimensions

Dimension	Metric	Definition	Index	Sets
Popularity	Median views per video	Median per-video views; log(1 + x)	P	B, X
Engagement	Like rate (LR)	Sum of likes / sum of views x 100	E	A, B, X
Engagement	Comment rate (CR)	Sum of comments / sum of views x 100	E	A, B, X
Engagement	Mean engagement rate (ER)	Mean of per-video (likes + comments) / views x 100	E	A, B, X
Consistency	Upload frequency (f)	Videos per month over the observation window; log(1 + x) as predictor	C	A, B, X
Consistency	View variation (CV)	SD / mean of per-video views x 100	C	A, B, X
Quality (proxy)	Audience appreciation	(0.6 LR + 0.4 CR) / 2	Q	-
Other behaviour	Mean video duration	Mean duration in seconds; log(1 + x)	-	A, B, X
Other behaviour	Short-video share	Share of videos of 60 s or less	-	A, B, X
Dimension scores	P, E, C, Q	Min-max normalised dimension values (Eq. 1)	-	C
Audience size	Subscriber count (and ratios)	Channel subscriber count	T	X only

Metrics are computed per channel from the de-duplicated records of the (up to) 100 most recent videos. LR, CR, and ER denote like rate, comment rate, and mean engagement rate. Index: dimension in which the metric enters the composite index (P, E, C, Q); T: used only to define the target. All formulas are implemented in the supplementary analysis script.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Institut Citra Buana Indonesia for the moral and material support provided throughout this research. It is hoped that this support will bring meaningful progress to the institution in particular and to society at large.

REFERENCES

[1] U. Pujianto, "Comparative analysis of bagging and boosting algorithms on the classification of the popularity of educational-themed YouTube videos," in *2021 7th International Conference on Electrical, Electronics, Information Engineering, and Technology Breakthrough for Greater New Life (ICEEIE)*, Jakarta, Indonesia, 2021, pp. 1-6.

[2] P. Bhattacharya, "A tale of machine learning based YouTube ad view prediction: analysis and interpretation," *IET Conference Proceedings*, vol. 2023, no. 44, pp. 337-343, 2023.

[3] A. Shutsko, "User-generated short video content in social media: a case study of TikTok," in *Lecture Notes in Computer Science*, vol. 12195, 2020, pp. 108-125.

[4] R. Lijo, "Comparing educational and dissemination videos in a STEM YouTube channel: a six-year data analysis," *Heliyon*, vol. 10, no. 3, p. e24856, Feb. 2024.

[5] Z. Meng, "Chinese video creator identities - a cross-platform social media perspective," *Platform*, vol. 9, pp. 24-42, 2022.

[6] W. Sałabun, "The Characteristic Objects Method: a new distance-based approach to multicriteria decision-making problems," *Journal of Multi-Criteria Decision Analysis*, vol. 22, no. 1-2, pp. 37-50, 2015, doi: 10.1002/mcda.1525.

[7] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "A multi-dimensional evaluation of synthetic data generators," *IEEE Access*, vol. 10, pp. 11147-11158, 2022.

[8] K. Subha, "Leveraging spark-based machine learning algorithm for audience sentiment analysis in YouTube content," *Intelligent Data Analysis*, vol. 28, no. 5, pp. 1395-1405, 2024.

[9] B. R. Babu, M. S. Khan, A. Agrawal, M. Gupta, and D. Choudhary, "Hybrid NLP framework for enhanced sentiment analysis and topic detection on YouTube," *International Journal of Basic and Applied Sciences*, vol. 14, no. 1, pp. 304-313, 2025.

[10] A. K. Jha, V. Singh, A. Kumar, and K. K. Singh, "Influencer recommendation system: choosing the right influencer using a network analysis approach," *Marketing Intelligence & Planning*, vol. 41, no. 8, pp. 1197-1212, 2023.

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License

Vol. 2 No.1 November 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v2i1.47

[11] M. Noerhadi, Budiman, and Sardjono, "Comparative analysis of machine learning algorithms for Indonesian Twitter sentiment classification on the Jakarta-Bandung high-speed rail project," *Bulletin of Intelligent Machines and Algorithms*, vol. 1, no. 1, 2025, doi: 10.65780/bima.v1i1.3.

[12] Darsiti, "A comprehensive machine learning approach for predicting beats per minute (BPM) in music using audio features," *Bulletin of Intelligent Machines and Algorithms*, vol. 1, no. 1, 2025, doi: 10.65780/bima.v1i1.2.

[13] D. Widiyanto, "High-precision credit card fraud detection on imbalanced data using Random Forest and 1D convolutional neural networks," *Bulletin of Intelligent Machines and Algorithms*, vol. 1, no. 1, 2025, doi: 10.65780/k6hexq72.

[14] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010.

[15] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, art. 15, 2012.

[16] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023.

[17] G. Varoquaux, "Cross-validation failure: small sample sizes lead to large error bars," *NeuroImage*, vol. 180, pp. 68-77, 2018.

[18] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *Journal of Machine Learning Research*, vol. 11, pp. 1833-1863, 2010.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, PMLR vol. 70, 2017, pp. 1321-1330.

[20] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1-3, 1950.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[22] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794.

[23] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 3146-3154.

[24] F. Pedregosa et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

Published by Maheswari Publisher
Creation disseminated under Creative Commons Attribution 4.0 International License