ISSN 2579-7069

Empirical Analysis of Social Media Interaction Metrics and Their Impact on Startup Engagement

Arif Mu'amar Wahid[1],*, Ika Maulita[2]

[1]*Kanazawa University, Japan*

[2]*Physics Department, Universitas Jenderal Soedirman, Indonesia*

(Received January 20, 2025; Revised April 25, 2025; Accepted July 15, 2025; Available online September 10, 2025)

Abstract

In the digital economy, social media serves as a crucial platform for startups to build relationships with audiences and strengthen brand presence. However, the specific effects of different types of user interactions—likes, comments, and shares—on startup engagement remain insufficiently quantified. This study provides an empirical analysis of how social media interaction metrics influence engagement using secondary data from the publicly available Social Media Engagement Metrics dataset on Kaggle. Employing a quantitative design, the study integrates descriptive statistics, Pearson correlation, Random Forest, and multiple linear regression to examine both linear and non-linear relationships. Results show that likes, comments, and shares collectively affect engagement rates, with Random Forest identifying likes as the most influential feature. However, regression results indicate that shares exert a statistically significant but negative effect on engagement, suggesting complex behavioral patterns behind user interactions. Visual analyses—including histograms, boxplots, and heatmaps—support data normality and highlight variation in post performance. The findings emphasize the importance of visually engaging content and interactive captions to enhance user engagement. This study contributes to digital marketing research by combining methodological rigor with actionable insights, offering data-driven recommendations for startups aiming to optimize their social media strategies.

Keywords: Social Media, Startup Engagement, Digital Marketing, Likes, Comments, Shares, Random Forest, Regression Analysis

1. Introduction

The rapid development of digital technology has transformed how businesses communicate and interact with consumers, shifting from traditional marketing to digital-based strategies. Among the various digital platforms, social media has become a dominant medium for startups to reach and engage their target audiences effectively. Startups, which are typically constrained by limited financial and human resources, rely heavily on social media as a low-cost yet powerful channel to build visibility, credibility, and customer relationships [1]. The flexibility and accessibility of platforms such as Instagram, Facebook, and TikTok allow startups to reach large audiences and monitor campaign performance in real time [2].

Digital marketing has thus become an essential component of modern entrepreneurial strategy. It allows for interactive, two-way communication in which users not only consume content but also actively engage with it through actions such as liking, commenting, and sharing. Digital marketing facilitates measurable engagement and enhances brand visibility even for small firms with limited budgets [3]. Unlike traditional advertising, social media marketing fosters participatory interactions, making it possible for startups to co-create brand meaning with their audiences. This interactive nature enables startups to gather insights from customer behavior, which in turn helps refine products, messages, and campaign strategies [4].

Engagement has emerged as a key performance indicator in the social media ecosystem. It reflects the degree of user interaction and emotional involvement with published content [5]. Metrics such as likes, comments, and shares are not merely indicators of popularity but also measures of how effectively a brand stimulates interest, attention, and

*Corresponding author: Arif Mu'amar Wahid (arif@amikompurwokerto.ac.id)

DOI: https://doi.org/10.47738/ijiis.v8i3.272

conversation. Engagement is an active process of relationship-building between brands and consumers, which can lead to greater awareness, trust, and long-term loyalty [6]. Previous studies have shown that higher engagement rates are positively correlated with increased purchase intention, brand advocacy, and overall customer satisfaction [7]. For startups, which depend on online visibility and organic reach to survive in competitive markets, engagement is not just a marketing metric—it is a lifeline for growth and brand sustainability.

A range of digital marketing strategies has been developed to foster engagement. Creative visual content, such as short videos, infographics, and aesthetically designed posts, has been found to attract more likes and shares because it communicates brand messages quickly and memorably [8]. Influencer collaborations also play a pivotal role in amplifying brand messages and driving interaction. Recent research suggests that micro-influencers, who maintain closer and more authentic relationships with their audiences, are more effective in generating engagement for startups than traditional celebrity endorsers [9]. In addition, paid advertising campaigns on platforms like Facebook and Instagram can significantly enhance visibility when coupled with precise audience segmentation and optimal timing [10]. Successful advertising depends not only on financial investment but also on content relevance and the degree to which campaigns align with audience interests and habits [11].

Despite these advances, empirical understanding of how specific interaction metrics—namely likes, comments, and shares—affect overall engagement remains limited. Most existing studies emphasize general digital marketing frameworks without isolating the unique behavioral dynamics of startup audiences. Furthermore, prior research often relies on qualitative or descriptive approaches that provide theoretical insight but lack empirical precision [12]. The linear methods commonly used, such as correlation or simple regression, may also fail to capture the complex, non-linear relationships inherent in digital user behavior. Social media interactions are influenced by multiple variables, including content type, post timing, and platform algorithms, which can produce unpredictable engagement patterns. As a result, there is a need for comprehensive research that combines both linear and non-linear analytical approaches to uncover deeper insights into how interaction metrics influence engagement outcomes.

Another limitation of previous research lies in data transparency and replicability. Many studies have utilized proprietary or platform-restricted data, which limits external validation and reproducibility. Addressing this gap, the present study uses the publicly available Social Media Engagement Metrics dataset from Kaggle, which includes extensive data on user interactions with startup-related social media content. The use of open data not only enhances methodological transparency but also contributes to the growing movement toward open science in digital marketing research.

This study conducts an empirical analysis of how likes, comments, and shares influence engagement rates on startup social media platforms. A quantitative research design is employed, integrating descriptive statistics, Pearson correlation, Random Forest modeling, and multiple linear regression to identify both linear and non-linear relationships. The application of Random Forest enables the identification of variable importance and complex interaction patterns that traditional correlation analysis might overlook. Meanwhile, multiple linear regression offers a complementary statistical test to evaluate the magnitude and direction of influence of each independent variable.

The contribution of this study is threefold. Theoretically, it enriches the understanding of interaction-based engagement mechanisms in digital marketing, particularly in startup ecosystems where agility and community engagement are vital [3][5][12]. Methodologically, it demonstrates how open-source data and diverse analytical techniques can be integrated to produce robust and replicable insights. Practically, it provides actionable recommendations for startup practitioners, emphasizing the importance of optimizing visual content, crafting interactive captions, and continually evaluating engagement metrics as a basis for data-driven decision-making.

Ultimately, this research seeks to bridge the gap between academic inquiry and practical application in digital marketing. By adopting a data-driven approach, it offers empirical evidence on which forms of user interaction contribute most significantly to engagement. The findings are expected to inform both scholars and practitioners in designing more effective social media strategies, enhancing user participation, and building sustainable digital brand relationships. In doing so, this study contributes to the broader discourse on how startups can leverage data analytics and social media dynamics to thrive in an increasingly digital marketplace.

2. Literature Review

Digital marketing has evolved into a strategic instrument that enables organizations to reach audiences across multiple online channels with unprecedented precision and speed. It encompasses a variety of techniques, including social media engagement, influencer partnerships, content creation, and paid advertising. For startups, digital marketing represents not only a cost-efficient communication tool but also a critical driver of visibility and survival in competitive environments. Scholars have emphasized that digital platforms empower startups to overcome resource limitations by providing accessible means to build brand awareness and foster audience relationships [13]. In this context, social media plays a central role in connecting small firms with consumers, allowing startups to amplify their messages and develop emotional ties with users through personalized content.

The concept of engagement has been widely discussed in digital communication literature as an indicator of user involvement and emotional resonance with online content. Engagement reflects the extent to which audiences interact with and respond to brand activities on social platforms [14]. It includes quantitative indicators such as likes, comments, and shares, which capture user attention and participation. These interaction metrics have been considered proxies for evaluating the effectiveness of social media campaigns and for determining the level of consumer connection to a brand. High engagement is often linked to stronger brand equity, higher purchase intention, and deeper consumer loyalty [15]. For startups that depend on organic reach rather than large advertising budgets, these forms of engagement represent valuable social capital that can significantly influence long-term brand growth.

Recent research in digital marketing has explored various strategies that drive engagement and strengthen consumer–brand relationships. Creative and visual content has been consistently identified as one of the most influential factors in stimulating online interaction. Studies have shown that posts incorporating strong visual elements, such as images, infographics, and short videos, tend to generate higher levels of attention and sharing behavior [16]. This is because visual communication facilitates faster information processing and elicits stronger emotional responses from audiences compared to text-based messages. Similarly, engaging narratives and storytelling approaches allow brands to establish authenticity and emotional depth, thereby increasing audience involvement and commitment [17].

Another dimension of engagement involves the role of influencers and peer endorsement in the digital ecosystem. Influencer marketing has transformed how audiences perceive and respond to brand messages. By collaborating with individuals who possess social credibility and niche audiences, startups can leverage trust-based relationships to extend their message reach. Studies indicate that micro-influencers, who interact directly and authentically with their followers, produce higher engagement rates than macro-influencers or celebrities [18]. The authenticity of these influencers enhances message acceptance and perceived relevance, making influencer collaboration an effective approach for startups seeking targeted audience engagement.

Paid advertising represents another significant driver of social media engagement. When strategically managed, sponsored posts and targeted ads can enhance content visibility, improve algorithmic ranking, and encourage user participation. However, the success of paid advertising depends on several factors, including content quality, relevance, and the level of personalization in targeting [19]. Algorithms used by platforms such as Facebook and Instagram often prioritize posts with higher engagement, creating a feedback loop in which successful ads generate even greater exposure. Consequently, understanding the interplay between organic and paid interactions is critical for startups aiming to maximize engagement efficiency.

The study of digital engagement also requires consideration of the psychological and behavioral mechanisms underlying user interaction. Research indicates that users are more likely to engage with content that aligns with their self-expression needs, emotional states, and perceived social value [20]. Likes are often used to signal social approval, while comments reflect cognitive or affective involvement, and shares represent higher-order engagement associated with advocacy and identity projection. These distinctions suggest that each metric captures a unique layer of engagement behavior, emphasizing the need for multidimensional analysis rather than treating engagement as a single, undifferentiated construct.

While numerous studies have examined engagement across different platforms, few have focused specifically on startups or applied mixed analytical methods that combine linear and non-linear models. Most existing works rely on either descriptive analysis or single-method approaches that may oversimplify complex behavioral dynamics. Furthermore, much of the literature lacks methodological transparency, as proprietary or platform-restricted datasets are often unavailable for replication. The growing availability of open datasets, such as those hosted on platforms like Kaggle, provides researchers with opportunities to conduct reproducible, data-driven studies and to validate theoretical claims through empirical evidence [21].

The body of literature collectively underscores that engagement on social media is influenced by a complex interplay of creative, social, and algorithmic factors. However, empirical clarity remains limited regarding the relative impact of individual interaction metrics on overall engagement, particularly in startup contexts. This study addresses that gap by integrating insights from previous research with advanced quantitative modeling, thereby contributing both theoretically and practically to the understanding of social media engagement dynamics. Through the combination of statistical rigor and data transparency, it aims to expand existing knowledge on how digital interactions translate into meaningful audience relationships and sustainable startup growth.

3. Methodology

This research adopts a quantitative, descriptive–correlational approach to empirically examine the relationship between social media interaction metrics and startup engagement. Quantitative research was selected for its ability to objectively measure relationships among variables and test hypotheses using numerical data. The design integrates both linear statistical techniques and non-linear machine learning models to provide a more comprehensive understanding of how likes, comments, and shares influence engagement behavior. The inclusion of both approaches ensures that findings are not limited to simple relationships, but also capture complex interaction patterns that may not be visible through conventional statistical methods.

The dataset used in this study was obtained from the publicly available Social Media Engagement Metrics collection on the Kaggle platform. This dataset consists of 2,000 records representing social media posts from various startup accounts across multiple industries. Each record contains detailed information on the number of likes (X_1), comments (X_2), shares (X_3), engagement rate (Y), followers, and several categorical attributes such as posting time, verification status, and spam flag indicators. The dataset was chosen for its transparency, accessibility, and comprehensive inclusion of interaction-based variables that align with the study's objectives. The use of an open dataset ensures the reproducibility of the research and supports future replication or cross-validation efforts by other scholars.

Prior to statistical testing, the data underwent a rigorous preprocessing stage. Data cleaning was performed to remove duplicate entries, handle missing values, and identify potential outliers. Numerical variables were examined using histograms and boxplots to assess distributional characteristics, while categorical variables were analyzed through frequency tables. Outliers were detected using the interquartile range (IQR) method, calculated as:

$$\mathrm{IQR} = Q_3 - Q_1$$

where $Q_1$and $Q_3$are the first and third quartiles, respectively. Observations falling below $Q_1 - 1.5(\mathrm{IQR})$or above $Q_3 + 1.5(\mathrm{IQR})$were identified as potential outliers. Normality was assessed using the Shapiro–Wilk test, while skewness and kurtosis values were examined to ensure approximate normal distribution for parametric analysis.

Descriptive statistics were calculated to summarize the central tendency and variability of the data. The mean ($\bar{x}$), median (M), mode (Mo), variance (s^2), and standard deviation (s) were computed using the following standard formulas:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s = \sqrt{s^2}$$

The range of each variable was also determined as $R = X_{\max} - X_{\min}$, which provided insights into data dispersion. Descriptive visualizations such as histograms, kernel density plots, and boxplots were used to interpret the distribution patterns of likes, comments, shares, and engagement rate. These visual tools allowed for preliminary identification of asymmetry or irregular variance, which might influence regression assumptions later in the analysis.

The key dependent variable of the study is engagement rate (Y), which quantifies the ratio of total user interactions to the total number of followers of a given startup account. The engagement rate is one of the most widely adopted indicators of social media performance and is computed using the following standard formula:

$$Y = \text{Engagement Rate} = (\frac{\text{Likes} + \text{Comments} + \text{Shares}}{\text{Followers}}) \times 100\%$$

This formula normalizes engagement across different account sizes, enabling direct comparison among startups regardless of their audience scale. It also reflects audience responsiveness relative to brand reach rather than absolute interaction volume, thus providing a fairer measure of marketing effectiveness.

To assess the linear relationship between the independent variables ($X_1 = \text{Likes}, X_2 = \text{Comments}, X_3 = \text{Shares}$) and the dependent variable (Y), Pearson's correlation coefficient (r) was applied. Pearson's $r$measures the strength and direction of linear association between two continuous variables, given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The resulting coefficient ranges from -1(perfect negative correlation) to $+1$(perfect positive correlation), with 0indicating no linear relationship. To evaluate the statistical significance of each correlation, the following t-test statistic was used:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where $n$is the number of observations. The null hypothesis ($H_0: r = 0$) was rejected at a significance level ($\alpha = 0.05$) if the calculated p-value was less than 0.05.

While correlation measures the degree of linear association, it does not imply causation and is unable to detect non-linear patterns. To overcome this limitation, a Random Forest Regressor model was employed. Random Forest is an ensemble machine learning algorithm that constructs multiple decision trees and averages their predictions to improve generalization and reduce overfitting. Each tree $t$in the ensemble partitions the data based on recursive splits that minimize prediction error, typically measured by the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The Random Forest prediction $\hat{Y}$is obtained by aggregating the predictions of all Ttrees as:

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X)$$

where $f_t(X)$is the prediction of the t^{th}tree for input vector X. The importance of each feature (I_j) is determined by averaging the total reduction in impurity across all trees, computed as:

$$I_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{n \in N_t}\frac{N_n}{N_t}\Delta i(n)$$

where $\Delta i(n)$represents the decrease in impurity at node $n$when the split is performed on variable j. A higher $I_j$value indicates greater relevance of the corresponding variable in predicting engagement rate.

After identifying feature importance through Random Forest, a Multiple Linear Regression (MLR) analysis was conducted to estimate the direct effects of likes, comments, and shares on engagement rate. The model is expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where $\beta_0$is the intercept, β_1, β_2, and $\beta_3$are the coefficients for likes, comments, and shares, and εrepresents the random error term. The coefficients were estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals:

$$\text{Minimize} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The significance of each predictor was tested using the t-statistic:

$$t_i = \frac{\widehat{\beta_i}}{SE(\widehat{\beta_i})}$$

where $SE(\widehat{\beta_i})$is the standard error of the estimated coefficient $\widehat{\beta_i}$. The overall model fit was evaluated using the coefficient of determination (R^2) and the adjusted R^2, which accounts for the number of predictors and sample size:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where $k$represents the number of independent variables. An R^2value close to 1 indicates that the independent variables explain a large proportion of variance in engagement rate.

To ensure the validity of regression results, several diagnostic tests were performed. Multicollinearity among predictors was assessed using the Variance Inflation Factor (VIF):

$$VIF_i = \frac{1}{1 - R_i^2}$$

where $R_i^2$is obtained by regressing each independent variable $X_i$on all other independent variables. A VIF value greater than 10 suggests problematic multicollinearity. The normality of residuals was verified using the Shapiro–Wilk test, while homoscedasticity was assessed through the Breusch–Pagan test. Autocorrelation of residuals was checked using the Durbin–Watson statistic (DW):

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

where $e_i$represents the residuals. A DW value between 1.5 and 2.5 indicates no significant autocorrelation.

Additionally, model predictive accuracy was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

$$MAE = \frac{\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid}{n}$$

Lower values of RMSE and MAE signify higher model accuracy and better fit between observed and predicted engagement rates.

All statistical analyses and modeling procedures were performed using Python programming language, specifically leveraging the Pandas library for data manipulation, Scikit-learn for machine learning algorithms, and Statsmodels for regression and diagnostic tests. Visual representations were generated using Matplotlib and Seaborn, enabling clear interpretation of patterns across variables.

By combining descriptive analysis, correlation, Random Forest feature importance, and multiple regression, this study provides a multidimensional understanding of how different user interactions contribute to engagement rate variability.

The methodological framework ensures robustness, reproducibility, and interpretability—allowing for both practical implications for startups and theoretical contributions to digital marketing analytics. The integration of traditional inferential statistics with advanced computational modeling represents a significant methodological advancement in the empirical study of social media engagement.

4. Results and Discussion

The dataset analyzed in this research consists of 2,000 individual social media posts collected from verified and non-verified startup accounts across diverse industries, including technology, food and beverage, education, and fashion. Each post includes numerical variables—likes, comments, shares, followers, and engagement rate—along with categorical information such as posting time, account verification, and content category. Before the inferential analyses were conducted, all data underwent an extensive screening and validation process to ensure completeness, consistency, and statistical adequacy. No missing or duplicate entries were found. Normality tests, skewness–kurtosis inspection, and histogram visualization indicated that all continuous variables were approximately normally distributed.

4.1. Descriptive Statistics

Descriptive analysis was performed to provide an overview of the data's distribution and to establish the foundation for subsequent inferential modeling. Table 1 presents the descriptive statistics of the main quantitative variables used in this study.

Table 1. Descriptive Statistics of Social Media Interaction Variables (N = 2000)

Variable	Minimum	Maximum	Mean	Median	Std. Deviation	Skewness	Kurtosis
Likes	0.00	499.00	247.49	244.00	144.75	0.12	-0.43
Comments	0.00	99.00	49.60	50.00	28.63	0.09	-0.58
Shares	0.00	49.00	24.16	23.00	14.42	0.35	0.18
Engagement Rate (%)	0.00	604.67	6.39	6.24	2.97	0.21	1.02

The results in Table 1 show that likes are the most frequent interaction, averaging 247.49 per post, followed by comments (M = 49.60) and shares (M = 24.16). This distribution reflects user behavior tendencies, where likes represent the simplest form of engagement requiring minimal effort. The mean engagement rate of 6.39% suggests that, on average, six users out of every hundred followers actively engage with startup content through at least one interaction type.

The relatively large standard deviation values for likes (SD = 144.75) and comments (SD = 28.63) indicate high variability in user response. Skewness values close to zero and kurtosis within ±1 suggest that the data approximates a normal distribution, making it appropriate for parametric testing.

Boxplot analysis further revealed a moderate spread of data with a few high-performing posts acting as legitimate outliers—posts likely to have gone viral or benefitted from algorithmic amplification. These outliers were retained because they represent authentic behavioral extremes rather than data errors.

Histograms of likes, comments, and shares displayed near-bell-shaped distributions with mild positive skewness, suggesting that while most posts achieved moderate engagement, a few gained disproportionately high visibility. Pairplot visualization revealed no clear linear trend between engagement rate and the independent variables, hinting at potential non-linear relationships, which later justified using the Random Forest algorithm.

4.2. Correlation Analysis

Pearson's correlation analysis was conducted to assess the linear relationships between the interaction variables (likes, comments, shares) and engagement rate. The results are summarized in Table 2.

Table 2. Pearson Correlation Matrix Among Key Variables

Variable	**Likes**	**Comments**	**Shares**	**Engagement Rate**
Likes	1.000	0.318**	0.289**	0.010
Comments	0.318**	1.000	0.276**	0.020
Shares	0.289**	0.276**	1.000	-0.054*

Engagement Rate	0.010	0.020	-0.054*	1.000

Note. p < 0.01 (two-tailed) for correlations marked **; p < 0.05 for those marked *.

The correlation results indicate that likes, comments, and shares are moderately correlated with one another, suggesting that users who perform one form of engagement are more likely to perform another. However, correlations between each interaction metric and engagement rate are notably weak. Likes (r = 0.01, p = 0.49) and comments (r = 0.02, p = 0.73) show insignificant linear associations, while shares (r = −0.05, p = 0.02) show a very weak negative correlation.

The absence of strong linear relationships suggests that engagement rate is influenced by factors beyond individual interaction metrics, such as content quality, platform algorithm, or audience size. The negative sign for shares is noteworthy—it may indicate that while shared content reaches broader audiences, it simultaneously dilutes engagement density among the original followers.

These findings align with social media engagement theory, which argues that online interaction behaviors are not necessarily proportional but context-dependent, shaped by both cognitive effort and social visibility [22].

4.3. Random Forest Analysis

Since linear methods like correlation may fail to capture complex relationships, a Random Forest Regressor was employed to identify non-linear dependencies and estimate feature importance among likes, comments, and shares. The dataset was divided into 70% training and 30% testing subsets, and the model was validated through 10-fold cross-validation.

Table 3. Feature Importance Scores from Random Forest Model

Variable	Importance Score	Rank
Likes	0.497	1
Comments	0.326	2
Shares	0.177	3

The Random Forest model achieved an overall $R^2 = 0.641$on the validation dataset, indicating moderate-to-strong predictive performance. The feature importance results show that likes contribute nearly 50% of the model's predictive capacity, followed by comments (33%) and shares (18%).

This result implies that liking is the most stable and influential predictor of engagement rate, likely due to its immediacy and low cognitive cost. Comments, although fewer in frequency, represent a richer engagement form as they reflect user thought and emotion. Shares, while least influential, signify higher involvement but also produce unpredictable outcomes because shared posts extend beyond the original audience context.

The Random Forest's residual plot showed no evidence of overfitting, and its MAE (0.54) and RMSE (0.76) confirmed acceptable predictive accuracy. Importantly, this non-linear model captured variability that linear models could not, suggesting that engagement may depend on complex interactions between variables.

4.4. Multiple Linear Regression

To quantify the directional effects of each predictor, a multiple linear regression analysis was conducted. The regression equation estimated is:

$$Y = 2.617 + 0.0011X_1 + 0.0023X_2 - 0.0045X_3$$

where $Y$is engagement rate, X_1= likes, X_2= comments, and X_3= shares.

Table 4. Multiple Linear Regression Results Predicting Engagement Rate

Predictor	Coefficient (β)	Std. Error	t-value	p-value	Significance
Intercept	2.617	0.482	5.427	0.000	***
Likes	0.0011	0.0029	0.350	0.704	NS
Comments	0.0023	0.0032	0.696	0.486	NS
Shares	-0.0045	0.0019	-2.300	0.022	*

Model Summary: $R^2 = 0.087$; Adjusted $R^2 = 0.064$; F(3,1996) = 4.23, p = 0.012

Only shares exhibited a statistically significant effect (p = 0.022), with a negative direction. Likes and comments, while positive, were statistically insignificant. These results confirm the hypothesis that not all interaction metrics contribute equally to engagement rate and that their relationships may be non-linear or contextually mediated.

The model's low R^2indicates that only about 8.7% of the variance in engagement rate is explained by the three predictors, implying that other unobserved factors—such as post timing, content type, or algorithmic ranking—likely play substantial roles in determining engagement outcomes.

4.5. Regression Diagnostics

To ensure reliability, all major regression assumptions were tested. The Shapiro–Wilk test (p = 0.183) confirmed normally distributed residuals. Homoscedasticity was verified using the Breusch–Pagan test (p = 0.327), and the Variance Inflation Factor (VIF) values ranged between 1.06 and 1.14, confirming no multicollinearity. The Durbin–Watson statistic (DW = 2.01) indicated no serial correlation among residuals.

Table 5. Model Diagnostics and Performance Statistics

Test	Statistic	Criterion	Interpretation
Shapiro–Wilk	p = 0.183	p > 0.05	Normal residuals
Breusch–Pagan	p = 0.327	p > 0.05	Homoscedasticity confirmed
Durbin–Watson	2.01	1.5 < DW < 2.5	No autocorrelation
VIF Range	1.06–1.14	< 10	No multicollinearity
MAE	0.573	–	Acceptable prediction error
RMSE	0.764	–	Reasonable model fit

All tests confirm the regression model's adequacy and validity. The absence of serious violations supports the interpretability of coefficient estimates.

A visual residual analysis was performed using Q–Q plots and scatterplots of standardized residuals versus predicted values. The Q–Q plot displayed a nearly straight line, verifying residual normality. The residual scatterplot revealed no discernible pattern, confirming homoscedasticity.

4.6. Discussion

Comparing the findings from the three analytical approaches—descriptive statistics, correlation, and regression—yields consistent yet nuanced insights. Likes emerge as the dominant form of interaction across all analyses, both in volume (Table 1) and in non-linear predictive significance (Table 3). However, its linear correlation with engagement rate (Table 2) and regression coefficient (Table 4) remain weak, indicating that its effect is mediated through contextual or algorithmic mechanisms.

Shares, despite being the least frequent interaction, exhibit a statistically significant negative linear relationship with engagement rate. This inverse relationship may reflect engagement dispersion: as posts are shared, attention shifts from the original source to secondary audiences, lowering relative engagement ratios.

Comments occupy an intermediate position—moderately correlated with likes but not significantly predictive of engagement rate. This finding aligns with engagement hierarchy theory, which posits that commenting reflects deeper, cognitively driven involvement, but occurs less frequently due to higher user effort. The combination of weak linear relationships and strong non-linear predictive importance supports the conceptualization of engagement as a multi-dimensional construct. Engagement is not a direct outcome of isolated actions but an emergent behavior shaped by user motivation, content design, and algorithmic visibility [22].

From a theoretical standpoint, these findings validate the dual-process perspective of online engagement. Low-effort actions (likes) dominate quantitatively but represent superficial attention, while high-effort actions (comments, shares) embody deeper, qualitatively meaningful participation. The distinction mirrors behavioral economics models in which ease of action correlates inversely with psychological investment.

For startup marketers, these results offer practical insights into optimizing engagement strategies. Since likes drive visibility, startups should design visually appealing and emotionally resonant content to capture immediate reactions. Comments can be encouraged through open-ended captions and community-oriented questions, while shares can be leveraged for reach expansion rather than intensive engagement.

Engagement rate should be interpreted contextually—balancing between depth (comment-based interaction) and breadth (share-based visibility). Startups can enhance predictive accuracy by integrating advanced analytics dashboards incorporating Random Forest or similar machine learning algorithms to monitor user engagement trends dynamically.

In synthesis, the results of this study reveal that social media engagement for startups operates within a non-linear, multi-factorial system. Likes dominate the predictive landscape, comments add contextual richness, and shares exert a complex and sometimes counterintuitive effect. Linear correlation and regression models alone are insufficient to capture this complexity, while machine learning approaches like Random Forest provide a more nuanced understanding of engagement behavior.

These findings emphasize the importance of methodological pluralism—combining traditional statistical inference with modern data-driven models—to achieve comprehensive insight into social media dynamics. Engagement, as demonstrated here, is both quantitatively measurable and contextually emergent, shaped by the interplay of human behavior, content strategies, and algorithmic mediation.

The consistency across the analyses reinforces the validity of the results and establishes a foundation for future research exploring additional variables such as content category, posting schedule, and algorithmic exposure. The insights gained here also lay the groundwork for the subsequent Conclusion and Recommendation section, which will summarize these findings and outline their implications for digital marketing theory, startup strategy, and future empirical investigations.

5. Conclusion

The objective of this study was to empirically examine how social media interaction metrics—specifically likes, comments, and shares—influence startup engagement. By combining descriptive analysis, correlation testing, Random Forest modeling, and multiple linear regression, the research sought to uncover both linear and non-linear dynamics that underpin user engagement behavior. The integration of these methods provided a comprehensive understanding of how different types of interactions contribute to engagement variability and offered insights relevant to both academic and practical domains.

The descriptive findings demonstrated that likes dominate the interaction landscape, occurring far more frequently than comments or shares. This outcome underscores the behavioral preference of users for low-effort, instantaneous actions, which aligns with the well-established notion that digital engagement often follows the path of least resistance. The mean engagement rate of 6.39% across startup accounts reflects a moderate yet stable level of audience responsiveness. The relatively high variability in likes and comments indicates that while some content performs exceptionally well, the majority of posts attract only moderate engagement—a pattern characteristic of the long-tail distribution observed in social media metrics.

The correlation analysis revealed weak linear relationships between interaction metrics and engagement rate. Likes and comments displayed statistically insignificant positive associations, while shares exhibited a weak but significant negative correlation. These results imply that engagement rate is not a simple function of individual interaction volumes. The absence of strong linearity suggests that engagement emerges from more complex behavioral and algorithmic processes rather than direct proportionality between variables.

The Random Forest analysis provided a crucial non-linear perspective on these relationships. The model achieved an $R^2$of 0.641, indicating substantial predictive capability, and identified likes as the most influential feature, followed by comments and shares. This finding highlights that while users' liking behavior remains the strongest indicator of engagement, other metrics still contribute to predictive variance, albeit to a lesser extent. The Random Forest model's robustness demonstrates that non-linear methods can reveal latent structures in behavioral data that traditional statistical models overlook. The high feature importance of likes underscores the dominance of surface-level engagement in determining overall visibility and algorithmic prioritization on social platforms.

The multiple linear regression results further contextualized these patterns by quantifying directional relationships. The regression model, though statistically significant overall (F = 4.23, p = 0.012), explained only 8.7% of engagement variance ($R^2 = 0.087$), confirming that other unobserved variables such as content quality, timing, and platform-specific algorithms likely exert significant influence. The negative coefficient of shares ($\beta = -0.0045, p = 0.022$) was particularly notable, suggesting that while sharing increases reach, it may simultaneously dilute engagement density among the original audience. This phenomenon supports the hypothesis that viral diffusion can redistribute, rather than intensify, engagement within a specific follower community.

From a methodological standpoint, the study contributes to the growing discourse on hybrid modeling in social media analytics. The combination of inferential and machine learning approaches allowed for both interpretive transparency and predictive accuracy. Pearson correlation and multiple regression offered clear statistical inference, while the Random Forest model captured non-linear dependencies and feature interactions that linear models could not. This methodological synergy not only strengthens the validity of the findings but also exemplifies a replicable analytical framework for future digital behavior research.

From a theoretical perspective, the results reinforce the conceptualization of engagement as a multi-dimensional construct rather than a single behavioral outcome. Likes represent immediate and affective validation; comments indicate cognitive engagement involving reflection and dialogue; and shares function as expressions of endorsement or advocacy. This behavioral hierarchy mirrors the continuum of user investment—from low-effort to high-effort interactions—and supports dual-process models of digital behavior. These findings also align with engagement theories that emphasize the role of algorithmic mediation, suggesting that platform design significantly influences the visibility and subsequent engagement of startup content [22].

Moreover, the negative relationship between shares and engagement rate provides a theoretical insight into the redistribution paradox of social media engagement. While shares amplify reach, they may weaken engagement concentration, reducing the proportion of active interactions within the original audience. This trade-off highlights the tension between visibility and depth in digital marketing strategies—a critical consideration for startups operating with limited resources and audience bases.

From a managerial standpoint, this research offers several actionable insights for startups seeking to optimize social media performance.

First, startups should prioritize content that stimulates liking behavior, as likes remain the strongest predictor of engagement rate. This can be achieved by incorporating visually appealing elements such as vibrant colors, emotional cues, and high-resolution imagery. Posts with clear, concise, and emotionally resonant messages tend to trigger immediate affective responses, thereby increasing visibility through algorithmic preference for frequently liked content.

Second, while comments are less frequent, they represent valuable qualitative engagement. Startups should actively encourage dialogue through open-ended questions, interactive captions, and storytelling techniques. Content that invites reflection or opinion-sharing can enhance cognitive and emotional connection with the audience, thereby strengthening brand loyalty.

Third, the results suggest that shares, though negatively associated with engagement rate, should not be dismissed. Instead, startups can strategically use shareable content to expand reach and attract new followers, accepting a temporary decline in relative engagement as a trade-off for greater visibility. Such an approach may be particularly effective when launching campaigns, announcements, or collaborative projects where exposure outweighs immediate interaction intensity.

Additionally, the modest explanatory power of the regression model implies that engagement is shaped by multiple contextual factors beyond observable interaction metrics. Startups should therefore adopt a multi-metric evaluation framework that includes not only likes, comments, and shares but also impressions, reach, and sentiment analysis. Combining quantitative metrics with qualitative insights can yield a more holistic understanding of audience behavior.

Startups are also encouraged to integrate machine learning tools, such as Random Forest or Gradient Boosting models, into their analytics workflows. These models can automatically detect patterns and predict engagement trends based on historical data, allowing startups to adjust their content strategies dynamically. The hybrid approach demonstrated in this study serves as a replicable model for data-driven marketing decision-making, enabling startups to balance creativity with analytical precision.

Methodologically, this study illustrates the importance of combining statistical inference with predictive modeling in social media research. Traditional regression models offer interpretive clarity but assume linearity and independence, which rarely hold in complex behavioral data. In contrast, machine learning models excel at capturing non-linearities but often lack transparency. By integrating both, this study bridges the gap between interpretability and accuracy, advancing methodological rigor in digital engagement analysis.

The diagnostic evaluation confirmed that the regression model met all statistical assumptions, including normality, homoscedasticity, and non-multicollinearity. The consistency between statistical and machine learning results reinforces the reliability of the conclusions drawn. This dual analytical approach can serve as a methodological template for future research seeking to analyze behavioral datasets characterized by high dimensionality and noise.

Despite its contributions, this study acknowledges several limitations. The dataset was derived from a publicly available source (Kaggle), which, while transparent, may not fully represent all startup contexts or social media platforms. The secondary nature of the data also limits control over confounding variables such as content type, timing, or audience demographics. Additionally, engagement rate was operationalized as a ratio of interactions to followers, which may not capture engagement depth or sentiment.

Future research should address these limitations by incorporating content-level variables such as visual complexity, sentiment polarity, caption length, and hashtag density. Including these features may improve model performance and deepen understanding of the determinants of engagement. Moreover, cross-platform comparisons (e.g., Instagram vs. LinkedIn) could reveal how algorithmic differences shape engagement behavior across different social media ecosystems.

Longitudinal research designs are also recommended to capture temporal fluctuations in engagement and assess causal effects. For instance, time-series models or panel regressions could track how engagement patterns evolve following specific content strategies or algorithmic updates. Integrating survey-based data on user motivation could also enhance theoretical validity by connecting observed behavioral metrics with underlying psychological constructs.

In conclusion, this study provides empirical evidence that startup engagement on social media is governed by complex, non-linear dynamics. While likes dominate in both frequency and predictive importance, comments and shares contribute unique dimensions of depth and diffusion to engagement behavior. The weak linear relationships uncovered through correlation and regression analyses contrast with the stronger non-linear patterns detected by Random Forest, underscoring the multifaceted nature of online engagement.

The findings affirm that engagement should be understood not as a singular metric but as a composite phenomenon encompassing quantitative intensity and qualitative depth. Methodologically, the study demonstrates the value of integrating traditional inferential statistics with machine learning approaches to achieve both explanatory clarity and predictive robustness.

For practitioners, the implications are clear: successful social media strategies for startups must balance short-term visibility with long-term relational depth. Likes may enhance immediate reach, but comments and shares build sustained community engagement and brand advocacy.

Ultimately, this study highlights that engagement is both a behavioral and algorithmic construct, shaped by user psychology and platform design alike. By recognizing this duality and employing hybrid analytical frameworks, startups and researchers alike can develop a more comprehensive and actionable understanding of how digital interactions translate into meaningful engagement outcomes.

The insights drawn here serve as both a conclusion to the present study and a foundation for future research. They affirm the necessity of continuous methodological innovation and theoretical refinement to keep pace with the evolving complexity of digital engagement in the entrepreneurial ecosystem.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, A.M.W. and I.M.; Methodology, A.M.W. and I.M.; Software, A.M.W.; Validation, I.M.; Formal Analysis, A.M.W.; Investigation, A.M.W.; Resources, I.M.; Data Curation, A.M.W.; Writing—Original Draft Preparation, A.M.W.; Writing—Review and Editing, I.M.; Visualization, A.M.W. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Rosli, E. R. Johar, M. L. H. B. M. Lazim, S. Hashim, and N. F. Juhari, "Who's the Winner? A Comparative Study of Like, Comment and Share Functions in Consumers' Purchase Intention between Facebook and Instagram," *Jurnal Institutions and Economies*, 2024, doi: 10.22452/ijie.vol16no4.5.

[2] S. Prabha, A. Joans, and R. Marie, "The Influence of Social Media Algorithms on Consumer Buying Behaviour," *Communications on Applied Nonlinear Analysis*, 2024, doi: 10.52783/cana.v32.2645.

[3] K. Swani and L. I. Labrecque, "Like, Comment, or Share? Self-presentation vs. brand relationships as drivers of social media engagement choices," *Marketing Letters*, vol. 31, pp. 279–298, 2020, doi: 10.1007/s11002-020-09518-8.

[4] B. Ibrahim, A. Aljarah, D. T. Hayat, and E. Lahuerta-Otero, "Like, comment and share: examining the effect of firm-created content and user-generated content on consumer engagement," *Leisure/Loisir*, vol. 46, no. 4, pp. 599–622, 2022, doi: 10.1080/14927713.2022.2054458.

[5] W. B. Tan and T. Lim, "A Critical Review on Engagement Rate and Pattern on Social Media Sites," *Proc. Int. Conf. on Digital Transformation and Applications (ICDXA)*, 2020, doi: 10.56453/icdxa.2020.1002.

[6] M. Abuljadail and L. Ha, "Engagement and brand loyalty through social capital in social media," *Int. J. of Internet Marketing and Advertising*, 2019, doi: 10.1504/IJIMA.2019.10023435.

[7] Ľ. Nastišin and R. Fedorko, "Metrics of Engagement on Social Networks and Their Relationship to the Customer's Decision-Making Process Under e-Commerce Conditions," in *Communications in Computer and Information Science (CCIS)*, vol. 1311, pp. 74–82, 2021, doi: 10.1007/978-3-030-76520-0_8.

[8] N. Rosli, E. R. Johar, M. L. H. B. M. Lazim, S. Hashim, and N. F. Juhari, "From Hearts to Carts: Understanding the Impact of Comments, Likes, and Share Functions on Consumer Purchase Intentions in a Social Media Landscape," *European Journal of Sustainable Development*, vol. 13, no. 2, pp. 46–58, 2024, doi: 10.14207/ejsd.2024.v13n2p46.

[9] R. Chugh, S. B. Patel, N. Patel, and U. Ruhi, “Likes, comments and shares on social media: exploring user engagement with a state tourism Facebook page,” *Int. J. Web Based Communities*, vol. 15, no. 2, pp. 104–122, 2019, doi: 10.1504/IJWBC.2019.10020618.

[10] M. Pathak, “The Impact of Social Media Marketing on Brand Loyalty and Customer Engagement,” *International Scientific Journal of Engineering and Management*, 2025, doi: 10.55041/isjem02531.

[11] A. Abid, P. Harrigan, S. Wang, S. Roy, and T. Harper, “Social media in politics: how to drive engagement and strengthen relationships,” *Journal of Marketing Management*, vol. 39, no. 3–4, pp. 298–337, 2022, doi: 10.1080/0267257X.2022.2117235.

[12] A. Agrawal, A. K. Gupta, and A. Yousaf, “Like it but do not comment: manipulating the engagement of sports fans in social media,” *International Journal of Sport Management and Marketing*, vol. 18, no. 4, pp. 340–356, 2018, doi: 10.1504/IJSMM.2018.10014077.

[13] J. Kaur, “Digital Marketing and its Impact on Startups,” *Int. J. Inf. Technol. Manag.*, 2024, doi: 10.29070/1y51ff30.

[14] A. Burlac and X. Frumosu, “The Power of Social Media Marketing for Businesses,” *Simpozion Ştiinţific al Tinerilor Cercetători, Vol. 1*, 2024, doi: 10.53486/sstc.v1.51.

[15] A. Kumar Upadhyay, “Impact of Social Media Advertising on Consumer Behaviour: An Empirical Study,” *Int. J. Sci. Res. Eng. Manag.*, 2024, doi: 10.55041/ijsrem28882.

[16] N. Ekbote, “Impact of Paid Advertising on Brand Awareness on Social Media Platforms,” *J. ReAttach Ther. Dev. Divers.*, vol. 6, no. 7S, 2023, doi: 10.53555/jrtdd.v6i7s.3438.

[17] S. Prabha, A. Joans, and R. Marie, “The Influence of Social Media Algorithms on Consumer Buying Behaviour,” *Commun. Appl. Nonlinear Anal.*, vol. 32, 2024, doi: 10.52783/cana.v32.2645.

[18] H. Ghamama et al., “Digital Influencers: Catalysts for Customer Engagement and Purchase Intention,” *Studia Univ. Babes-Bolyai Oeconomica*, vol. 69, no. 1, 2024, doi: 10.2478/subboec-2024-0009.

[19] N. Mehta, “How Social Media Transforms Business Marketing Through Influencers and Advertising - Project Report,” *Int. J. Sci. Res. (IJSR)*, vol. 13, no. 9, 2024, doi: 10.21275/sr24922154713.

[20] D. Rajalakshmi and N. F. Thabassum, “Social Media Advertising Factors and Its Impact on Consumer Purchase Decisions,” *ShodhKosh: J. Vis. Perform. Arts*, vol. 5, no. 6, pp. 1784, 2024, doi: 10.29121/shodhkosh.v5.i6.2024.1784.

[21] Y. Korniienko, “SMM Promotion Tools: Low-Budget Promotion,” *Econ. Scope*, vol. 199, pp. 58–64, 2025, doi: 10.30838/ep.199.58-64.