Analysis of the Best Social Media Platforms for Promotion Using Machine Learning and RFE Feature Selection: A Comparative Study of Gradient Boosting, XGBoost, CNN, and SVR

Maulina Putri [1]*, Aria Hendrawan [2]**
* Department of Informatics Engineering, Semarang University
** Faculty of Information and Communication Technology, Semarang University
g231210064@student.usm.ac.id [1], ariahendrawan@usm.ac.id [2]

Article Info

Article history:

Received 2025-12-17
Revised 2025-12-26
Accepted 2026-01-08

Keyword:

Machine Learning,
KDD,
Social Media Promotion,
CNN,
Feature Selection RFE.

ABSTRACT

This study aims to identify the most effective social media platforms for digital marketing. The use of social media for promotion continues to grow, yet many businesses still struggle to determine which platforms have the greatest impact. Therefore, this study compares the performance of various machine learning platforms to predict the best platform. The algorithms used are Gradient Boosting Regressor, XGBoost Regressor, Convolutional Neural Network (CNN), and Support Vector Regression (SVR) to estimate digital conversion potential based on user reviews, ad reach, and content trend patterns. A Knowledge Discovery in Databases (KDD) workflow is used to identify the most important key factors. This process includes data preprocessing, TF-IDF feature extraction, sentiment analysis, feature engineering, and feature elimination (RFE). The results showed that the CNN algorithm excelled in prediction, with the highest R^2 score of 0.74 and the lowest RMSE of 14.78. CNN predictions showed YouTube topping the list in terms of conversion potential, followed by Facebook and TikTok. These results highlight the higher promotional effectiveness of video-based platforms and the importance of machine learning in digital marketing decision-making. However, this study is limited by its reliance on static user review and ad reach data, which may not fully capture the dynamic changes of social media platforms.

I. INTRODUCTION

With the advancement of digital technology, social media has become a vital part of marketing strategies. Social media platforms like Facebook, Twitter, Instagram, YouTube, and TikTok use a variety of different features and algorithms to reach their audiences. As a result, businesses find it challenging to choose the best social media platform for promotion. The use of social media for promotion heavily depends on creative communication strategies that build positive relationships with consumers through storytelling and direct interaction [1]. Social media influencers are also highly effective in enhancing brand loyalty through two-way interaction [2]. Promotions on platforms like Instagram, Facebook, and TikTok are popular for reaching younger audiences and maximizing engagement [3]. High-quality communication strategies on social media can increase user engagement and strengthen public trust in digital information [4].

This study aims to analyze and compare the most effective social media platforms for promotion using machine learning methods. The methods employed include Gradient Boosting, XGBoost, Convolutional Neural Network (CNN), and Support Vector Regression (SVR), with Recursive Feature Elimination (RFE) used for feature selection to identify the most influential features affecting promotional effectiveness. The results are expected to assist businesses in selecting the most effective social media platforms and serve as a reference for applying machine learning in digital marketing analysis. According to the study conducted by [5], entitled A Study of Using Social Media for Luxury Brand Promotion, it is recommended that luxury brands strengthen the synergy between physical and online stores, build digital storefronts, and establish connections with online shopping malls. By emphasizing quality in product design and marketing strategies, brand identity and premium image must be maintained. Furthermore, in 2022, the study [6], entitled Social Media Health Promotion and Audience Engagement showed that the success of health campaigns on social media depends on two-way interactions between the organization and its audience. Therefore, an effective health promotion strategy on social media must prioritize the dissemination of information through educational content, interactive messages, and self-confidence-boosting materials.

With advances in data analytics technology, social media has transformed from merely a platform for organic promotion into a data-driven ecosystem supported by machine learning (ML). The application of ML in the Small and Medium Enterprise (SME) sector enables automated analysis, audience segmentation, optimal upload timing, and more targeted content recommendations. This implementation improves cost efficiency and expands the reach of digital promotion. Additionally, the integration of machine learning and Natural Language Processing (NLP) has become a key element in digital marketing strategies, as it allows large-scale personalization and predicts user behavior trends [7]. Methodologically, the feature selection process is an essential step before training classification models. Algorithms such as Gradient Boosting and Extreme Gradient Boosting (XGBoost) can be utilized to rank feature importance, while iterative methods like Recursive Feature Elimination (RFE) are effective in reducing data dimensionality, accelerating the training process, and improving model generalization. Proper feature selection helps reduce noise, prevent overfitting, and clarify the impact of each variable on promotional performance indicators, such as click-through rate (CTR) and conversion rates [8].

Even while research on the efficacy of social media marketing is expanding, the majority of studies that are now available concentrate on engagement or optimization measures within a particular platform and industry context. within a single platform and industry context. Previous research has extensively addressed the role of content quality, influencer marketing, and audience interaction in increasing engagement and brand loyalty [1], [2], [4]. Other studies have applied machine learning techniques to optimize social media campaigns; however, they often rely on limited feature sets, such as engagement or demographic variables, without integrating heterogeneous data sources across platforms [7].

Additionally, there is still a lack of research on comparison modeling across various social media sites utilizing systematic feature selection techniques. Specifically, prior research has not adequately addressed the explicit application of Recursive Feature Elimination (RFE) to identify the most influential predictors from integrated sentiment analysis, ad reach, and content trend features. To objectively identify the best social media platforms for digital promotion, there is a research need to create a data-driven, cross-platform framework that concurrently incorporates multidimensional promotion metrics and compares several machine learning models.

II. METHOD

This study applies a comparative modeling approach to evaluate the performance of various machine learning algorithms in determining the most effective social media platforms for digital promotion. The four algorithms compared include Gradient Boosting Regressor, Extreme Gradient Boosting (XGBoost Regressor), Convolutional Neural Network (CNN), and Support Vector Regression (SVR). Prior to the model training stage, feature selection was performed using the Recursive Feature Elimination (RFE) method to identify the variables that most influence promotion effectiveness. Furthermore, the performance of each model was evaluated using four main metrics, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score, which assess the accuracy, stability, and generalization ability of the model in prediction.

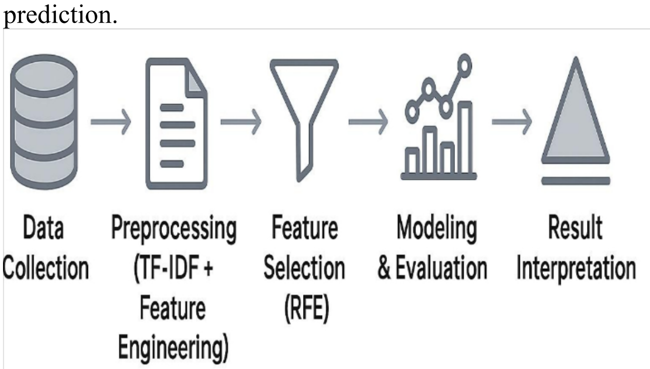

Figure 1. Steps of the KDD Process.

This figure comprehensively illustrates the workflow of the Knowledge Discovery in Databases (KDD) process implemented in this study. The process begins with data collection, obtained from digital trend and social media reports such as DataReportal, Detiknet, and user reviews on the Google Play Store. The data then undergoes preprocessing, including data cleaning, text normalization, stop-word removal, and feature representation using the Term Frequency–Inverse Document Frequency (TF-IDF) method. Feature engineering is subsequently performed to construct derived variables representing elements of social media promotion. The process continues with feature selection using Recursive Feature Elimination (RFE) to choose the most relevant feature subset while eliminating redundant or less informative features. RFE improves the predictive performance of the models by focusing on the most important features [9].

Model training was conducted by training four algorithmic models, namely Extreme Gradient Boosting (XGBoost Regressor), Gradient Boosting Regressor, Support Vector Regression (SVR), and Convolutional Neural Network (CNN). Each algorithm was trained using the feature-selected data and subsequently evaluated to determine the best performance in predicting the effectiveness of social media platforms as promotional tools. Model evaluation was performed using four main metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. The model with the lowest MAE, MSE, and RMSE values and the highest R^2 score was considered to have the best performance. The comparison results were then used to identify the most efficient algorithm for selecting the optimal social media platform for promotional activities. The final stage involves result interpretation, which includes analyzing the performance of the best model and its strategic implications in utilizing social media as a digital promotion tool.

A. Data Collection

The data used in this study consist of three main categories, namely user reviews from Google Play Store, social media advertising reach data, and social media content trend data for 2025. User reviews were obtained through web scraping of the official applications of five major social media platforms, namely Facebook, Instagram, TikTok, X (Twitter), and YouTube, each consisting of 200 reviews, resulting in a total of 1,000 user reviews. These reviews contain text comments and star ratings (scale 1–5). These ratings serve as indicators of user satisfaction and implicitly reflect the effectiveness of the platforms in digital promotional activities. Accurate and representative data collection is a fundamental step to ensure the quality of machine learning models. The data collection process must consider completeness, diversity, and class balance to minimize bias and improve the reliability of analysis results[10]. Furthermore, proper and representative data collection is a crucial stage that can reduce bias and enhance the generalization ability of machine learning models [11]. Therefore, utilizing automated web scraping to collect user reviews provides a consistent and reproducible data base for subsequent text analysis and modeling. Due to technical limitations of scraping tools, which do not yet fully support consistent timeframe filtering across platforms, user reviews are collected based on the most recent data available at the time of the scraping process. Thus, review data represents current user perceptions, not a specific historical period.

Meanwhile, advertising reach data was collected from relevant industry sources such as DataReportal Indonesia Digital 2025 and the We Are Social Report, providing estimates of potential audience size and demographic characteristics of users on each platform. Social media content trend data was obtained from the 2025 Indonesian digital marketing research publications, which include information on the most preferred content types, user engagement levels, and dominant brand communication strategies on each platform.

B. Data Preprocessing and Feature Engineering

The preprocessing stage in this study aims to transform unstructured social media reviews into a more consistent and easily processable text format. The cleaning procedure includes converting all characters to lowercase, removing numbers and punctuation, and filtering “stopwords” with an extended list to improve contextual accuracy. Additionally, a fallback mechanism is applied to ensure that at least one token remains available if all words are filtered during the cleaning process. This strategy aligns with previous studies showing that the order and quality of preprocessing steps significantly affect the effectiveness of Natural Language Processing (NLP) models [12]. The cleaned text data is then converted into numerical representations using TF-IDF with a feature limit (max_features) to prevent the curse of dimensionality, which can subsequently be combined with additional numerical variables. Recent studies also confirm that TF-IDF remains a robust baseline approach in social media text analysis and is easily integrated with other numerical features [13].

TABLE I.
BEFORE–AFTER CLEANING TEXT

Platform	Raw Review (Excerpt)	Cleaned_Reviews
Facebook	1. Durasi video salah tekan langsung gak	durasi video salah tekan langsung gak di
Facebook	Alhamdulillah	alhamdulillah
Instagram	Akun saya gak bisa di bukak	akun gak bukak terkena baned
Instagram	Instagram makin hari makin jelek banget. Sering ngelag gak jelas	instagram jelek ngelag
TikTok	Aku mau nya masuk cepat dongg tiktok nya	masuk cepat dongg tiktok
TikTok	Aplikasi ini sangat bagus	bagus
X Twitter	Akun premium pun bisa dibanned	akun premium dibanned
X Twitter	kok sekarang aplikasi ini sering keblock akun secara tiba tiba tanpa ada alasan ,aneh ??!!	keblock akun alasan aneh
Youtube	sangat membantu	membantu
Youtube	senang nonton di disini serba lengkap	senang nonton serba lengkap

This table presents the original user reviews and their cleaned versions after preprocessing for further analysis in the machine learning pipeline.

The feature engineering pipeline was continued with sentiment analysis, calculation of keyword frequency related to promotion and user engagement, as well as determination of content trend weights. The sentiment of each review was analyzed using Text Blob, producing a polarity score ranging from −1 to +1 (negative values classified as negative sentiment, 0 as neutral, and positive values as positive sentiment). Sentiment labels were then counted and their percentages were calculated for each platform. Content trend weights were combined with user ratings to calculate the Engagement Rate using:

$$Engagement\ Rate\ = \frac{TrendWeight \times Rating}{5} \times 100$$

Next, the Engagement Rate was used to estimate digital campaign conversion through:

$$Conversion\ = \frac{EngagementRate\ \times\ AdReach}{100}$$

where Ad Reach is expressed in millions of users. All features from the previous TF-IDF stage, keyword frequency, sentiment, trend weights, ad reach, and estimated conversions were combined into a comprehensive dataset for model training, enabling the system to fully represent digital marketing behavior.

C. Feature Selection

The feature engineering process produced 16 attributes representing various aspects of user interactions and content characteristics, including TF-IDF weights, sentiment scores, keyword frequency, content trend weights, and advertising reach data. To simplify data dimensionality, improve model generalization, and accelerate training, feature selection was performed using the Recursive Feature Elimination (RFE) method.

RFE is a wrapper method that uses an estimator to evaluate the importance of each feature and gradually removes less relevant features until the optimal number of features is obtained. This approach has been proven effective across various research domains as it can identify features with the highest predictive contribution, especially in datasets with heterogeneous feature types [14].

In implementation, all features were first converted into numeric format, and missing values were handled using simple imputation. A stable base model resistant to inter-feature correlations was then established. Random Forest Regressor was chosen due to its robustness to data scale and ability to capture nonlinear relationships among variables. In this study, RFE was configured to select the ten best features from all available attributes. The selection procedure was then verified to ensure that the chosen features were suitable for the data characteristics and their relationship with the target variable.

TABLE 2.
RFE FEATURE SELECTION RESULTS

NO	Feature Name	Ranking RFE	Selection Status
1	Rating	1	Selected
2	Advertising reach	1	Selected
3	Trend weights	1	Selected
4	Engagement_Rate	1	Selected
5	konten	1	Selected
6	iklan	1	Selected
7	followers	1	Selected
8	trending	1	Selected
9	like	1	Selected
10	video	1	Selected
11	share	2	Not Selected
12	engagement	3	Not Selected
13	brand	4	Not Selected
14	Deskripsi	5	Not Selected

Of the 16 initial features, 10 were deemed most predictive for the target variable. The selected features were rating, ad reach (millions), trend weight, engagement rate, content, ad, followers, trending, like, and video. Share, engagement, brand, and description features were not selected due to their lower predictive contribution. Numerical attributes like rating, ad reach, engagement rate, and trend weight provide significant quantitative information on content and ad performance, according to an analysis of the chosen features. The model was able to capture user interaction patterns and content popularity by using TF-IDF to convert text-based variables such as followers, trending, video, brand, share, and like into numerical representations. By eliminating less relevant features, the model benefits from dimensionality reduction, improved generalization, and faster training. In keeping with earlier research showing the efficacy of RFE for heterogeneous datasets, this selection retains only the features with the highest predictive contribution, reducing overfitting risk while maintaining strong model performance.

D. Data Splitting

The data splitting stage was conducted after the feature selection process to ensure that the predictive model was trained using a subset of data different from the one used in testing. This strategy aims to prevent data leakage and ensure that model performance evaluation remains objective. In this stage, features selected through the RFE method were used as predictor variables (X), while the Conversion (million) variable was designated as the target (y). All features were converted into numeric format, and missing values were handled to maintain data consistency and integrity.

The train–test split was then applied with 80% as training data and 20% as testing data, using a fixed random_state value to ensure replicability of the experimental results. Selecting the appropriate data split ratio is critical because the training–testing split proportion directly affects model accuracy and generalization, whereas an improper ratio may introduce bias in performance evaluation [15]. Moreover, studies [16] indicate that a 70–80% range for training data generally provides an optimal balance between the model's learning capacity and the representation of the test data distribution, supporting the use of an 80:20 ratio in this study.

E. Model Training

In the model training stage, this study utilized four machine learning algorithms, namely Gradient Boosting Regressor, XGBoost Regressor, Convolutional Neural Network (CNN), and Support Vector Regression (SVR) to predict conversion values based on features selected through the RFE process. The choice of Gradient Boosting and XGBoost considered their ability to handle non-linear regression patterns and iterative optimization mechanisms in tree ensemble structures. Several studies report that XGBoost can provide consistent predictions on complex datasets through a staged boosting process [17], while Gradient Boosting is widely used in modern regression due to its ability to iteratively reduce errors and capture non-linear relationships among variables [18]. The effectiveness of boosting approaches is also supported by findings that hybrid deep learning models can improve prediction accuracy on complex dynamic data, as seen in ensemble models for financial time series forecasting [19].

Moreover, the CNN model was used as a deep learning approach to extract deeper feature representations from both numeric and text-based variables. To strengthen performance comparison, this study also implemented Support Vector Regression (SVR), which is known to be effective in handling non-linear regression using kernel functions. The relevance of SVR is reinforced by studies showing that blurred optimization-based SVR can achieve high prediction accuracy on fluctuating datasets, such as airport cargo volume forecasting [20].

Although Convolutional Neural Networks (CNN) are typically used for processing image or sequential data, this study utilizes a CNN as a fully connected neural network to perform regression on heterogeneous tabular data. The model takes as input numerical features and TF-IDF representations of text reviews selected through the RFE method. To capture complex non-linear relationships between features while reducing the risk of overfitting, the model is built with multiple dense layers equipped with dropout regularization and the Rectified Linear Unit (ReLU) activation function. With this architecture, the model is able to learn high-level feature interactions without relying on temporal or spatial relationships, making it methodologically appropriate to the characteristics of the data used in the study.

All algorithms were trained using the previously processed training–testing dataset. Training for Gradient Boosting, SVR, and evaluation metric calculations was performed using the scikit-learn library, while the XGBoost model was executed using the XGBoost library. The CNN model was constructed with a fully connected architecture based on Keras/TensorFlow. Each model was trained using basic parameter configurations appropriate for its algorithm, and the prediction outputs were saved in Excel format for further analysis and comparative performance evaluation across algorithms.

F. Model Evaluation

To assess the predictive performance of each regression model, this study employed evaluation metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE provides an estimate of the average absolute error, while MSE and RMSE are more sensitive to large errors (outliers), and R^2 is used to evaluate the proportion of target variability explained by the model. Recent literature on regression model evaluation emphasizes that using a combination of these metrics is necessary to obtain a more comprehensive and robust performance assessment against data variations [21].

G. Model Performance Comparison

Model selection and performance comparison were conducted after all regression algorithms were trained. Predictive performance was evaluated using R^2, RMSE, and MAE, with the best model determined based on the combination of the highest R^2 and the lowest RMSE/MAE. This approach aligns with practices in regression literature, which indicate that a high R^2 value along with low RMSE and MAE reflects an accurate and reliable model [22]. Comparative metric visualizations were used to provide an overview of the reliability of each algorithm, while the model with the best performance was selected as the final model for further analysis and cross-algorithm comparison.

H. Result Interpretation

Result interpretation was conducted after selecting the best model based on performance evaluation (highest R^2 and lowest RMSE/MAE). Conversion predictions from this model were linked to each platform to calculate the average conversion per platform, enabling the identification of platforms with the highest potential performance. The summary of predictions was visualized using a heatmap, providing a comparative overview of the relative effectiveness of each platform. All interpretation data was stored for further analysis and to support data-driven decision-making in digital campaign strategies.

III. Results And Discussion

This section presents the empirical results obtained from data processing, feature extraction, RFE feature selection, model training, and evaluation of the applied regression algorithms. The results focus on the research outputs, including user behavior analysis, digital promotion effectiveness on each platform, and comparative accuracy of models in predicting conversion values.

A. User Review Sentiment Analysis

Sentiment analysis was applied to user reviews from five social media platforms, categorized as positive, neutral, and negative. The analysis results indicate that each platform exhibits distinct patterns of user perception.

TABLE 3.
SENTIMENT DISTRIBUTION BY PLATFORM

Platform	Negative (n, %)	Netral (n, %)	Positif (n, %)	Total
Facebook	3 (1.5 %)	171 (85.5%)	26 (13%)	200
Instagram	2 (1%)	168 (84%)	30 (15%)	200
TikTok	1 (0.5%)	157 (78.5%)	42 (21%)	200
X Twitter	5 (2.5%)	180 (90%)	15 (7.5%)	200
Youtube	3 (1.5%)	177 (88.5%)	20 (10%)	200

The sentiment distribution across all platforms shows a dominance of neutral sentiment, ranging from 78.5% to 90%. TikTok has the highest proportion of positive sentiment at 42%, while the highest negative sentiment is found on X Twitter, at 2.5%.

B. Keyword and Content Preference Analysis

Keyword frequency analysis was conducted to identify the main focus of user reviews regarding the effectiveness of social media platforms as digital promotion tools. Keywords such as ads, video, like, content, and followers were counted on each platform. The results provide an overview of the most frequently discussed topics and the interaction characteristics of each platform. These patterns were then visualized in a heatmap to illustrate the intensity of keyword occurrences across the social media platforms.

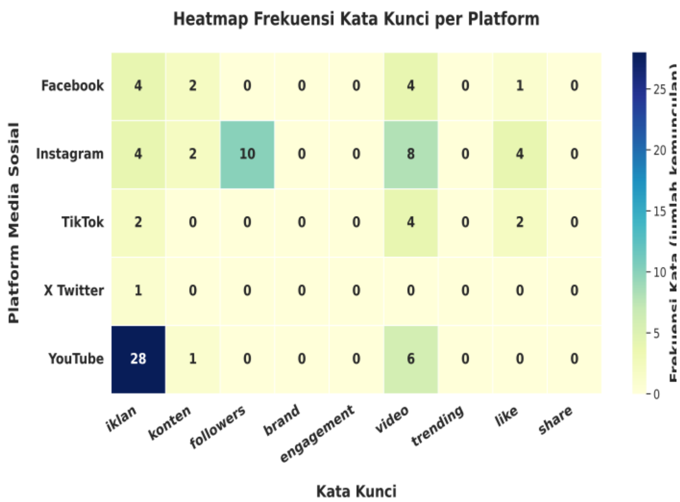

Figure 2. Keyword Frequency Heatmap

The heatmap visualization shows that each social media platform has a distinct keyword frequency pattern. YouTube recorded the highest frequency for the keyword "ads" (28), followed by "video" (6), indicating that user conversations on this platform are heavily related to ad exposure and the dominance of video-based content. Instagram exhibited relatively high frequencies for the keywords "followers" (10), "video" (8), "ads" (4), and "like" (4), reflecting the platform's focus on audience growth, visual content, and user interaction.

Facebook exhibited a more balanced pattern, with the keywords "ads" (4) and "video" (4) appearing equally dominant, accompanied by mentions of "content" (2) and "like" (1). This indicates that user discussions on Facebook still revolve around content consumption and promotional activities. TikTok had a relatively lower keyword frequency, but the keyword "video" (4) remained prominent, emphasizing TikTok's primary character as a video-focused platform.

In contrast, X (Twitter) shows very limited discussion intensity, with only one occurrence of the keyword “advertisement” (1) in the analyzed keyword set. Overall, the heatmap results confirm that each platform has a distinctive communication character: YouTube focuses on advertising and video content, Instagram stands out in terms of engagement and audience, Facebook combines promotional and content discussions, TikTok remains focused on video, while X (Twitter) shows relatively minimal keyword activity. These differences provide an important basis for designing a more specific digital promotion strategy that is in line with the characteristics of each platform.

C. Engagement Rate and Conversion Analysis

To evaluate the effectiveness of each platform in digital promotion activities, this study uses two main indicators, namely Engagement Rate and Conversion. Engagement Rate is obtained from the calculation of content trends and user ratings, while conversion is calculated based on the combination of engagement level and ad reach.

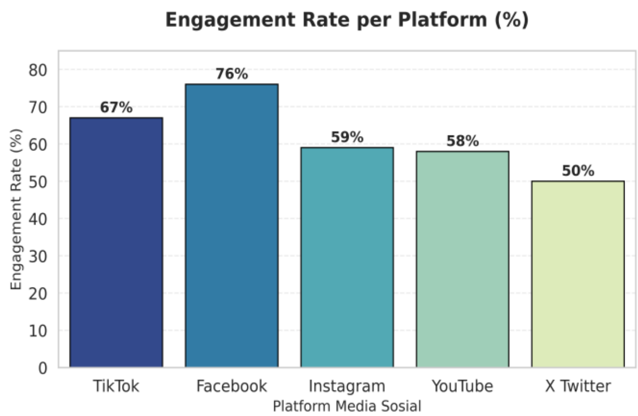

Figure 3. Engagement Rate Results

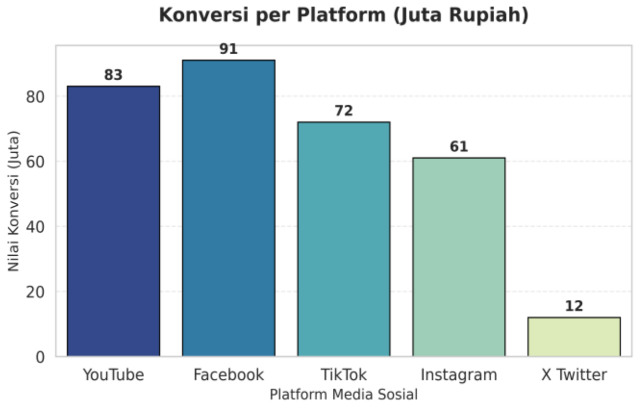

Figure 4. Conversion Results

Data visualizations show that Facebook has the highest engagement and conversion rates compared to other platforms, followed by TikTok and YouTube, while Instagram ranks in the middle and X (Twitter) performs the lowest. This pattern indicates that platforms with robust visual and video content ecosystems tend to generate optimal engagement and conversions, making them more worthy of prioritization in digital marketing strategies.

D. Model Performance Comparison

This section presents the evaluation results of the four algorithms used, namely Gradient Boosting Regressor, XGBoost Regressor, CNN, and SVR, aiming to identify the most accurate model in predicting conversion values for each platform.

TABLE 4.
MODEL EVALUATION MATRIX

MODEL	MAE	MSE	RMSE	R^2
Gradient Boosting	16.64	844.70	29.06	- 0.01
XGBoost	14.97	580.17	24.09	0.31
SVR	20.89	814.85	28.55	0.3
CNN	13.67	218.48	14.78	0.74

The evaluation results for the four models show significant differences in their performance. CNN has the best prediction for conversion because it has the best performance, with an RMSE of 14.78 and an R^2 of 0.74. This means that CNN is able to predict complex and non-linear data. It is followed by XGBoost in performance with an R^2 of 0.31 and an RMSE of 24.09. This means that XGBoost is able to predict the target variable with some degree of variability, although it does not predict as accurately as CNN. In contrast, Gradient Boosting has an even lower performance than the simple baseline model with an R^2 of -0.01 and an RMSE of 29.06. Meanwhile, SVR has the worst overall performance with an R^2 of 0.3 and an RMSE of 28.55, which means that this model is unable to model the data with any degree of success, meaning that it has more limitations in trying to predict the data.

To strengthen the comparative interpretation of model performance, the R^2 and RMSE values were also visualized in the following graphs to provide a clearer performance contrast among the evaluated algorithms.

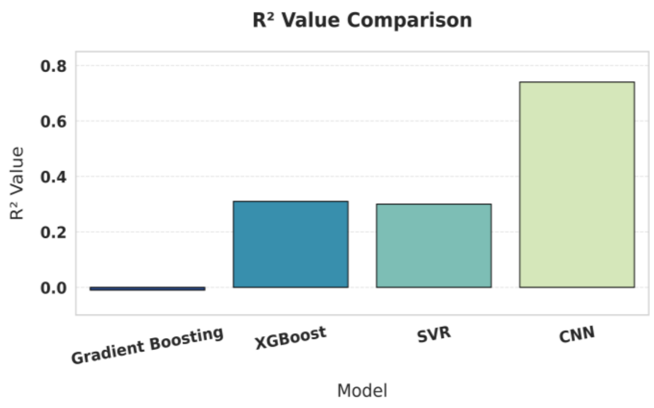

Figure 5. R^2 Value Comparison

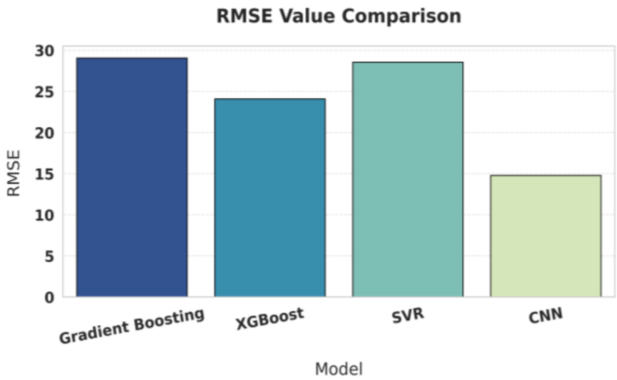

Figure 6. RMSE Value Comparison

The visualization of R^2 and RMSE further confirms the findings presented in the previous analysis. Among the models evaluated, the CNN-based deep neural network with fully connected layers achieved the highest prediction accuracy, whereas Gradient Boosting, XGBoost, and SVR showed relatively weaker performance. The model was particularly effective in capturing complex and nonlinear relationships across heterogeneous features. Nevertheless, owing to its architectural complexity, this deep neural network is more susceptible to overfitting. Consequently, careful data management is required, and additional validation techniques, such as cross-validation or testing on a larger dataset, are recommended to obtain more robust and reliable results.

E. Predicted Conversion Interpretation Platform

After evaluating the R^2, RMSE, and MAE metrics, CNN was identified as the best-performing model, and digital conversion predictions were conducted per platform to assess the effectiveness of each social media channel. This process involved mapping the test data indices back to their original platforms and then calculating the average conversion values predicted by the CNN model for each platform.

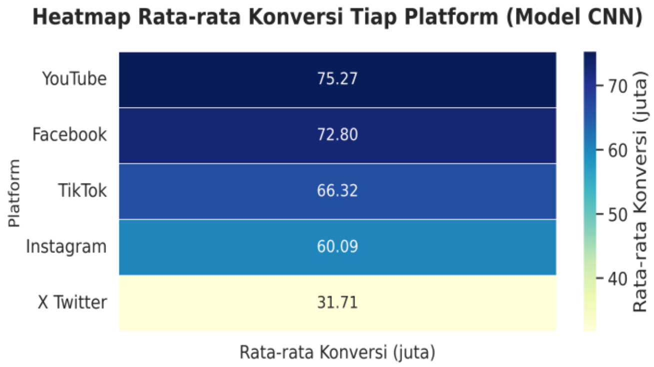

Figure 7. Average Conversion Results (Million)

The visualization shows clear differences in conversion intensity among the platforms, where areas with darker colors represent higher conversion values. YouTube emerges as the platform with the highest conversion at 75.27 million, indicating that long-form video content and high user consumption significantly contribute to promotional effectiveness. Facebook follows with 72.80 million, supported by strong visual content and broad advertising reach. TikTok records 66.32 million, demonstrating the effectiveness of trend-based and short video content in driving engagement. Instagram attains 60.09 million and ranks in the middle, while X (Twitter) has the lowest conversion at 31.71 million, reflecting the limitations of text-based platforms in driving conversions. Overall, based on the CNN model predictions with the highest accuracy in this study, the heatmap indicates that YouTube is the most effective platform for digital promotion, followed by Facebook and TikTok.

Regardless of the score, the advantages of YouTube, Facebook, and TikTok encompass their user characteristics, primary content types, and market needs in today's digital age. YouTube is dominated by users actively seeking in-depth product reviews, making it ideal for industries like technology, education, tourism, and other products that require thorough consideration before purchasing [23]. With a larger and more demographically diverse user base, Facebook benefits SMBs, retail brands, and service-based industries that require broad audience reach and organized ad targeting. Meanwhile, TikTok attracts a younger audience with short attention spans and high engagement behavior, making it highly relevant for promotional campaigns for fast-moving consumer goods, fashion, lifestyle, and lifestyle products [24].

Studies [25] show that the user growth dynamics and demographics of each social media platform differ, and these differences significantly impact the success of digital marketing strategies. YouTube, TikTok, and Facebook reportedly have distinct audiences despite their large user bases. Users currently use all three platforms as sources of product information and as a consideration before purchasing.

Furthermore, research shows that visual and video-based content can increase emotional engagement, build trust, and strengthen user loyalty, especially on platforms like YouTube and TikTok. TikTok is considered to have strong marketing potential due to its interactive short-form video format and its ability to capture cross-generational attention through trend-driven content. Overall, audience characteristics and industry needs indicate that video-based platforms generally exhibit greater conversion potential than text-based platforms like Twitter.

These findings align with the research objectives and confirm that video-based platforms have a higher capacity to generate conversions compared to other platforms. These results are consistent with previous studies [7], which show that machine learning algorithms can identify social media platforms with the highest conversion contributions. The study also concluded that video-dominated platforms tend to achieve higher engagement rates and campaign performance. These findings are consistent with the CNN model predictions in this research, where YouTube, Facebook, and TikTok, as visual content-based platforms, achieved the highest conversion values. The main difference lies in the scope of the study, as prior research focused on optimizing MSME campaigns through automated systems, whereas this study examines cross-platform comparisons considering user behavior, sentiment, and content trends. Despite the different contexts, both studies emphasize that platform characteristics and content type are crucial factors in determining digital promotion success.

IV. CONCLUSION

This study was conducted to identify the most effective social media platforms for digital promotion by applying machine learning methods and RFE feature selection. The four regression algorithms used in this study include Gradient Boosting Regressor, XGBoost Regressor, Support Vector Regression (SVR), and Convolutional Neural Network (CNN). Performance evaluation based on MAE, MSE, RMSE, and R^2 metrics indicates that CNN is the best-performing model, marked by the highest R^2 value of 0.74 and the lowest RMSE and MAE values.

The interpretation of predictions using CNN shows significant differences in conversion across platforms. YouTube is recorded as the platform with the highest conversion potential, followed by Facebook and TikTok, while Instagram falls into the mid-range category, and X (Twitter) has the lowest conversion. These findings indicate that video-dominated platforms have stronger promotional effectiveness compared to text-based platforms, given the high consumption of visual content and user interaction.

However, this study has several limitations. First, the data used relies heavily on user reviews, which are subjective, potentially biased, and contain sentiment. Furthermore, lexicon-based sentiment analysis methods are limited in understanding informal language contexts. This study also used static ad reach and content data, which does not fully represent the dynamic changes in algorithms within social media platforms.

From a practical perspective, the results of this study can help business practitioners make decisions about promotional media selection. YouTube is recommended for advertising on educational activities and product storytelling, while Facebook and TikTok are preferred for advertising on engagement-oriented activities and viral content. Future research is recommended to process data on a larger scale, using a deep learning-based sentiment analysis approach and real-time data to generate predictions that are more responsive to social media changes.

REFERENCES

[1] W. Huang, "How Can a Brand Utilize Social Media to Achieve Impressive Promotion," *Commun. Humanit. Res.*, vol. 16, no. 1, pp. 116–120, 2023, doi: 10.54254/2753-7064/16/20230564.

[2] M. N. Najmi Nuji, A. Ali, W. N. Wan Noordin, B. A. N. Mohamed Thaheer, and V. Mathiew, "Of Trust and Influence: A Look At Social Media Influencers and Brand Promotion," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 13, no. 6, pp. 2152–2170, 2023, doi: 10.6007/ijarbss/v13-i6/15444.

[3] N. S. Amdan *et al.*, "Optimalisasi Pemanfaatan Media Sosial dalam Promosi Kampung Wisata Pekijing," *J. Pengabdi. Pada Masy.*, vol. 7, no. 1, pp. 72–79, 2022, doi: 10.30653/002.202271.26.

[4] M. Cinelli *et al.*, "Promoting engagement with quality communication in social media," *PLoS One*, vol. 17, no. 10 October, pp. 1–19, 2022, doi: 10.1371/journal.pone.0275534.

[5] J. Li, "A Study of Utilizing Social Media for Luxury Brand Promotion," vol. 35, pp. 373–379, 2024.

[6] S. Jiang, "Social Media Health Promotion and Audience Engagement : The Roles of Information Dissemination , Organization-Audience Interaction , and Action Confidence Building Social Media Health Promotion and Audience Engagement : The Roles of Information Dissemination , Organization-Audience Interaction , and Action Confidence Building," no. November, 2022, doi: 10.1080/10410236.2022.2150809.

[7] W. E. Kedi, C. Ejimuda, and C. Idemudia, "Machine learning software for optimizing SME social media marketing campaigns," vol. 5, no. 7, pp. 1634–1647, 2024, doi: 10.51594/csitrj.v5i7.1349.

[8] N. C. Ramadhan, H. H. H, T. Rohana, and A. M. Siregar, "TIN : Terapan Informatika Nusantara Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur Xgboost Untuk Klasifikasi Kanker Payudara TIN : Terapan Informatika Nusantara," vol. 5, no. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408.

[9] N. Sulistianingsih, S. Agrippina, A. Yusuf, F. N. Sari, and M. P. Juniarta, "JTIM : Jurnal Teknologi Informasi dan Multimedia Analisis Pengaruh Recursive Feature Elimination Terhadap Kinerja Model Prediksi Dini Diabetes Mellitus di RS PKU Muhammadiyah Bima," vol. 7, no. 3, pp. 548–560, 2025.

[10] W. J. Ladeira, F. D. O. Santini, T. Rasul, S. Elhajjar, N. Yasin, and S. Akhtar, "Big data analytics and the use of artificial intelligence in the services industry : a meta- analysis," vol. 2069, 2024, doi: 10.1080/02642069.2024.2374990.

[11] J. Lucas, J. M. Alvarez, S. Fidler, and M. T. Law, "Optimizing Data Collection for Machine Learning," pp. 1–28.

[12] M. A. Palomino, "applied sciences Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," 2022.

[13] R. Roshan, I. A. Bhacho, S. Zai, and I. Ahmad, "Comparative Analysis of TF – IDF and Hashing Vectorizer for Fake News Detection in Sindhi : A Machine Learning and Deep Learning Approach †," 2023.

[14] A. M. Priyatno, T. Widiyaningtyas, I. Engineering, and U. N. Malang, "A Systematic Literature Review : Recursive Feature," vol. 9, no. 2, pp. 196–207, 2024, doi: 10.33480/jitk.v9i2.5015.INTRODUCTION.

[15] R. Oktafiani, U. T. Yogyakarta, A. H. Hermawan, U. T. Yogyakarta, D. Avianto, and U. T. Yogyakarta, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning Pengaruh _ Komposisi _ Split _ Data _ Terhadap _ Performa _ Klasifikasi Penyakit _ Kanker _ Payudara _ Menggunakan Algoritma _ Machine _ Learning," no. August, 2023, doi: 10.34128/jsi.v9i1.622.

[16] V. R. Joseph, "Optimal ratio for data splitting," no. April, 2022, doi: 10.1002/sam.11583.

[17] A. Purnomo, A. G. Putrada, and R. Habibi, "MDI and PI XGBoost regression-based methods : regional best pricing prediction for logistics services," vol. 22, no. 5, 2024, doi: 10.12928/TELKOMNIKA.v22i5.26037.

[18] M. Alam, I. M. Rafiqul, M. M. H. Sizan, and A. A. Akash, "The Integration of Machine Learning in information technologies: Future Trends and predictions," pp. 75–84, doi: 10.32996/jcsts.

[19] K. He, Q. Yang, L. Ji, J. Pan, and Y. Zou, "Financial Time Series Forecasting with the Deep Learning Ensemble Model," pp. 1–15, 2023.

[20] C. Yang, J. Shao, Y. Liu, P. Jou, and Y. Lin, "Application of Fuzzy-Based Support Vector Regression to Forecast of International Airport Freight Volumes," pp. 1–18, 2022.

[21] E. Rizkiani and D. Brahma, "Perbandingan Performa Algoritma Metode Bagging dan Boosting pada Prediksi Konsentrasi PM 10 di Jakarta Utara," vol. 01, pp. 74–81, 2024.

[22] M. A. Sembiring, F. W. Sembiring, and S. Informasi, "Analisa Kinerja Model Regresi Dalam Machine Learning," vol. 8, no. 1, pp. 144–152, 2024.

[23] P. Karunia, D. R. Cahyaningrum, and E. Apriani, "Entrepreneurship and Tourism Digital Marketing ' s Big Potential on Tiktok Platform : Users , Audiences , And Social Media Competitions Journal of Management , Entrepreneurship and Tourism," vol. 3, no. 2, pp. 211–221, 2025.

[24] D. Rosalia, "Strategi Komunikasi Digital dalam Pemasaran: Studi Peran Media Sosial dalam Keputusan Pembelian Konsumen," vol. 9, no. 2, pp. 488–499, 2025.

[25] D. Ardiyanto, "Komunikasi Promosi Di Era Digitalisasi," vol. 4, pp. 4414–4421, 2025.