Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 227 (2023) 823–832

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023)

Machine learning models to predict the engagement level of Twitter posts: Indonesian e-commerce case study

Dinda Thalia Andariesta[a]*, Meditya Wasesa[a]

[a] *School of Business and Management, Institut Teknologi Bandung, Bandung, 40132, Indonesia*

Abstract

The growing utilization of social media platforms enables direct interaction between companies and consumers. However, the expanding range of interactions and real-world data complexities necessitate the development of more sophisticated decision models. To address this, the current research focuses on constructing machine learning models, namely multinomial logistic regression, decision tree, k-nearest neighbor, and random forest, to forecast the engagement level of Twitter posts from three prominent e-commerce platforms in Indonesia: Bukalapak, Blibli, and Tokopedia. The analysis comprises a dataset of 12,786 unique tweets, accumulating 11,870,254 favorites and 2,735,886 retweets over a seven-month period from February 1 to August 31, 2021. The prediction models are built upon three theoretical constructs with seven features, encompassing interactivity (e.g., links, hashtags), vividness (e.g., images, short videos, long videos), and temporal factors (e.g., day of post, last post time). Factors such as post frequency, interactive posting elements, and static visual elements emerge as significant features for predicting the engagement level of Twitter posts. Results demonstrate that the random forest model outperforms singular classifier models, including multinomial logistic regression, decision tree, and k-nearest neighbor models, in terms of precision, recall, and F1 score.

Keywords: Twitter; social media; engagement; predictive analytics; machine learning.

* Corresponding author.
E-mail address: dinda_thalia@sbm-itb.ac.id

1877-0509
10.1016/j.procs.2023.10.588

1. Introduction

The rapid internet adoption goes hand in hand with the increasing use of social media platforms such as microblogs, social networks, blogs, forums, and photo-sharing platforms. One of the most popular social media platforms is Twitter. As of April 2021, there are approximately 396 million active users of Twitter worldwide [1]. Indonesia is the sixth-largest country in terms of active Twitter users, with an audience reach of 15.1 million users [2]. In response, marketers have seriously utilized Twitter for various purposes such as advertisement, promotion, research, and customer care [3]. Both in the academic and business worlds, the topic of how to use social media platforms for business purposes has become increasingly attractive.

The widespread use of social media has transformed the way companies reach out to and engage with their customers [4]. Companies use Twitter and other social media platforms to enhance positive relationships with their consumers. Content or posts on the social media platform encourage further interactions with consumers [5]. Through comments, likes, and shares of the original content on social media, followers (i.e., consumers) add value to the content while influencing the engagement level on the social media platform [6]. Correspondingly, many studies have focused on investigating post characteristics, such as vividness, interactivity, content styles, and temporal factors, as determinants for engagement [7]. Regardless of the potential benefits of social media marketing, how platforms should be used, how content should be designed, and how consumers will expectedly react are still open for further research [7].

Although many studies have examined the influence of post characteristics on engagement, many of the studies have used regression models. Given the high variety of interactions and numerous features available in social media interactions, developing a consumer engagement model is a complex task. Since the relationships among variables increase, the real-world phenomenon becomes harder to predict and more challenging to handle using regression analysis [3]. Therefore, novel approach such as machine learning offers high potential for contributing new knowledge on consumer behavior, where regression analysis like ordinary least squares (OLS) regression [4]–[6], [8], negative binomial regression [7], and logistic regression [3] are widely used. In response, this study focuses on developing machine learning-based models (i.e., multinomial logistic regression, decision tree, k-nearest neighbor (KNN), and random forest) to predict the engagement level on Twitter posts of three prominent Indonesia's e-commerce platforms, namely, Tokopedia, Bukalapak, and BliBli.

2. Literature Review

Customer online engagement refers to the users' psychological state characterized by interactive, co-creative user interactions with a focal agent (i.e., brand page, fan page) and object (i.e., post and content of the post) [9]. In an online environment, engagement is measured using metrics such as click-through rates, page views, or various action measures that vary across platforms [10]. For example, on Facebook, such acts involve likes, comments, shares, and reactions, each representing a different level of interaction [4]. While on Twitter, those actions may include favorites, replies, and retweets [3]. Engagement and interactions are intertwined concepts, where interactions on social media can trigger engagement [11]. Therefore, low user interactions represent a low engagement rate and weak social media marketing [3].

In the literature, there are two opposing concepts of interactivity [12]. First, interpersonal interactivity refers to continuous communication between individuals or organizations in one or two directions [13]. Second, machine interactivity [14] is characterized as the degree to which users can change the messages they receive [15]. The sharing of messages between users enables interpersonal interactivity. While machine interactivity can be obtained using links or hashtags that provide additional information when clicked on. The links can increase user engagement by providing easy access to interesting content [12]. However, clicking on the link implies navigating away from the social media page, increasing the risk of users not commenting on the post anymore [5]. On the other hand, the hashtags positively affect the number of likes and comments [6].

Another important social media post characteristic is vividness. Vividness refers to the way an environment presents information to the senses [15]. Posts on Twitter can be made in various formats, including text-only, images or photos (static visuals), and animated photos or videos (dynamic visuals). Dynamic visuals are more vivid than static visuals and text-only posts, and the use of images and videos will gain more attention than text-only posts. The previous studies found that a higher level of vividness triggers more likes [8] and higher engagement levels [7].

Apart from interactivity and vividness, temporal factors such as days when posting (weekday or weekend) and posting frequency might also influence social media engagement [8].

Table 1 summarizes previous studies on social media engagement. In terms of methods, most of them employed regression analysis such as OLS regression [4]–[6], [8], negative binomial regression [7], and logistic regression [3]. Besides regression analysis, other studies applied analysis of variance (ANOVA) to assess the significance of several post characteristics affecting user engagement [10], [16]. While those two techniques have been popular, there are limited studies using machine learning. To the best of our knowledge, only Aydin et al. (2021) employed artificial neural network (ANN) to predict engagement based on post characteristics [3].

Table 1. Literature review on the impact of social media post characteristics on engagement.

Study	Independent Variable				Dependent Variable	Social Media Platform	Method	Number of Posts	Data Period	Industry Context
	1	2	3	4						
[3]	v	v	v	v	Engagement rate	Facebook, Twitter	Logistic regression, decision tree, ANN, random forest	1,130	Four months in 2017 and 2018	Durable goods and FMCG
[4]		v		v	Engagement rate, the ratio of like, share, reaction, comment	Facebook	OLS regression	2,627	Apr - Jun 2016	Retailer
[5]	v	v		v	Number of likes and comments	Facebook	OLS regression	164	Mar - Apr 2011	Travel Agency
[6]	v	v	v	v	Number of likes, comments, shares	Facebook	OLS regression	792	Apr - May 2014	Apparel and food retailer
[7]	v	v	v	v	Engagement rate, the ratio of like, share, and comment	Facebook	Negative binomial regression	5,035	Jan - Mar 2012	FMCG
[8]	v	v	v	v	Number of likes and comments	Facebook	OLS regression	355	May 2010 - Feb 2011	Diverse
[10]	v	v	v		Number of likes, comments, shares	Facebook	ANOVA	1,030	Mar - May 2014	Diverse
[16]		v			Like ratio	Facebook	ANOVA	560	Dec 2014	Automotive
This study	v	v		v	Engagement level	Twitter	Multinomial logistic regression, decision tree, KNN, random forest	12,786	Feb - Aug 2021	e-commerce

Note: Independent variable 1 = Interactivity, 2 = Vividness, 3 = Content type, 4 = Temporal factors.

Furthermore, most studies were carried out in Europe and the US. with empirical cases from fast-moving consumer goods (FMCG) [3], [7] and retailing industry [4], [6]. Studies investigating cases from developing countries are hard to find. Correspondingly, this study chooses three leading Indonesia's e-commerce platforms as the research context. This study focuses on developing Twitter post's engagement prediction models of three leading Indonesia's e-commerce platforms using machine learning-based models, namely multinomial logistic regression, decision tree, KNN, and random forest. In terms of method, the extension of binary logistic regression allows classifying multi-categorical variables called multinomial logistic regression [17]. Next, decision tree models can handle nonlinear relationships [18]. However, outliers may affect the developed decision tree, leading to an overfitting problem [3]. Therefore, a combination of several decision trees called random forest is developed to overcome this problem [3]. Lastly, KNN is a popular distance-based algorithm because it has good classification performance and simple implementation [19].

3. Research Method

Fig. 1 shows the research framework that guides the development of the classification model. The framework consists of four steps: (1) data collection, (2) data preparation, (3) model development, and (4) model evaluation. In

the first step, we collect the Twitter post data from the respondents' e-commerce official Twitter account. In the second step, we extract the prediction models' variables. Then, we use stratified sampling to split the data into training and testing datasets. In the third step, we use the training dataset to train the machine learning models and perform 10-fold cross-validation. During the model construction, the model parameters were automatically tuned via a cross-validation algorithm. We use the area under the precision-recall curve (PR AUC) score to monitor the training performance and select the best model parameters. In the last step, the best prediction models are used to predict the testing (out-sample) datasets. Finally, this study evaluates the prediction models using three prediction accuracy metrics, including precision, recall, and F_1 score.

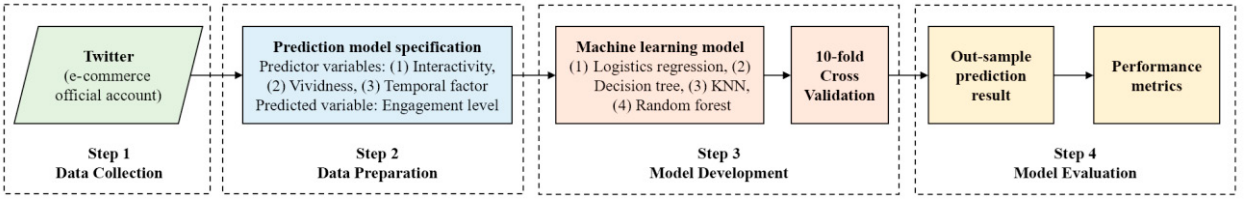

Fig. 1. Research framework.

3.1. Data Collection

In the data collection step, we mine data from three e-commerce's official Twitter account for seven months (February 1 – August 31, 2021) using Twitter API. We take evidence from three leading e-commerce companies in Indonesia, namely Tokopedia (total followers as of August 31, 2021 = 926,228), Bukalapak (total followers as of August 31, 2021 = 219,827), and BliBli (total followers as of August 31, 2021 = 538,091). Based on the data collected, a total of 12,786 posts were obtained, with 11,870,254 favorites and 2,735,886 retweets from three Twitter accounts' timelines (i.e., @tokopedia, @bukalapak, @bliblidotcom).

3.2. Data Preparation

In this step, we identify the available features of a Twitter post (see Fig. 2). Then, we extract the predictor and predicted variables (see Table 2). The predictor variables consist of six categorical variables (i.e., link, hashtag, image, short video, long video, and day of post) and a numerical variable (i.e., last post time). The use of links, hashtags, images, short videos, and long videos is represented by binary variables, namely "Yes" and "No". The day of posts indicates whether the post was made during the "Weekday" or the "Weekend". The last post time represents the time difference between current and previous posts that can indicate the posting frequency.

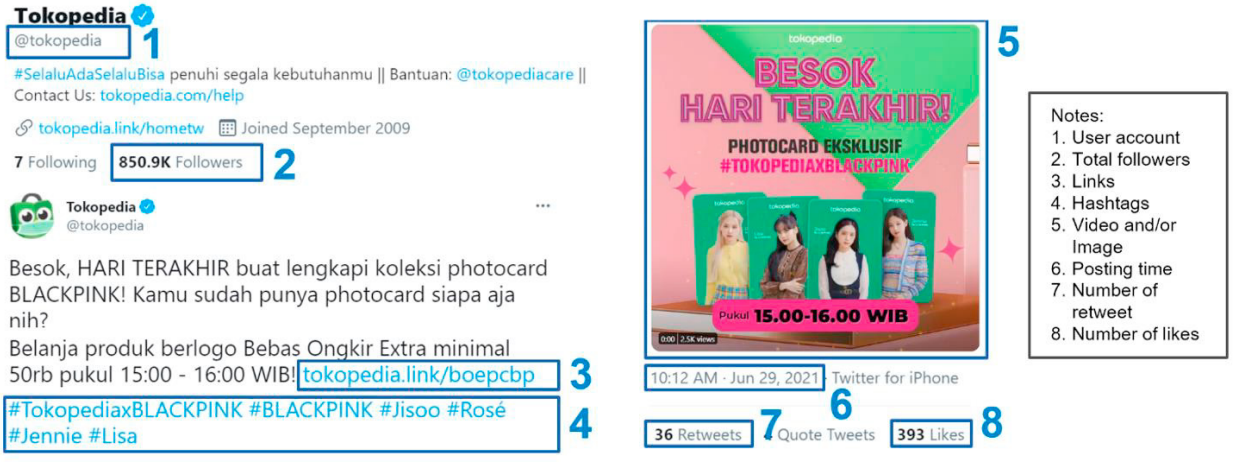

Fig. 2. Identification of variables on a Twitter post.

As in [20], the marketing research analyzed millions of posts from various profiles to obtain engagement rates across social media platforms. They calculate the engagement rates in Twitter from the total engagements divided by the total followers and multiplied by 100 to get the percentage value. Then, they categorize the engagement rates into four levels of engagement based on a statistical percentile (i.e., 25th, 50th, 75th, and 100th percentile). Thus, we benchmark the results of four engagement levels for Twitter, namely low (engagement rate = 0 - 0.02%), good (engagement rate = 0.02 - 0.09%), high (engagement rate = 0.09 - 0.33%), and very high engagement (engagement rate > 0.33%). In this study, we calculate the total engagements by adding the total number of retweets and likes of a Twitter post. Then, the engagement rate was obtained from the total engagements over total followers at the posting time and multiplied by 100.

The results of variable extraction in Table 2 indicate that the classes are imbalanced. Therefore, we split the dataset into 80% for the training dataset and 20% for the testing dataset using the stratified sampling method. Then, we perform the resampling technique. Stratified sampling divides the population into strata, selects samples from each stratum, and combines them to estimate population parameters. This method is used to balance the class on the training and testing dataset [3].

Table 2. Variables for the prediction model.

Variable	Data Type	Definition	Data Code	Tokopedia		Bukalapak		BliBli	
				n	%	n	%	n	%
Predictor Variables									
a. Interactivity									
Link	Categorical	Links to the website or the social media post	No	3,054	51.26%	2,811	79.81%	1,797	54.36%
			Yes	2,904	48.74%	711	20.19%	1,509	45.64%
Hashtag	Categorical	Use hashtags in posts	No	2,661	44.66%	3,186	90.46%	1,210	36.60%
			Yes	3,297	55.34%	336	9.54%	2,096	63.40%
b. Vividness									
Image	Categorical	Use the image in posts	No	4,729	79.37%	3,073	87.25%	1,034	31.28%
			Yes	1,229	20.63%	449	12.75%	2,272	68.72%
Short video	Categorical	Use video 0-10 seconds	No	5,919	99.35%	3,518	99.89%	3,301	99.85%
			Yes	39	0.65%	4	0.11%	5	0.15%
Long video	Categorical	Use video > 10 seconds	No	5,822	97.72%	3,510	99.66%	3,254	98.43%
			Yes	136	2.28%	12	0.34%	52	1.57%
c. Temporal factor									
Day of post	Categorical	Posting during weekday or weekend	Weekend	1,046	17.56%	715	20.30%	769	23.26%
			Weekday	4,912	82.44%	2,807	79.70%	2,537	76.74%
Last post time	Double	Time difference with the previous post	-	5,958	100.00%	3,522	100.00%	3,306	100.00%
Predicted Variable									
Engagement level	Categorical	%Engagement rate = Total engagements (likes + retweets) / Total followers *100	Low (0-0.02%)	2,118	35.55%	3,044	86.43%	3,075	93.01%
			Good (0.02-0.09%)	2,195	36.84%	301	8.55%	180	5.44%
			High (0.09-0.33%)	829	13.91%	138	3.92%	41	1.24%
			Very high (>0.33%)	816	13.70%	39	1.11%	10	0.30%

3.3. Model Development

This research focuses on developing prediction models with a multi-class classification that can accurately predict the engagement level of Twitter posts. In terms of classifiers category, we investigate both singular and ensemble models. For singular models, we used multinomial logistic regression, decision tree, and KNN, whereas, for the ensemble model, we employed random forest.

a. Multinomial Logistic Regression

We employed logistic regression to describe the relationship between the independent variables and the dichotomous or multi-categorical dependent variable [17]. Since the level of engagement is more than two categories, the appropriate logistic regression method is multinomial logistic regression. We formalize the model as follow:

$$P(Y_i) = \frac{e^{\alpha+\beta_{i1}x_1+\beta_{i2}x_2+\ldots+\beta_{in}x_n}}{1+e^{\alpha+\beta_{i1}x_1+\beta_{i2}x_2+\ldots+\beta_{in}x_n}} \quad (1)$$

where: $P(Y_i)$ is the probability of class i (class 1: low, class 2: good, class 3: high, class 4: very high), e is 2.71828 (natural number), α is Y_i intercept, β_{in} is the regression coefficients, and x_n $(n = 1,2,\ldots,7)$ is the predictor variables.

b. Decision Tree

The decision tree is a classification procedure through a tree-like structure consisting of three primary segments: a root node, the hidden nodes, and the terminal nodes or leaves [3], [18]. As a non-parametric method, decision trees have good adaptability with various datasets and can deal with nonlinear relationships [18]. The quality of split for every node evaluates using the Gini index, where the estimated probability of misclassification of the Gini index can be written as follow [18]:

$$i(t) = 1 - \sum_{j=1}^{K} P_j(t)^2 \quad (2)$$

where: $i(t)$ is the estimated probability of misclassification in node t, $P_j(t)$ $(j = 1,2,\ldots,K)$ is the probability corresponding to class j in node t, and K is the number of classes ($K = 4$, namely low, good, high, very high).

c. K-Nearest Neighbor (KNN)

KNN computes the distance between samples or instances [19], [21]. Then, the algorithm will shortlist the samples closer to the new data sample called the nearest neighbor and conduct classification for each value. Most KNN algorithms calculate the distance using Euclidean distance, which can be written as follow [21]:

$$Dist(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (3)$$

where: x is the new data sample, y is the training data sample, x_i $(i = 1,\ldots,N)$ is the attribute of new data samples x, y_i $(i = 1,\ldots,N)$ is the attribute of training data samples y, and N is the number of attributes.

d. Random Forest

The random forest consists of several decision trees where each decision tree is trained on a different set of records and variables randomly selected [3]. Thus, a random forest can accommodate multiple classifiers, and the ensemble prediction is made by combining the results of a single tree [22]. The ensemble score of random forest is as follow [22]:

$$\widehat{Y_i} = mode_{n=1\ldots N_{trees}} \widehat{Y_n} \quad (4)$$

where: $\widehat{Y_i}$ is the ensemble result, $mode$ is the most frequent result, N_{trees} is the number of trees in the ensemble, and $\widehat{Y_n}$ is the result of a single tree.

Based on the four classifiers, a total of 12 models are trained through the training dataset and evaluated using the test dataset. During the model training, we perform 10-fold cross-validation. This method was applied to find optimal parameters, and iterations were executed until each model yielded the highest PR AUC score of 10-fold cross-validation.

3.4. Model Evaluation

The accuracy of the models is evaluated using several metrics, namely precision, recall, and F_1 score. One study indicated that the precision-recall curve is more informative than the receiver operating characteristics (ROC) curve when evaluating classifiers on imbalanced data [23]. Both precision and recall allow us to evaluate a classifier's performance on the minority class. The F_1 score represents the harmonic mean of precision and recall. High precision, recall, and F_1 score indicate higher performance in classification models. We formalize the macro-average precision, recall, and F_1 score for multi-class classification as follows [24]:

$$Precision_M = \frac{\sum_{i=1}^{K} \frac{TP_i}{TP_i + FP_i}}{K} \tag{5}$$

$$Recall_M = \frac{\sum_{i=1}^{K} \frac{TP_i}{TP_i + FN_i}}{K} \tag{6}$$

$$F_1\ score_M = 2 \times \frac{Precision_M \times Recall_M}{Precision_M + Recall_M} \tag{7}$$

where: $Precision_M$ is the macro-average precision, $Recall_M$ is the macro-average recall, TP is true positive, FP is false positive, and FN is false negative, K is the number of classes, and $F_1\ score_M$ is the macro-average F_1 score.

4. Results and Discussion

Table 3 shows the results of 10-fold cross-validation for the machine learning classification models. The results summarize the PR AUC of the validation fold on each model, the mean of PR AUC, and the model development time. The results show that constructing a decision tree model requires the least processing time of 15.654 s, 16.974 s, and 18.323 s for Tokopedia, Bukalapak, and Blibli, respectively. However, the performance is below random forest and KNN. The random forest model leads to the best PR AUC of 0.595, 0.818, and 0.795 for Tokopedia, Bukalapak, and Blibli, respectively. Nevertheless, the models require more time to develop. The cross-validation results indicate that the multinomial logistic regression models perform worst with PR AUC between 0.366 and 0.461. As the results of 10-fold cross-validation, we concluded that the ensembled model requires a longer time for model development than the singular models, but it can provide higher classification performance.

Table 3. Results of 10-fold cross validation.

	Multinomial Logistic Regression			Decision Tree			K-Nearest Neighbor			Random Forest		
	1	2	3	1	2	3	1	2	3	1	2	3
Fold 1	0.404	0.365	0.464	0.449	0.410	0.603	0.458	0.496	0.313	0.609	0.810	0.795
Fold 2	0.423	0.362	0.485	0.392	0.391	0.570	0.451	0.489	0.323	0.597	0.815	0.791
Fold 3	0.391	0.380	0.446	0.431	0.348	0.574	0.467	0.493	0.314	0.580	0.822	0.842
Fold 4	0.420	0.371	0.464	0.428	0.399	0.614	0.460	0.473	0.309	0.612	0.802	0.764
Fold 5	0.412	0.363	0.461	0.446	0.401	0.556	0.460	0.512	0.302	0.607	0.822	0.823
Fold 6	0.404	0.360	0.483	0.425	0.419	0.617	0.479	0.473	0.317	0.601	0.824	0.740
Fold 7	0.395	0.358	0.447	0.351	0.399	0.504	0.461	0.477	0.324	0.589	0.821	0.807
Fold 8	0.420	0.368	0.457	0.367	0.413	0.566	0.464	0.476	0.314	0.607	0.833	0.783
Fold 9	0.401	0.363	0.447	0.429	0.412	0.589	0.444	0.475	0.325	0.576	0.795	0.773
Fold 10	0.420	0.370	0.454	0.384	0.367	0.615	0.454	0.494	0.326	0.576	0.839	0.828
Mean PR AUC	0.409	0.366	0.461	0.410	0.396	0.581	0.460	0.486	0.317	0.595	0.818	0.795
Model development time (seconds)	108.36	131.59	154.37	15.65	16.97	18.32	41.66	45.18	49.21	343.93	426.16	415.48

Note: 1 = Tokopedia, 2 = Bukalapak, 3 = Blibli.

Table 4 shows the performance of the models in the out-sample contexts. The random forest models show the

best engagement prediction performance in all three e-commerce contexts. The random forest provides the best recall of 0.77 in the Tokopedia context, 0.786 in the Bukalapak context, and 0.843 in the Blibli context. It means that the models have a low false-negative rate. However, the precisions are relatively low for Blibli with a high false-positive rate. The second-best model is KNN that yielded a recall of 0.574 for Tokopedia, 0.715 for Bukalapak, and 0.794 for Blibli. Meanwhile, multinomial logistic regression shows the most inferior performance.

These findings show that the random forest, which employs an ensemble classifier, outperforms other singular classifiers (i.e., multinomial logistic regression, decision tree, and KNN). The random forest offers various advantages, such as robustness against overfitting and dealing with high-dimensional problems [22]. Furthermore, as a non-parametric method, the random forest does not require distributional assumption for the training dataset [18]. However, the complexity and model constructing time of a random forest increase with the number of trees and training samples.

Table 4. Performance metrics of the prediction models.

	Multinomial Logistic Regression			Decision Tree			K-Nearest Neighbor			Random Forest		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Tokopedia												
Class: Low	0.668	0.597	0.630	0.806	0.802	0.804	0.676	0.670	0.673	0.672	0.715	0.693
Class: Good	0.703	0.674	0.688	0.825	0.549	0.659	0.729	0.319	0.444	0.757	0.390	0.514
Class: High	0.328	0.133	0.189	0.291	0.313	0.301	0.315	0.566	0.405	0.616	0.976	0.755
Class: Very High	0.311	0.620	0.414	0.388	0.712	0.502	0.429	0.742	0.544	0.647	1.000	0.786
Macro-Average	0.502	0.506	0.504	0.577	0.594	0.586	0.537	0.574	0.555	0.673	0.770	0.718
Bukalapak												
Class: Low	0.911	0.708	0.797	0.927	0.335	0.492	0.977	0.345	0.510	0.968	0.693	0.808
Class: Good	0.211	0.133	0.163	0.167	0.283	0.210	0.195	0.550	0.288	0.341	0.700	0.459
Class: High	0.089	0.286	0.136	0.095	0.536	0.161	0.121	0.964	0.214	0.457	0.750	0.568
Class: Very High	0.029	0.375	0.054	0.022	0.625	0.043	0.082	1.000	0.152	0.080	1.000	0.148
Macro-Average	0.310	0.375	0.340	0.303	0.445	0.360	0.344	0.715	0.464	0.461	0.786	0.581
Blibli												
Class: Low	0.968	0.387	0.553	1.000	0.213	0.351	0.994	0.288	0.446	0.994	0.581	0.733
Class: Good	0.094	0.583	0.162	0.070	0.667	0.127	0.099	0.889	0.178	0.149	0.917	0.257
Class: High	0.022	0.250	0.040	0.038	0.500	0.071	0.063	1.000	0.119	0.101	0.875	0.182
Class: Very High	0.010	0.500	0.019	0.024	1.000	0.048	0.063	1.000	0.118	0.167	1.000	0.286
Macro-Average	0.273	0.430	0.334	0.283	0.595	0.384	0.305	0.794	0.440	0.353	0.843	0.498

In general, the prediction models show a balance between precision and recall in the Tokopedia context only. While in Bukalapak and Blibli contexts, the models show a high false-positive rate. Table 2 shows that the prediction class imbalance in Bukalapak and Blibli contexts is more severe than in Tokopedia. Thus, there are a lot of negative samples that were identified as false-positive. Otherwise, there are fewer positive samples that could be identified as false-negative. Our case study needs a balance of precision and recall because it relates to the companies' social media efforts.

When we take a closer look at the social media activities, each e-commerce shows different patterns of Twitter posts. During the same period of seven months, Tokopedia posted more frequently than Bukalapak and Blibli. Tokopedia posted 5,958 tweets during February – August 2021, 69% higher than Bukalapak and 80% higher than Blibli. In addition, Tokopedia also uses more interactive elements (i.e., links and hashtags) and dynamic visual elements (i.e., short and long video). Meanwhile, Blibli is the one who uses images and hashtags the most in their tweets (68% and 63% of the total tweets use images and hashtags, respectively). In terms of the posting day, all e-commerce posts more frequently on weekdays than on weekends (76 – 82% of the total tweets are posted on

weekdays).

Table 5 shows the relative variable importance of the best model, i.e., the random forest model. Last post time that represented the posting frequency consistently ranked in the top tiers of variable importance in all e-commerce contexts. This finding is in line with the findings of the earlier study [3]. For Tokopedia, the factor is followed by images, posts on weekdays, and links. At the same time, interactive elements such as links and hashtags highlight the significant influence in Bukalapak and Blibli. Furthermore, static visual elements (images) emphasize influence in the engagement model for all e-commerce. In contrast, dynamic visual elements (short and long videos) are less influential in all cases. The findings point out that post frequency, interactive posting elements, and static visual elements played an important role in attracting the engagement of a company's Twitter posts. Hence, marketing and social media managers can utilize our prediction model to assess the prospective engagement level of the company's social media efforts and continuously improve their social media engagement strategies.

Table 5. Relative variable importance of random forest model.

	Tokopedia	Bukalapak	Blibli
Last post time	100.00%	100.00%	100.00%
Use images	6.33%	6.48%	8.08%
Posting on weekdays	4.84%	6.13%	6.60%
Use links	4.82%	7.93%	14.26%
Use hashtags	4.11%	8.26%	5.76%
Long videos	1.19%	0.50%	1.87%
Short videos	0.46%	0.61%	0.33%

5. Conclusion

We provide a solution for predicting the engagement level of Indonesia's e-commerce businesses based on their social media activities, i.e., Twitter account, using machine learning models. We analyzed a total of 12,786 unique Twitter posts with 11,870,254 favorites and 2,735,886 retweets from the official Twitter account of three of the largest e-commerce platforms in Indonesia, namely Tokopedia (i.e., @tokopedia), Bukalapak (i.e., @bukalapak), and Blibli (i.e., @bliblidotcom) over seven months period (i.e., February 1 – August 31, 2021). In developing the prediction models, we incorporate three constructs with seven features, including interactivity (i.e., link, hashtag), vividness (i.e., image, short video, long video), and temporal factors (i.e., day of post, last post time). Post frequency, interactive posting elements, and static visual elements played an important role as dominant predictors of the engagement of the Twitter post. Our study shows that the random forest model provides the best prediction performance (i.e., precision, recall, and F_1 score) than the singular classifier models such as multinomial logistic regression, decision tree, and k-nearest neighbor models.

From a practical point of view, the proposed machine learning model can support marketing and social media managers in predicting the engagement level of Twitter posts so that they improve their corresponding social media strategy. To the best of our knowledge, this study is one of the earliest research to predict the engagement level of Indonesia's e-commerce businesses based on their social media activities using the machine learning approaches. Nevertheless, this study has several limitations that open opportunities for further research. Our study only focuses on Twitter, thus, one can attempt to develop predictive models for other social media platforms, such as Facebook, YouTube, Instagram, etc. [4]–[7]. Next, we only focus on three constructs as the predictors for the machine learning models, namely vividness, interactivity, and temporal factors. Further research can incorporate other constructs, such as content type [6]–[8], in developing the prediction models. Lastly, more advanced data pre-processing approaches development methods can be further researched to handle the imbalanced data cases that often exist in real-world situations.

References

[1] H. Tankovska, "Most popular social networks worldwide as of April 2021, ranked by number of active users," *Statista*, 2021. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed Jul. 05, 2021).

[2] H. Tankovska, "Leading countries based on number of Twitter users as of April 2021," *Statista*, 2021. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed Jul. 05, 2021).

[3] G. Aydin, N. Uray, and G. Silahtaroglu, "How to engage consumers through effective social media use-guidelines for consumer goods

companies from an emerging market," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 4, pp. 1–23, 2021, doi: 10.3390/jtaer16040044.

[4] I. Antoniadis, S. Paltsoglou, and V. Patoulidis, "Post popularity and reactions in retail brand pages on Facebook," *Int. J. Retail Distrib. Manag.*, vol. 47, no. 9, pp. 957–973, 2019, doi: 10.1108/IJRDM-09-2018-0195.

[5] F. Sabate, J. Berbegal-Mirabent, A. Cañabate, and P. R. Lebherz, "Factors influencing popularity of branded content in Facebook fan pages," *Eur. Manag. J.*, vol. 32, no. 6, pp. 1001–1011, 2014, doi: 10.1016/j.emj.2014.05.001.

[6] C. D. Schultz, "Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages?," *Electron. Commer. Res. Appl.*, vol. 26, pp. 23–34, 2017, doi: 10.1016/j.elerap.2017.09.005.

[7] I. Pletikosa Cvijikj and F. Michahelles, "Online engagement factors on Facebook brand pages," *Soc. Netw. Anal. Min.*, vol. 3, no. 4, pp. 843–861, 2013, doi: 10.1007/s13278-013-0098-8.

[8] L. De Vries, S. Gensler, and P. S. H. Leeflang, "Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing," *J. Interact. Mark.*, vol. 26, no. 2, pp. 83–91, 2012, doi: 10.1016/j.intmar.2012.01.003.

[9] R. J. Brodie, L. D. Hollebeek, B. Jurić, and A. Ilić, "Customer engagement: Conceptual domain, fundamental propositions, and implications for research," *J. Serv. Res.*, vol. 14, no. 3, pp. 252–271, 2011, doi: 10.1177/1094670511411703.

[10] P. Luarn, Y. F. Lin, and Y. P. Chiu, "Influence of Facebook brand-page posts on online engagement," *Online Inf. Rev.*, vol. 39, no. 4, pp. 505–519, 2015, doi: 10.1108/OIR-01-2015-0029.

[11] R. Dolan, J. Conduit, J. Fahy, and S. Goodman, "Social media engagement behaviour: a uses and gratifications perspective," *J. Strateg. Mark.*, vol. 24, no. 3–4, pp. 261–277, 2016, doi: 10.1080/0965254X.2015.1095222.

[12] S. Burton and A. Soboleva, "Interactive or reactive? Marketing with Twitter," *J. Consum. Mark.*, vol. 28, no. 7, pp. 491–499, 2011, doi: 10.1108/07363761111181473.

[13] W. Macias, "A beginning look at the effects of interactivity, product involvement and web experience on comprehension: Brand web sites as interactive advertising," *J. Curr. Issues Res. Advert.*, vol. 25, no. 2, pp. 31–44, 2003, doi: 10.1080/10641734.2003.10505147.

[14] D. L. Hoffman and T. P. Novak, "Marketing in hypermedia computer-mediated environments: Conceptual foundations," *J. Mark.*, vol. 60, no. 3, pp. 50–68, 1996, doi: 10.2307/1251841.

[15] J. Steuer, "Defining Virtual Reality: Dimensions Determining Telepresence," *J. Commun.*, vol. 42, no. 4, pp. 73–93, 1992, doi: 10.1111/j.1460-2466.1992.tb00812.x.

[16] T. F. Trefzger, C. V. Baccarella, and K.-I. Voigt, "Antecedents of brand post popularity in Facebook: The influence of images, videos, and text," *Proc. 15th Int. Mark. Trends Conf.*, no. January, pp. 1–8, 2016.

[17] J. Liang, G. Bi, and C. Zhan, "Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R," *Ann. Transl. Med.*, vol. 8, no. 16, pp. 982–982, 2020, doi: 10.21037/atm-2020-57.

[18] X. E. Pantazi, D. Moshou, and D. Bochtis, *Intelligent Data Mining and Fusion Systems in Agriculture*. Academic Press, 2020.

[19] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.

[20] G. Mee, "What is a Good Engagement Rate on Twitter?," *Social*, 2020. https://scrunch.com/blog/what-is-a-good-engagement-rate-on-twitter (accessed May 10, 2021).

[21] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, vol. 391, no. xxxx, pp. 234–242, 2020, doi: 10.1016/j.neucom.2018.11.101.

[22] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 88, no. June 2019, p. 102051, 2020, doi: 10.1016/j.jag.2020.102051.

[23] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, pp. 1–21, 2015, doi: 10.1371/journal.pone.0118432.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.