Unveiling Post Engagement Rate Using Predictive Analytics Model for Instagram Accounts of Coffee Shops in West Jakarta

Sachio Dariell Sutanto*1
School of Business and Management, Institut Teknologi Bandung1,2
Email: sachio_dariell@sbm-itb.ac.id

Abstract
During the COVID-19 Pandemic, telemedicine has had a high use rate since the government issued a policy limiting people's movement to reduce transmission. It is because telemedicine can be a way to solve this issue since it can provide and support health care when distance separates the users only using electronic communication technologies. However, through this remote assessment, patients and physicians lose the opportunity to interact in person, which might impact doctor-patient communication. Furthermore, covid-19 cases in Indonesia have shown a downward trend, making the government have no regulation limiting people's movement. Now, people can see their physicians directly for some conditions. Therefore, this research will analyze factors influencing customer intention to consult through telemedicine applications and determine which factors more significantly influence customer intention. This research is conducted through a qualitative approach by semi-structured interviews and a quantitative approach by an online survey of customers who ever experience using telemedicine applications. The researcher uses open coding to analyze the interview result and descriptive statistics and PLS-SEM to analyze the survey result. The author gets seven respondents from interviews and 317 from online surveys. The results indicate that perceived benefit, satisfaction, saving time, saving cost, and performance expectancy influence customer intention to use telemedicine applications continuously. Furthermore, satisfaction and saving costs are the factors that significantly influence customer intention to use telemedicine applications continuously. The finding of this research is expected to give insight into telemedicine applications in Indonesia about making their strategies to enhance their customer intention to use continuously.

Keywords: intention to use continuously; telemedicine application; satisfaction

A. INTRODUCTION
Over the past decade, there has been a new and emerging industry that gradually changes the behavior of people into consuming more and more of their products and services, that is the Coffee Shop Industry. The global industry of coffee shop has grown rapidly, and it seems to continue growing, especially in recent years when the Covid-19 pandemic has begun to turn into an endemic. The rapid growth of this industry creates fierce competition that forces coffee shop owners to constantly seek innovation or strategies to differentiate themselves from their competitors, allowing them to attract loyal customers, and drive sustainable growth. This global trend is not limited to certain regions, as we can see a similar phenomenon in countries like Indonesia.

The number of coffees consumed by Indonesian people is increasing yearly, even when the country is still affected by the COVID-19 pandemic, as reflected in the number of coffee consumption from 2019/2020 to 2020/2021. At the same time, the number of coffee shops in Indonesia also grew from 1083 in 2016 to 2937 in 2019 (Toffin and Mix, 2020:9). The growing coffee consumption and the flourishing coffee shop industry in Indonesia have had a noticeable impact on various regions and West Jakarta is one of the areas that has experienced the rapid growth of coffee shops. West Jakarta was quite famous for several colonial relics that are still preserved and maintained by the government, which attracts people from outside West Jakarta to visit this area. In addition to the colonial relics, many private universities in Jakarta are in West Jakarta, e.g., Bina Nusantara University, Trisakti University, Universitas Kristen Krida Wacana, Tarumanegara University, etc. These universities in West Jakarta have the biggest student density at 5252 students/km2 (BPS, 2021).

The tourist attractions and universities allow coffee shops to be built in West Jakarta to accommodate the need for places to rest and hang out with friends or families. As coffee shops emerge in this thriving environment, it becomes essential to recognize that the successful establishment and growth of a coffee shop venture go beyond
physical infrastructure alone. Building a coffee shop includes the marketing effort to create a consumer base and grow the business. When we want to grow a business, one of the most frequently used methods to market its business is by using social media.

Social media is an internet-based channel of mass personal communication that facilitates interactions between users that has value primarily from user-generated content (Carr and Hayes, 2015:49), and it changed how information spreads around the world. Social media is now the top choice for businesses to engage with their customers and consumer base. It is also the best media to boost the number of customers that come to their coffee shops by creating social media campaigns. Social media offers unparalleled opportunities to boost the number of visiting customers to coffee shops. With a well-executed social media campaign, a coffee shop can leverage the power of viral marketing, potentially attracting dozens of customers lining up to the coffee shop. However, attracting customers on social media is more complex than it may seem, the post on social media should be visually appealing to the netizens that would see the posts. If we use social media correctly, it could be a fantastic tool to generate customer engagement and brand awareness, for all kinds of businesses (Sledge, 2014). However, it is important to note that simply posting what we believe is the best may not align with what consumers think is attractive. Nowadays, companies need to have the ability to understand the targeted customers’ value, and the reason for their attraction to the company (Thompson, 1998:21), to create a better social media interaction that aligns our product with our customers’ value.

In selecting specific social media platforms, coffee shops must consider several factors that contribute to effective customer engagement, and two of the main factors are the number of users and the engagement rate of the platform. In January 2023, the number of social media users in Indonesia was 167.0 million, equivalent to 60.4% of the total population (DataReportal, 2023). In Indonesia, the currently most used social media platform is WhatsApp, followed by Instagram and Facebook (DataReportal, 2023). Since WhatsApp is not a platform we can use to engage with followers and strangers on the internet, this research decided to use Instagram because of its ability to interact with followers, with the possibility to count the engagement rate for each post. The other reason why we choose Instagram is that the average engagement rate of all types of posts is better on Instagram than on Facebook.

Engagement Rate is a way to measure how much influence an Instagram account has on its followers. It is one of the metrics we can use to know how to optimize an endorsement on Instagram influencers (Arman and Sidik, 2019:5). The variable engagement rate will be crucial for coffee shops to understand whether their Instagram post is doing great in giving exposure to the coffee shop or not. Instagram is one of the most used social media platforms in the world, reaching 1.628 billion users on advertising reach in April 2023 (DataReportal, 2023). It is a platform that we can use to capture, edit, and share photos or videos in the form of Instagram posts. The post may be accompanied by captions, hashtags, location tags, and other features. These companions on the post are the data that we can collect to predict the engagement rate for an Instagram post.

In this research, we want to understand the interaction between the variables. Understanding the interaction means that we can create a prediction on what type of Instagram posts could generate the most engagement with the consumer base. If we can predict the most effective post for Instagram’s engagement rate, we can create an effective and efficient Instagram marketing plan. To understand the relationship between the variables, we will use linear regression, with a little tweak to accommodate the independent variables which are mostly categorical. Linear regression can be defined as a formula or a method to understand the correlation of a linear relationship between dependent and independent variables (Groß, 2003:4). The regression method uses a continuous variable as the dependent variable, and the independent variable may vary between categorical or continuous variables.

The information from social media can be used to forecast future outcomes, specifically by using the Linear Regression model (Asur and Huberman, 2010:499). Suppose we can use linear regression to create a predictive model for social media posts. In that case, we can forecast the outcomes of our engagement rate, creating a possibility of a better Instagram marketing plan using the given variables. This research develops a linear regression predictive analytics model to predict the expected engagement rate of Instagram posts using RStudio. RStudio is an application that could be used for R programming, typically for regression or other machine learning techniques. RStudio uses all functions available in R, and it uses a more user-friendly interface to allow the users to analyze data comfortably and generate the graphical outcome for the data (Kronthaler and Zöllner, 2021:1-2).

This research aims to understand the interaction between the engagement rate and other observable variables. Suppose we can understand the correlation between those variables. In that case, it means that we can create a prediction on what type of Instagram posts could generate the most engagement with the consumer base.
B. RESEARCH METHODS

In this study, we aim to investigate the impact of independent variables toward the value of engagement rate on the Instagram posts of several West Jakarta Coffee Shop’s Instagram Account. Based on our theoretical exploration we portray the following theoretical conceptual model at Figure 1.

![Figure 1. Theoretical Framework](image)

In line with the theoretical conceptual model definition, we propose the following hypotheses to guide our research process and address the objectives in this research:

H1: There is significant impact of Post Content Type toward the Engagement Rate.

H2: There is significant impact of Caption Type toward the Engagement Rate.

H3: There is significant impact of Location Type toward the Engagement Rate.

H4: There is significant impact of Hashtag Use toward the Engagement Rate.

H5: There is significant impact of Audio Use toward the Engagement Rate.

In this research, we will use Web Data Extraction, which is a method of data extraction that uses the data stored from a website source as the main source of data (Ferrara et al., 2014:301). To do that, we will use Apify to create a data extraction task on the Instagram Scraper API. In this task, we should define the input parameter for the Instagram account data that will be collected. The input parameter here is the Instagram Account’s URL, time frame, and max number of posts that will be extracted.

In collecting the data, we will use the time frame between the 1st of January 2022 and the 31st of December 2022. The maximum number of posts will be 100 posts per Instagram Account. After giving the parameters or limitations on the Apify Task, we will use the Start button to start the scraping task on the API. Then, after collecting the data for this research, we cleaned, transformed, and validated the data. The total amount of collected clean data is 417. In addition to collecting the data, there will be several manually categorized data, which are Post Type, Location Type, Audio Use, Caption Type, and the Engagement Rate calculated by using the formula.

\[
Engagement\ Rate = \frac{Likes + Comments}{Total\ Followers}
\]

After processing the data, we generate a multiple linear regression model representing the relationship between the dependent variable and the independent variables. We will divide the data into training and test dataset, then the training dataset will generate an equation that will serve as a model to allow make predictions or draw conclusions based on the independent variable. These model will be developed by using RStudio.

The division will be conducted by separating the collected data into 70% and 30% using a “set.seed” function to ensure that any iteration of the division will be the same when we clean the workspace in RStudio.
Then, we continue to evaluate the predictive analytics model by using the equation to calculate predictions, then testing it to the test dataset.

To create the prediction model, we will use the function “lm()” from RStudio to create the linear regression model using the training dataset, using the following code:

```r
linearregression <- lm(engagement_rate ~ caption_type + hashtag_use + location_type + post_type + audio_use, data = trainSet)
```

By executing the defined code, Rstudio has created a linear regression equation which will be represented by the variable “linearregression”. To further comprehend the equation, we will use the function “summary()” with “linearregression” as the called variable for the function to return the result of the regression and other statistical insights.

Then, to validate the reliability of the machine learning model, we will use RMSE and MAE as the metrics to measure reliability of the model. Both can be done in RStudio by using the respective function of “mean((modelEval$Actual - modelEval$Predicted)^2)” for predicting RMSE and we can then use the function “mean(abs(modelEval$Actual - modelEval$Predicted))” for predicting MAE.

RMSE is a measure to understand the overall difference between the predicted and actual values, which then indicates how well the prediction aligns with the actual data. A lower RMSE means that the model is quite accurate since the overall difference between the whole predicted value and the actual value is low.

MAE is a measure to understand the average absolute difference between the predicted and the actual values of a dataset. MAE indicates the average error of the prediction model in comparison to the actual value from the dataset. Like the RMSE, the lower value that MAE generates means that the prediction accuracy is better.

The research’s purpose is to understand the interaction between the engagement rate and other observable variables using linear regression prediction model. Suppose we can understand the correlations, it means that we can create a prediction on what type of Instagram posts could be generated in order to maintain a high engagement rate. The generated model will then answer the hypotheses and research question, thus fulfilling the research objective.

C. RESULTS AND ANALYSIS

From the summary function in RStudio, we decided to analyze each factor from the analyzed variables using train dataset to determine the significance of all variables, specifically each factor within the variables toward the dependent variable. In this research, we will use the significance level of 0.05 to understand the significance of each variable using the P Values of each factor to answer the proposed hypotheses. The equation summary shows that there are only two specific factors have a significant impact into the engagement rate value, the factors are an Endorsement Post from the Post Type variable, and the City of the Coffee Shop from the Location Type variable.

To further understand the result of linear regression equation from RStudio, the researchers converted the estimate value from the summary into a Multiple Linear Regression equation that can be seen in the formula below:

\[
y = 0.0137009x_1 + 0.0031128x_2 + 0.0021084x_3 + 0.0042954x_4 + 0.0042047x_5 + 0.0109943x_6 \\
+ 0.0071867x_7 + 0.0060929x_8 + 0.0020010x_9 + 0.0001449x_{10} + 0.0021084x_{11} \\
+ 0.0062277x_{12} + 0.0081009x_{13} - 0.0020412x_{14} + 0.0049077x_{15} + 0.0008522x_{16} \\
+ 0.0041713x_{17} - 0.0018726x_{18} - 0.0039268x_{19} - 0.0008522x_{20}
\]

Where:
- \(y\) = Engagement Rate
- \(x_1\) = post_type1 (endorsement post)
- \(x_2\) = post_type2 (events within the coffee shop or celebration post)
- \(x_3\) = post_type3 (photo of the menus)
\[ x_1 = \text{post\_type4 (ambience of the coffee shop)} \]
\[ x_2 = \text{post\_type5 (discount or promotion poster)} \]
\[ x_3 = \text{post\_type6 (video about the coffee shop)} \]
\[ x_4 = \text{post\_type7 (video of the menus)} \]
\[ x_5 = \text{post\_type8 (video of events within the coffee shop)} \]
\[ x_6 = \text{caption\_type1(grateful or playful interaction with the customers)} \]
\[ x_7 = \text{caption\_type2 (promoting purchase packages)} \]
\[ x_8 = \text{caption\_type3 (promoting specific menus)} \]
\[ x_9 = \text{caption\_type4 (promoting the ambience)} \]
\[ x_{10} = \text{caption\_type5 (promoting the coffee shop or the coffee shop location)} \]
\[ x_{11} = \text{caption\_type6 (promoting or explaining the services)} \]
\[ x_{12} = \text{caption\_type7 (celebrating special day)} \]
\[ x_{13} = \text{hashtag\_use1 (presence of hashtag within the caption)} \]
\[ x_{14} = \text{location\_type1 (city of the coffee shop)} \]
\[ x_{15} = \text{location\_type2 (coffee shop location)} \]
\[ x_{16} = \text{location\_type3 (every other location)} \]
\[ x_{17} = \text{audio\_use1 (presence of audio)} \]

In that equation, the coefficient each represents the effect of the predictor variables on the dependent variable Engagement Rate. Since the regression variables are categorical, the coefficient value will only be triggered if the specific categorical value is present. We tweaked the equation to show the first variable for post type, endorsement post type, at the expense of the intercept value.

For example, when the caption is caption\_type1, then only the variable \( x_1 \) has the value of 1, it increases the value of engagement\_rate by 0.0020010. Meanwhile, all other caption\_type variables have a 0 value, which does not affect the value of Engagement Rate. Similarly, other variables also contribute to the dependent variable based on their respective coefficients and the presence of other variables in the same categorical variable type.

From the same summary function, we can further analyze the significance of each factors of the variables, by using the P Values of each respective variables in the equation, then we reject or not reject each variables according to the given hypotheses. The result of analyzing the P Values can be seen on the table below.

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>T Statistics</th>
<th>P Values</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{1.1} )</td>
<td>3.395</td>
<td>0.001</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>( H_{1.2} )</td>
<td>0.863</td>
<td>0.389</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.3} )</td>
<td>1.321</td>
<td>0.187</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.4} )</td>
<td>1.908</td>
<td>0.057</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.5} )</td>
<td>1.040</td>
<td>0.299</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.6} )</td>
<td>1.619</td>
<td>0.107</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.7} )</td>
<td>1.110</td>
<td>0.268</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.8} )</td>
<td>0.930</td>
<td>0.353</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.9} )</td>
<td>0.020</td>
<td>0.536</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.10} )</td>
<td>0.039</td>
<td>0.969</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.11} )</td>
<td>0.036</td>
<td>0.525</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.12} )</td>
<td>1.685</td>
<td>0.093</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.13} )</td>
<td>1.887</td>
<td>0.080</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.14} )</td>
<td>-0.480</td>
<td>0.632</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.15} )</td>
<td>1.407</td>
<td>0.161</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.16} )</td>
<td>0.624</td>
<td>0.533</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.17} )</td>
<td>2.150</td>
<td>0.032</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.18} )</td>
<td>-1.742</td>
<td>0.083</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.19} )</td>
<td>-1.360</td>
<td>0.175</td>
<td>Rejected</td>
</tr>
<tr>
<td>( H_{1.20} )</td>
<td>-0.362</td>
<td>0.718</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

Source: research data, 2023

From the equation, we will then evaluate the machine learning model, we will use RMSE and MAE as the evaluation metrics for the prediction performance.
Table 2. The Performance Evaluation Of The Machine Learning Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
<th>Evaluation Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear</td>
<td>Post Content Type + Caption Type + Hashtag Type + Location Type + Audio Type</td>
<td>RMSE: 0.09779162</td>
</tr>
<tr>
<td>Regression</td>
<td></td>
<td>MAE: 0.005533414</td>
</tr>
</tbody>
</table>

Source: research data, 2023

D. CONCLUSIONS

In this research, we concluded that not all variables have a significant impact towards the engagement rate of an Instagram post. Although insignificant, the machine learning model may still use all variables from the equation to make predictions for the engagement rate of future Instagram posts. When we mention that not all variables have a significant impact, there are two factors that have a significant impact towards the engagement rate, these factors are Endorsement Post Type and City of the Coffee Shop Location Type. It is important to note that being significant here means the P values are lower than the level of significance, and the significance are tested using the training dataset, thus making a limitation from the already limited number of samples from the dataset.

Then, have evaluated the model and gained an insight that the model has the value RMSE and MAE is 0.09779162 and 0.005533414 respectively, that means the model is quite accurate in predicting the engagement rate for Instagram posts, having an MAE value of 0.55% from the engagement rate value.

Researcher then concluded several points according to the research result: 1) The predictive analytics model can be used to predict several Instagram posts at once. If we want to grow the number of customers engaging with our coffee shop business, we can use this prediction model to create an Instagram marketing post plan, avoiding the combination that hurts the potential engagement rate for the Instagram post; and 2) when a coffee shop wants to post on their Instagram account, they should consider using the endorsement post and use the city of their coffee shop to be put in the location input of the Instagram post for better average value of engagement rate.

In addition, the researchers want to give a recommendation for future research as an evaluation for this research, such as: 1) Using other machine learning methods like artificial neural network or other statistic models for the predictive analytics to improve the accuracy of the prediction. The limitation of Multiple Linear Regression is that we must do statistical training and comprehension of the statistics model, and there could be a missed detection of complex relationship between variables when creating a prediction model (Tu, 1996:1229); 2) Using a photograph or image classifier or other Machine Learning/Artificial Intelligence method to further reduce the subjectivity of the image within the post. The subjectivity here can come from different people interpreting the guidelines of classification differently, instead of using a machine that has one perspective toward the given guidelines to classify the images; 3) using Naive Bayes or other text classifier types to further reduce the subjectivity for the caption type. The text classification can be inconsistent without clear keywords or limitation toward what kind of word strings directly classify the captions of the Instagram post; and 4) the future researchers or business owners may expand the scope of the research to make it more sophisticated or be ready to be used as a predictive model for a larger area, not limited to West Jakarta. Making the dataset larger can also create a more sophisticated machine learning model, increasing the accuracy of the prediction.

REFERENCES


