Improving social media engagements on paid and non-paid advertisements: a data mining approach

Jen-Peng Huang

Information Management Department,
Southern Taiwan University of Science and Technology,
No. 1, Nan-Tai Street, Yungkang Dist.,
Tainan City 710, Taiwan
Email: jehuang@stust.edu.tw

Genesis Sembiring Depari*

Business and Management Department,
Southern Taiwan University of Science and Technology,
No. 1, Nan-Tai Street, Yungkang Dist.,
Tainan City 710, Taiwan
Email: genesissembiring@gmail.com
*Corresponding author

Abstract: The purpose of this research is to develop a strategy to improve the number of social media engagement on Facebook both for paid and non-paid publications through a data mining approach. Several Facebook post characteristics were weighted in order to rank the input variables importance. Three machine learning algorithms performance along with dynamic parameters were compared in order to obtain a robust algorithm in assessing the importance of several input factors. Random forest is found as the most powerful algorithm with 79% accuracy and therefore used to analyse the importance of input factors in order to improve the number of engagements of social media posts. Eventually, total page likes (number of page follower) of a company Facebook page are found as the most important factor in order to have more social media engagements both for paid and non-paid publications. We also propose a managerial implication on how to improve the number of engagements in company social media.

Keywords: social media; data mining; paid advertisement; non-paid advertisement; social media engagements.

Reference to this paper should be made as follows: Huang, J-P. and Depari, G.S. (2021) 'Improving social media engagements on paid and non-paid advertisements: a data mining approach', *Int. J. Data Analysis Techniques and Strategies*, Vol. 13, Nos. 1/2, pp.88–106.

Biographical notes: Jen-Peng Huang received his PhD from the Department of Computer Science, University of Oklahoma. Currently, he is a Professor at the Department of Information Management, Southern Taiwan University of Science and Technology. His research interests include data mining, database management and e-commerce.

Genesis Sembiring Depari received his Master degree from National Kaohsiung First University of Science and Technology. He is a full time Lecturer at Universitas Pelita Harapan Medan, Indonesia and currently still pursuing his PhD degree in Department of Business Management, Southern Taiwan University of Science and Technology. His research interests include social media and business data mining.

1 Introduction

Recently, customer engagements are considered more important in research and discussion (Harrigan et al., 2017). This phenomenon is caused by customer engagements being considered as a highly related factor to an indicator of brand success (Bijmolt et al., 2010; Bowden, 2009; van Doorn et al., 2010). In fact, many of those engagements appear online over social media (Malthouse and Hofacker, 2010). With the rapid development of information and technology, those engagements can be easily recorded and converted to valuable information. Unfortunately, the company usually omitted the opportunities and useful insight presented by those active customer engagements (Berthon et al., 2007).

Currently, social media has held a vital rule in bridging communication through many forms, for example, to build a high-quality liaison between the university and their students, the university should put more attention and efforts in social media communications (Clark et al., 2017). On the other hand, recently some companies have put more efforts in developing their online presence with hiring an online content editor, for instance, the *New York Times*, etc. Although many companies understand the vital role of being engaged on social media, they do not really understand how to organise it effectively and measurable (Hanna et al., 2011).

Based on *We Are Social* report (2018), there are 4.021 billion people interact on the Internet nowadays, and 3.196 Billion are active in social media. Need to be noted that Facebook is the most social media used today with more than 2.1 billion users over the world. This number shows a big opportunity in doing advertisement through Facebook. Moreover, Facebook provides a download feature that allows an admin page to download the advertisements data. Therefore, with these useful features, an admin can more understand the performance of each advertisement both for paid and non-paid advertisements. However, the question of how to make advertisements effective is a big issue today. Hence, in this study, we propose a data mining method to extract useful information inside of social media data.

Data mining method can be performed to dig more information regarding customers, people opinions, and finding influential people through leveraging social media data (Barbier and Liu, 2011). Machine learning, gathering information, statistics, database, and data visualisation is part of the data mining method (Larose and Larose, 2014). Therefore, to deal with this social media data, we examined the performance of three machine learning algorithms including support vector machine, random forest and deep learning algorithm.

The purpose of this study is to develop a strategy to improve social media engagements through paid and non-paid publications based on publication characteristics. We tested three potential machine algorithms such as support vector machine (SVM), deep learning (DL) and random forest (RF) in order to get a robust algorithm in dealing

with social media data. After selecting the most powerful algorithm, then we performed the algorithm to analyse the relevance of input variables in supporting the number of social media engagements. We either analyse the correlations of both customer engagements and post characteristics towards a number of people reached. Finally, this research comes up with a managerial implication on how to improve customer's engagement through a company social media brand and its different strategy to have more engagements.

2 Related work

Engagement is a user-initiated action (Gluck, 2013). Other researcher defined engagement as a multidimensional idea including behaviour, cognitive and emotional (Hollebeek, 2011). Social media engagements on Facebook can be expressed by liking, commenting and sharing (Khan, 2017). Those expressions may reflect the customers feeling or emotion regarding the content provided. Gummerus et al. (2012) found those customer engagement behaviours strongly affected by benefit received. The benefits would be the entertainment benefit, economic benefit, and social benefit. Therefore, generating more customers engagements could bring valuable insight into the decision making process.

Previous research in the field of customer engagement put more efforts on the process of engagement (Bowden, 2009), understanding customer to customer interaction (Chan and Li, 2010), customer empowerment (Cova and Pace, 2006.), and Customers attitude in Facebook (Debatin et al., 2009). However, only a few papers using data mining technique as an analysing strategy to understand more online customer engagement both for paid and non-paid publications through Facebook.

Data mining method is an approach that includes several stages such as data understanding, data preparation, modelling and measurement (Han et al., 2011). Machine Learning Algorithms offer a technical source of data mining used to analyse information from raw data (Witten et al., 2016). Moro et al. (2016) used data mining technique to predict the performance metric of a well-known Facebook brand company. Trainor et al. (2014) also employed data mining method in dealing with social media data especially users' inputs and was proven as a useful tool.

3 Methodology

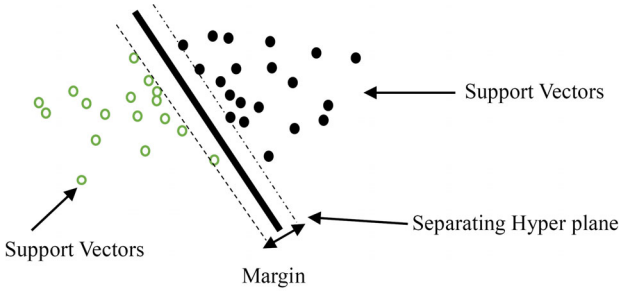
Support vector machine (SVM) is one of the machine learning algorithms. SVM was first introduced and developed by Boser et al. (1992) and Valdimir and Vapnik (1995). SVM classifiers were composed to predict the level or class to which a certain particle belongs, with transforming input space into n-dimensional feature space depends on a kernel (Steinwart and Christmann, 2008). The classification function is described as follows:

$$f(x) = \sum_{i=1}^{N} \propto_i y_i K\left(X, X_i\right) + b \tag{1}$$

$\{(x_i, y_i)\}_{i=1}^{N}$ is a training model with N support vectors (x_1), and the equal class (y_1); $\propto_i$ and b are parameters that adjustable in order to tuning model performance in training data

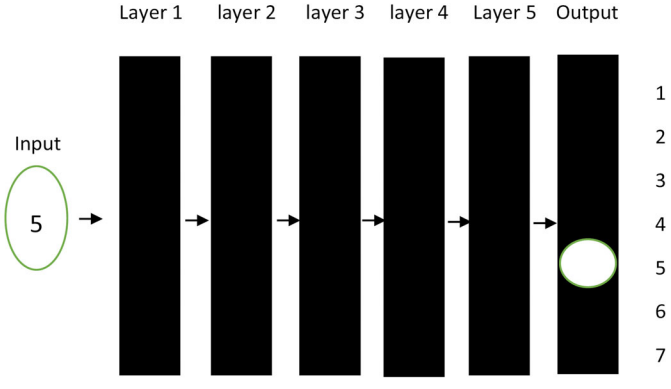
process; then $K(X_1, X_2)$ is named as kernel function. There are some widely used kernel function such, linear function, $k(x, y) = x.y$, polynomial function, $k(x, y) = (x.y + 1)^d$, and radial function, $\exp(-g\|x - y\|^2)$. The complete picture regarding how SVM works, described at figure.1 below. The dots that lie in the dash line are named support vectors, and then the line between dash lines named separating hyperplane. SVM using the kernel function the finding the optimal hyperplane to separate the data. SVM has been proven as a very important algorithm in dealing with many data mining issues (Mangasarian, 2001).

Figure 1 Support vector machine (see online version for colours)



Deep learning (DL) is part of the machine learning field which emphasising on learning representation from data and these learning representations are studied from the *neural network* model (Chollet, 2017). One important aspect of deep learning is its ability in dealing with large, complex, and unlabeled data (Najafabadi et al., 2015). On the other hand, deep learning was found able to produce more important results than the common method in finance (Heaton et al., 2016). This phenomena probably can be implemented to others field such as marketing, management, etc. Generally speaking, deep learning can be divided into three part of a sequence such as input, hidden layers, and output. The sequence is described in Figure 2.

Figure 2 A deep neural network (see online version for colours)



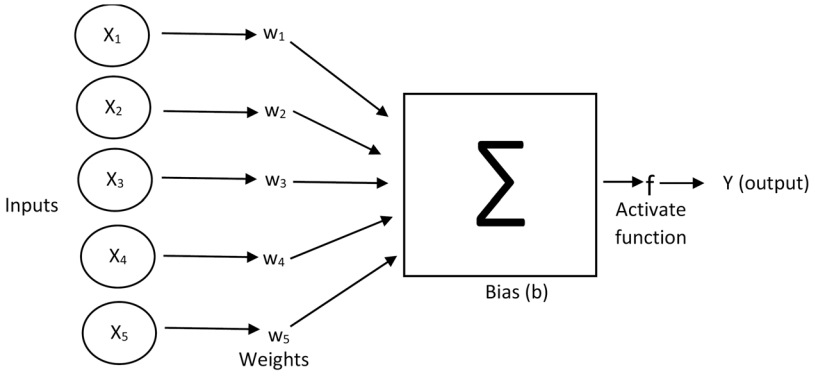
To understand more how deep learning calculation works, we describe the process in Figure 3. Need to be noted that if we only have one hidden layer, then the model is called

a neural network model. If we have multiple hidden layers, then we have a deep learning model (deep neural network). In this study, we run a deep learning model which has more than one hidden layers. In the first layer, there are input variables, then multiply by weight in the next layer. To calculate the output, we described the formula below:

$$y = A\left(\left(I_1.w_1\right)+\left(I_2.w_2\right)+\left(I_3.w_3\right)+\left(I_4.w_4\right)+\left(I_5.w_5\right)+\ldots+\left(I_n.w_n\right)+b\right) \tag{2}$$

Which y is output, A is activation function. There are several famous activation function usually used such binary step function, $f(x) = 1$, $x = \varnothing$, linear function, $f(x) = ax$, sigmoid, $f(x) = \frac{1}{1+e^{-x}}$, tanh, $\tanh(x) = \frac{2}{1+3^{-2x}} - 1$, ReLu, $f(x) = \max(\varnothing, x)$, and leaky Relu, $f(x) = ax$, $x < \varnothing = x$, $x \geq \varnothing$. In this research, we tested the neural network model using 2, 3, 4, 5, 6, and 7 hidden layers, however using four hidden layers found as the most accurate model. To determine the other parameter such as activation function, we performed a grid search method (results can be seen in the discussion section).

Figure 3 Neural network model



Decision tree in data mining method is widely used because of its flexibility and ease of use with single tree analysis. However, a single tree analysis often leads to biases and unstable prediction (Bhattacharyya et al., 2011). Random forests (RF) are a combination of tree estimator which each tree depends on the rate of the stochastic vector using equal distribution for all tree in the forest (Breiman, 2001). Therefore, with using many tresses, Random forest is well known regarding its ability to overcome the overfitting problem. In dealing with customer churn prediction, RF was found as a robust algorithm compared with artificial neural network, decision tree and class weighted core support vector machines (Xie et al., 2009). In line with previous research, RF was also performed better than ordinary linear regression and logistic regression in estimating customer retention and profitability (Larivière and van den Poel, 2005). In this study, we examined three machine learning algorithms along with dynamic parameters in order to have a robust algorithm in predicting social media engagements. Finally, we used a selected algorithm to rank the input importance of company social media data.

4 Experiment

This research is focused on how to improve social media engagements through data mining approach using two company's social media data. The first data was taken from UCI dataset consist of 500 publications of a well-known cosmetic company and the second data was gathered from an educational company from Indonesia consist of 216 publications. The only differences between cosmetic company data (the first dataset) and educational company data (the second dataset) are post categorical and publication strategy were taken by the company. The cosmetic company data using the paid publication as their strategy and put more data in categorical data such action, product, inspiration but for Indonesian company data, categorical publications data is not listed and the strategy to publish the publication only using organic strategy (non-paid publications).

This research was divided into two stages of experiments. The first stage was to examine three machine learning algorithms including support vector machine (SVM), deep learning (DL) and random forest (RF) in predicting social media performance metrics. The second stage was to find the most relevant input variables in improving engagements using the selected machine learning algorithm. The cosmetic company used paid ads strategy in their publications. Conversely, the Indonesian educational company only used non-paid publications (organic strategy) as its publication strategy. Therefore, we analysed the input variables importance both for paid publications and non-paid publications using the selected algorithm. The data is described in Table 1.

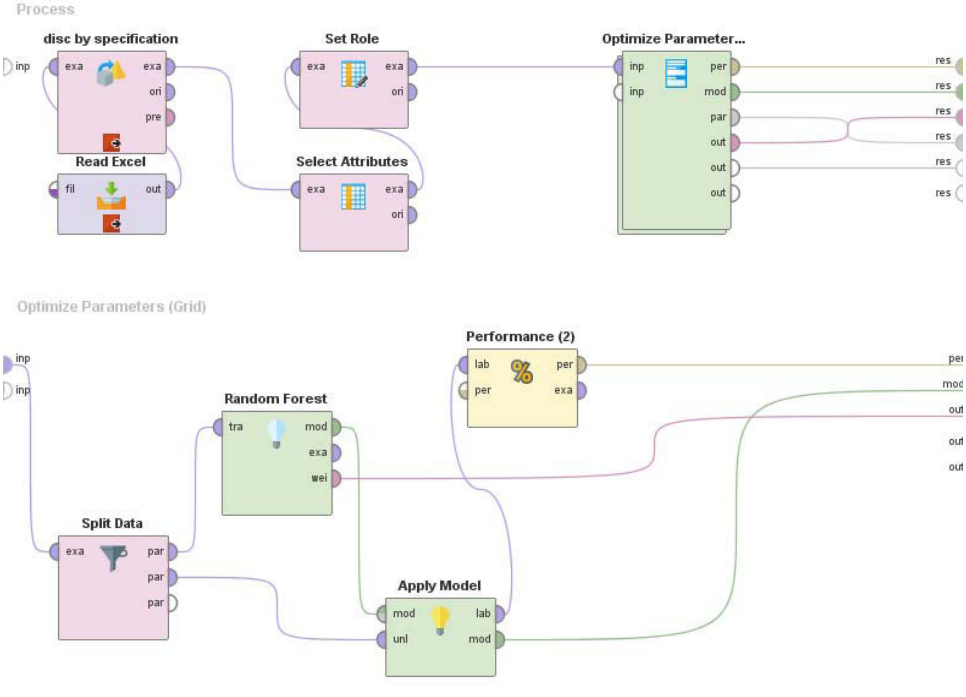
Table 1 Features

Feature	*Role*
Pages total likes	Input feature
Type	Input variables (photo, status, link, and video)
Category	Input variables (action, product, and inspiration)
Post month	Input variables (January–December)
Post weekday	Input feature (Sunday–Saturday)
Post hour	Input feature (1–24)
Paid	Input feature (1, 0)
Comments	Label or dependent variable
Likes	Label or dependent variable
Shares	Label or dependent variable
Total interactions (comments + likes + shares)	Label or dependent variable

In order to perform a data mining approach using a machine learning algorithm, we employed RapidMiner as a predictor tool. The data mining sequences through RapidMiner are shown in Figure 4. First, we feed in the data using Excel read package then we perform preprocessing data by divided the total interaction into two major groups such as low and high level of engagements. We divided the total interaction based on expert recommendation. The publication which has 0–240 engagements are considered as low publication engagements and the publication which has more than 240 engagements are considered as high publication engagements. Within this section also, we normalise

the input variables using the z-transformation technique. Second, in order to avoid overfitting problem, the data were divided into two groups such as test (25%) and training (75%) data using stratified sampling method then the results of the prediction value was compared with test data to have the model accuracy. Before predicting the model performance we optimise the algorithm parameters by conducting a grid search method.

Figure 4 Data mining sequence using RapidMiner (see online version for colours)



5 Result and discussion

5.1 Algorithms comparison

Grid search method was applied in obtaining optimum parameters of random forest. In order to reach high accuracy, a combination between 41 maximal depth and 11 number trees were found as the most optimum parameters. The other combination that has equal accuracy was 80 maximal depth and 100 number of trees. However, using more tree usually spends more times in processing. Therefore, in this research, we used 41 maximal depth and 11 trees to predict a number of engagements. Based on using the optimum parameters above, we got 79.00% of overall accuracy. In predicting high engagements, random forest reaches 58.33% accuracy and 81.82% in predicting low engagements.

Table 2 Grid search of random forest parameters

Maximal depth	*Number of trees*	*Accuracy*
41.0	11	0.79
80	100	0.79
60	70	0.78
1	1	0.77
1	11	0.77
1	21	0.77
1	31	0.77
100	21	0.77
1	41	0.77
1	51	0.77

Table 3 Grid search of SVM parameters

Kernel function	*Accuracy*
Polynomial	0.77
Rbf	0.76
Sigmoid	0.75
Precomputed	0.23

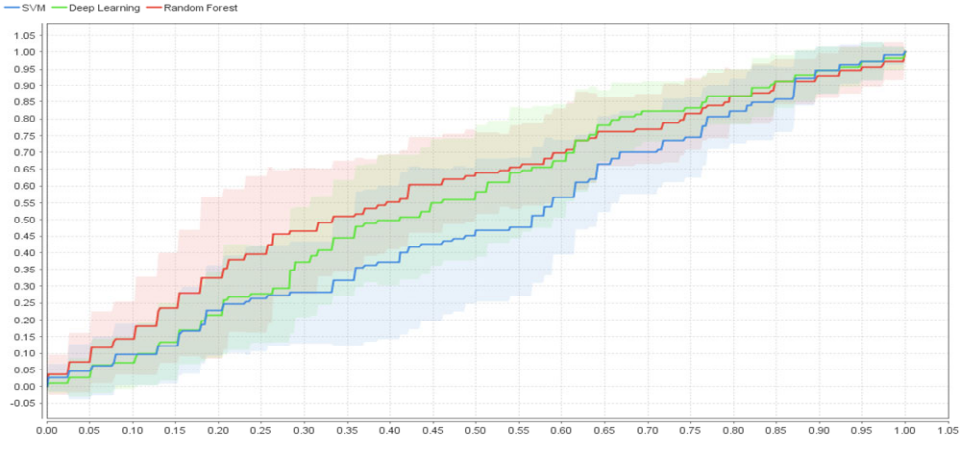
In using support vector machine, we tested 4 kernel functions such as polynomial, Rbf, sigmoid and precomputed. Based on these experiments, we found that the polynomial kernel function works better in term of accuracy (Table 3). Therefore, we used this activation in dealing with SVM prediction. We got 77% of overall accuracy in predicting the number of engagements. The third candidate algorithm was deep learning. Before performing the prediction, we also examined four activations in order to get a better prediction performance. Eventually, we found that tanh activation performed better than other activations function with 71%accuracy. Therefore, using this activation, deep learning able to reach 71% accuracy. Based on three examined algorithms, we conclude that in predicting the engagements, random forest works better than two other algorithms. However, in term of time processing, Random Forest also spends more time than two other algorithms (Table 4). Based on the different of testing and training accuracy, we can conclude that the model also free from overfitting problem.

Table 4 Algorithm comparison

	Overall accuracy (using training data)	*Overall accuracy (using test data)*	*Precision (pred high)*	*Precision (pred low)*	*Time (S)*
Support vector machine	77.5%	77%	0%	77%	1
Random forest	99%	79%	58.33%	81.82%	8
Deep learning	74.5%	71%	39.29%	83.33%	4

To have broader information regarding algorithm performance comparison, we assess the algorithm performance by drawing a ROC curve (Figure 5). Based on Figure 5, we can see that most of the time random forest works better than two other algorithms. When false positive rate (horizontal line) is low, random forest shows its robustness compared to deep learning and SVM. It means random forest has more true positive rate (vertical line) than SVM and deep learning. Therefore, we can conclude that in dealing with this social media data, random forest performs better than SVM and deep learning. Furthermore, we used random forest weight to analyse the importance of publications characteristics in order to improve the engagements both for publications with and without paid ads.

Figure 5 Roc curve of random forest, SVM and deep learning (see online version for colours)



5.2 Improving paid ads performance

We used Cosmetic company data to measure the performance of paid ads (data was explained in Section 3). The dependent variable is total interaction which is an accumulation of share, like and comment. The inputs variable are total page likes, post day, post month, categories, type, post hour, and paid advertisements (Table 1). Based on expert recommendation (social media analyst), we grouped total engagements into two categories such low level of and high level of engagements. Low level of engagements was range from 0 to 240, and a high level of engagements was range from 241 to 6334. Furthermore, we performed random forest (RF) to rank the relevance of input variables. Before we run random forest to analyse the importance of input variables, we employed the grid method to select the optimum parameters. The results of the grid search are shown in Table 5. The optimal maximal depth and number of trees were found as many 41 and 11. Therefore, we applied the results to the random forest model.

Table 5 Results of grid search

Maximal depth	*Number of trees*	*Accuracy*
41	11	0.79
80	100	0.79
60	70	0.78
.	.	.
.	.	.
.	.	.
100	1	0.63
11	1	0.6
41	1	0.58

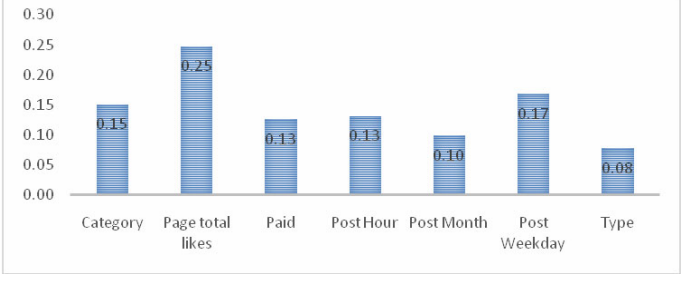
After applying the optimal parameters, we generated the confusion matrix result (Table 6). The overall accuracy is 79% and to predict the low and high number of engagements, random forest achieves 81.82% and 58.33% accuracy.

Table 6 Confusion matrix of paid publications

	True low	*True high*	*Class precision*
Pred. low	72	16	81.82%
Pred. high	5	7	58.33%
Class recall	93.51%	30.43%	

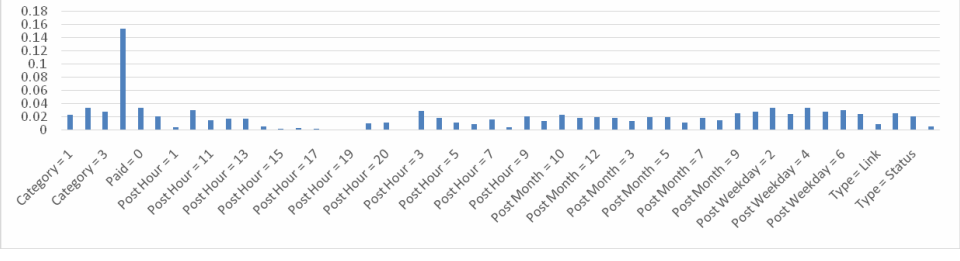
Based on the results of random forest analysis, page total likes was found as the most relevant factor that affects the number of engagements compared to the category of the publication, paid ads strategy, post hour, post month, post weekday, and type of publication (Figure 6). Paid ads strategy was found less relevance in attracting people to engage with only 13% relevant. These results are supported by the research conducted by Lee et al. (2018). They discovered publications which have humour and emotion inside likely create more engagements. Barreto (2013) also found that paid ads work poorly to catch user attention compared to a friend's recommendation on Facebook. Bacík et al. (2015) explored that using paid ads strategy on Facebook doesn't guarantee to have successful promotion activity. Therefore, this study concludes that In order to have more customer engagements in company social media brand, a manager has to pay more attention to have more page total likes/followers in advance.

Figure 6 Percentage of relevance inputs toward a number of engagements (paid publications) (see online version for colours)



In order to have deeper analysis regarding what category, post hour, post month, post weekday, and type that affect a number of engagements the most, we also analyse the importance of sub-inputs to support the number of engagements to each of paid publication. The results are shown in Figure 7. Category 2 (promoting product) is found as the most important factor compared to category 1 (action) and category 3 (inspiration) publications. In term of post hour, posting a publication on 10 am is found as the most relevance time. For the type of publication, publications that using a photo works better than a publication using link or video. The most important point here is that total page likes of the Facebook page are the most important factor in supporting more publications engagements. This insight was proven by the result of variable input importance which is total page likes has a dominant influence compared to other input variables. Therefore, we can conclude that in supporting the number of publication engagements through paid publication, having more page total likes is a must.

Figure 7 The importance of sub-inputs and inputs (paid publications) (see online version for colours)



5.3 Improving organic performance (non-paid ads)

The same treatment was conducted to measure the performance of non-paid ads. The random forest was employed to analyse the importance of each input variables in supporting the number of engagements. Before running the Random forest algorithm, we performed a grid search method in order to have optimal parameters. The parameters are a number of trees and maximal depth towards accuracy, it means more accurate a prediction is better. The results of the grid search are shown in Table 7. The optimal maximal depth and number of trees were found 60 and 41. Therefore, we applied this optimal value to random forest architecture.

The accuracy achieved after implementing the optimal parameters is shown in Table 8. Overall prediction accuracy is 86.05%. In order to predict low engagements, the performance of the model achieves 81.25% accuracy. The most powerful model's performance was reached by predicting the high level of engagements with 100% accuracy. The complete results are described in Table 8.

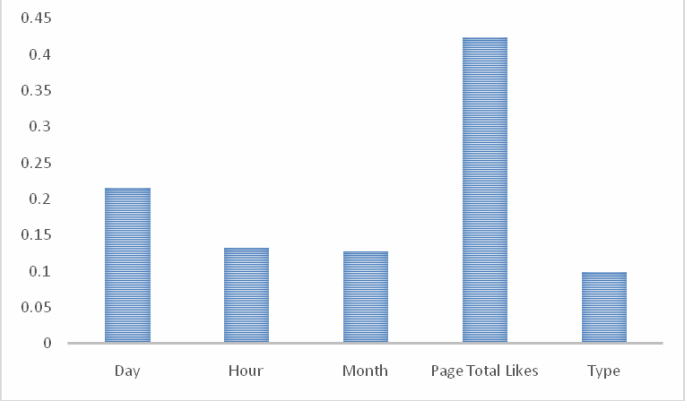
Table 7 Grid search results

Maximal depth	*Number of trees*	*Accuracy*
60	41	0.860465116
51	51	0.860465116
41	70	0.860465116
11	31	0.837209302
.	.	.
.	.	.
.	.	.
1	90	0.604651163
1	100	0.604651163
41	1	0.581395349

Table 8 Confusion matrix of non-paid publications

	True low	*True high*	*Class precision*
Pred. low	26	6	81.25%
Pred. high	0	11	100.00%
Class recall	100.00%	64.71%	

Figure 8 Percentage of relevance inputs toward a number of engagements (paid publications) (see online version for colours)

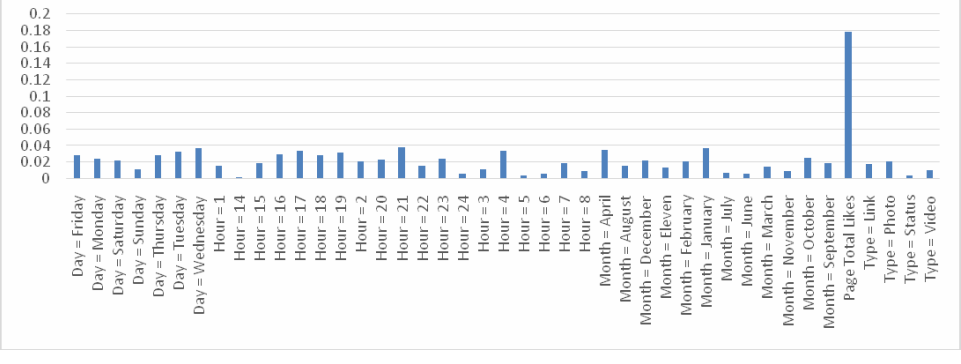


The results of the input variables importance are shown in Figure 8. In supporting the number of engagements of non-paid publications, page total likes also has the biggest relevance compared to other input factors. These results emphasise important information. In improving the number of engagements through Facebook page either with paid and non-paid publications, page total likes hold a very important role. Having more page total likes probably can improve the possibility to have more engagements in publications either through paid or non-paid strategy. This new insight strongly suggests the company Facebook page to put more efforts into attracting people to like their Facebook page. Need to keep in mind that whoever likes a Facebook page will

automatically become the follower of the Facebook page. Whoever becomes the follower of one Facebook page, will continuously see the Facebook page publications in their dashboard either through paid and non-paid publications strategy.

In order to have deeper analysis on what day, hour, type of publication and month that support the number of engagements, we go deeper to analyse the importance of sub-inputs and input variable using the Random forest. The results are shown in Figure 9. For non-paid publications, Wednesday is found as the most effective day in publishing a post. Therefore, publishing a promotion product on Wednesday probably having more engagements than other weekdays. For hour post, 21 pm is found as the most effective time to publish a post. Posting publications along with a picture also found working better than using a link or video in publications. However, the contribution is not as strong as having more page total likes in Facebook page. These results are confirmed by the influence gap between page total likes and the other input variables.

Figure 9 The importance of sub-inputs and inputs (non-paid publications) (see online version for colours)



5.4 Correlation

To prove that post engagement has a positive and significant correlation to a number of reached people, we run a correlation analysis of shares, likes, comments, post type, post hour, post weekday, post month, page total likes, and category, both for paid and non-paid posts data. Finally, on paid posts data which is from the cosmetic company found that customer engagements have a strong connection to a number of reach people. Likes, shares, and comments have 0.55, 0.46 and 0.43 correlation which is higher than the other inputs. Besides, the other input factors show a slight correlation toward a number of people reached. These results bring a very important insight that a number of customer engagements have a strong correlation with the number of people reached. Again, these results confirm that customer engagements are valuable because not only showing customers reaction regarding a post but also have a high correlation with a number of people reached.

Table 9 Correlation towards lifetime post total reach through the paid post

	Lifetime post total reach	*Like*	*Share*	*Comment*	*Page total likes*	*Type integer*	*Category*	*Post month*	*Post weekday*	*Post hour*	*Paid*
Lifetime post total reach	1										
Like	0.55	1.00									
Share	0.46	0.90	1.00								
Comment	0.43	0.84	0.87	1.00							
Page total likes	–0.08	0.05	–0.01	0.03	1.00						
Type integer	0.14	–0.04	0.01	–0.01	0.08	1.00					
Category	–0.14	0.13	0.15	0.03	–0.09	–0.18	1.00				
Post month	–0.10	0.02	–0.03	0.01	0.94	0.13	–0.13	1.00			
Post weekday	–0.05	–0.08	–0.05	–0.08	–0.01	0.02	–0.05	–0.02	1.00		
Post Hour	0.00	–0.02	–0.06	0.00	–0.14	–0.08	–0.11	–0.18	0.05	1.00	
Paid	0.15	0.11	0.08	0.08	0.01	0.03	–0.02	0.02	0.00	–0.07	1

Table 10 Correlation towards lifetime post total reach through the organic post

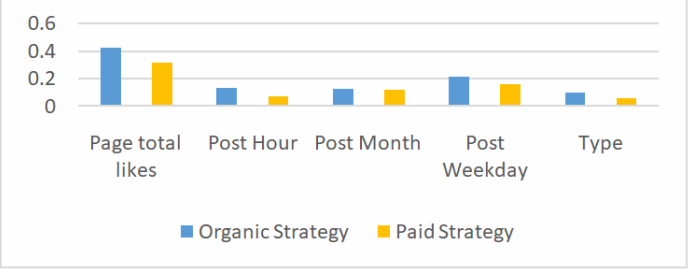
	Lifetime post total reach	*Like*	*Share*	*Comment*	*Page total likes*	*Hour*	*Type*	*Month*	*Day*
Lifetime post total reach	1.00								
Like	0.40	1.00							
Share	0.43	–0.08	1.00						
Comment	0.22	0.50	0.25	1.00					
Page total likes	0.27	0.42	0.16	0.17	1.00				
Hour	0.12	–0.02	0.15	0.01	0.00	1.00			
Type	–0.33	–0.40	–0.21	–0.20	–0.77	–0.05	1.00		
Month	0.00	–0.51	0.21	–0.11	–0.37	0.11	0.30	1.00	
Day	0.03	0.04	–0.11	0.09	–0.04	0.00	0.02	0.09	1.00

Similar results are shown by non-paid data (organic data). These data come from the Indonesian educational company that didn't use paid ads as their strategy. All of the posts published using organic strategy through their company Facebook pages. The results showed that likes, shares, comments and page total likes hold a high correlation towards a number of people reached. Within organic posts, the share was found with the highest correlation toward a number of people reach. This revealed important insight into the importance of shares through organic posts. Page total likes were also found highly correlated to a number of people reach that different to paid post that mentioned before. Page total likes found more important to reach more people in organic publications compared to paid publications. Therefore, these results are consistent with random forest importance results. Both random forest and correlation results emphasise that page total likes are an important input in order to reach more people and boost customer engagements.

5.5 Managerial implication

Since categorical of publications are not listed in the Indonesian educational data, we compare only the identical attributes. Thus, the cosmetic company data will not provide categorical data in this section. The comparison between the importance of input variables with paid and non-paid publications are shown in Figure 10. The comparison graph below provides very important information to the company that leveraging their existence in the online Facebook page. For organic publications, page total likes hold much more important rule than paid publications. It is proven by the percentage of relevance in supporting the number of engagements with 42% influence and 31.8% for paid publications. Therefore we conclude that, if a company only rely on organic strategy, the company should put more efforts into acquiring more people to like its company Facebook page. The second important factor is post weekday. Both for paid and non-paid publications, weekday of publications occupied the second relevance position after page total like. In term of publications type, it is found as least relevance to support the number of Facebook pages engagements both paid and organic publications. The other important result is the different relevance result between hour post of paid publication and non-paid publications. In paid publications, post hour occupies the fourth position, but inorganic publications, post hour occupies the third position in importance. Therefore we conclude that, if we want to use organic strategy as our strategy, then we should put more focus on analysing the most appropriate time.

Figure 10 Comparison of input variable importance (see online version for colours)



6 Conclusions and future research

Having more engagements on social media is important. With advance rapidly technology in information management nowadays, the owner of a Facebook page is allowed to download and gathering the full information of their company social media engagements. This feature probably will help the company to understand more their customers based on their engagements through company publications. Therefore in this study, we try to find the most important factors that affect and help the publications to have more engagements. Beside it, we either analyse the correlation of both customer engagements and post characteristic towards a number of people reached. We conclude several important results listed below:

1 Paid Advertisement through Facebook has a less relevance contribution in order to improve the number of social media engagements, conversely, page total likes are found more relevance in supporting the number of engagements. Therefore, having more page total likes on the Facebook page is important.

2 Paid and non-paid publications through two Facebook pages with different company and background likely require a high number of page total likes in order to achieve more engagements.

3 Organic publications require more effort in selecting the best time to post, it is proven by the relevance of hour in organic publications much bigger than the hour relevance of paid publications.

4 Type of publications has the least relevance both for paid and non-paid publications in supporting the number of Facebook page engagements.

5 In term of paid publications, the category of publications plays a more important rule that the type of publications. Therefore, to have more engagements, selecting the suitable categorical of publications probably acquiring more engagements.

6 In term of correlation, customer engagements in paid posts were found holding a positive and high correlation toward a number of people reached, therefore these results confirm the importance of having more customer engagements in company Facebook page.

7 In organic publications, not only customer's engagements were found having a high correlation. Page total likes also found highly correlated with a number of people reached. Therefore, this study concludes that organic posts are required to have more page total likes or followers in their company Facebook page in order to have more people reached and more customers engaged.

Finally, this study strongly suggests a company to put more efforts to have more page total like (follower) in its company Facebook pages. Therefore the question of how to get more Facebook company page (followers) through paid and non-paid post probably good area for further research.

References

Bacík, R., Fedorko, R., Kakalejcík, L. and Pudlo, P. (2015) 'The importance of Facebook ads in terms of online promotion', *Journal of Applied Economic Sciences*, Vol. 10, No. 5, p.35.

Barbier, G. and Liu, H. (2011) 'Data mining in social media', *Social Network Data Analytics*, pp.327–352, Springer, Boston, MA.

Barreto, A.M. (2013) 'Do users look at banner ads on Facebook?', *Journal of Research in Interactive Marketing*, Vol. 7, No. 2, pp.119–139.

Berthon, P.R., Pitt, L.F., McCarthy, I. and Kates, S.M. (2007) 'When customers get clever: managerial approaches to dealing with creative consumers', *Business Horizons*, Vol. 50, No. 1, pp.39–47.

Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C. (2011) 'Data mining for credit card fraud: a comparative study', *Decision Support Systems*, Vol. 50, No. 3, pp.602–613.

Bijmolt, T.H., Leeflang, P.S., Block, F., Eisenbeiss, M., Hardie, B.G., Lemmens, A. and Saffert, P. (2010) 'Analytics for customer engagement', *Journal of Service Research*, Vol. 13, No. 3, pp.341–356.

Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992) 'A training algorithm for optimal margin classifiers', in Haussler, D. (Ed.): *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp.144–152, ACM Press, New York, NY.

Bowden, J.L.H. (2009) 'The process of customer engagement: a conceptual framework', *Journal of Marketing Theory and Practice*, Vol. 17, No. 1, pp.63–74.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Chan, K.W. and Li, S.Y. (2010) 'Understanding consumer-to-consumer interactions in virtual communities: the salience of reciprocity', *Journal of Business Research*, Vol. 63, Nos. 9–10, pp.1033–1040.

Chollet, F. (2017) *Deep Learning with Python*, Manning Publications Co, USA.

Clark, M., Fine, M.B. and Scheuer, C.L. (2017) 'Relationship quality in higher education marketing: the role of social media engagement', *Journal of Marketing for Higher Education*, Vol. 27, No. 1, pp.40–58.

Cova, B. and Pace, S. (2006) 'Brand community of convenience products: new forms of customer empowerment – the case 'My Nutella the Community', *European Journal of Marketing*, Vol. 40, Nos. 9–10, pp.1087–1105.

Debatin, B., Lovejoy, J.P., Horn, A.K. and Hughes, B.N. (2009) 'Facebook and online privacy: attitudes, behaviors, and unintended consequences', *Journal of Computer-Mediated Communication*, Vol. 15, No. 1, pp.83–108.

Gluck, M. (2013) *Digital Ad Engagement: An Industry Overview and Reconceptualization*, Interactive Advertising Bureau.

Gummerus, J., Liljander, V., Weman, E. and Pihlström, M. (2012) 'Customer engagement in a Facebook brand community', *Management Research Review*, Vol. 35, No. 9, pp.857–877.

Han, J., Pei, J. and Kamber, M. (2011) *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam, Netherlands.

Hanna, R., Rohm, A. and Crittenden, V.L. (2011) 'We're all connected: the power of the social media ecosystem', *Business Horizons*, Vol. 54, No. 3, pp.265–273.

Harrigan, P., Evers, U., Miles, M. and Daly, T. (2017) 'Customer engagement with tourism social media brands', *Tourism Management*, Vol. 59, pp.597–609.

Heaton, J.B., Polson, N.G. and Witte, J.H. (2016) *Deep Learning in Finance*, arXiv preprint arXiv:1602.06561.

Hollebeek, L. (2011) 'Exploring customer brand engagement: definition and themes', *Journal of Strategic Marketing*, Vol. 19, No. 7, pp.555–573.

Khan, M.L. (2017) 'Social media engagement: what motivates user participation and consumption on YouTube?', *Computers in Human Behavior*, Vol. 66, pp.236–247.

Larivière, B. and van den Poel, D. (2005) 'Predicting customer retention and profitability by using random forests and regression forests techniques', *Expert Systems with Applications*, Vol. 29, No. 2, pp.472–484.

Larose, D.T. and Larose, C.D. (2014) *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, USA.

Lee, D., Hosanagar, K. and Nair, H.S. (2018) 'Advertising content and consumer engagement on social media: evidence from Facebook', *Management Science*, Vol. 64, No. 11, pp.5105–5131.

Malthouse, E. and Hofacker, C. (2010) 'Looking back and looking forward with interactive marketing', *Journal of Interactive Marketing*, Vol. 24, No. 3, pp.181–184.

Mangasarian, O.L. (2001) 'Data mining via support vector machines', *IFIP Conference on System Modeling and Optimization*, July, pp.91–112, Springer, Boston, MA.

Moro, S., Rita, P. and Vala, B. (2016) 'Predicting social media performance metrics and evaluation of the impact on brand building: a data mining approach', *Journal of Business Research*, Vol. 69, No. 9, pp.3341–3351.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E. (2015) 'Deep learning applications and challenges in big data analytics', *Journal of Big Data*, Vol. 2, No. 1, p.1.

Steinwart, I. and Christmann, A. (2008) *Support Vector Machines*, Springer Science & Business Media, Germany.

Trainor, K.J., Andzulis, J.M., Rapp, A., and Agnihotri, R. (2014) 'Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM', *Journal of Business Research*, Vol. 67, No. 6, pp.1201–1208.

Valdimir, V. and Vapnik, N. (1995) *The Nature of Statistical Learning Theory*, Springer Science & Business Media, Berlin, Germany.

Van Doorn, J., Lemon, K.N., Mittal, V., Nass, S., Pick, D., Pirner, P. and Verhoef, P.C. (2010) 'Customer engagement behavior: theoretical foundations and research directions', *Journal of Service Research*, Vol. 13, No. 3, pp.253–266.

We Are Social (2018) *Global Digital Report 2018*, Erisim [online] https://wearesocial.com/blog/2018/01/global-digital-report-2018.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. (2016) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, USA.

Xie, Y., Li, X., Ngai, E.W.T. and Ying, W. (2009) 'Customer churn prediction using improved balanced random forests', *Expert Systems with Applications*, Vol. 36, No. 3, pp.5445–5449.