Social Media Content Analysis Study through Multiple Linear Regressions & Topic Modelling Analysis

Anita, Hoe Xiu Ming and Sallyana
Undergraduate in Singapore Management University

Abstract

In the recent years, the popularity of social media platforms has increased tremendously and this has driven the rise of social media advertising industry. While many have attempted to measure the performance of social media contents, few have investigated the factors affecting the performance of these contents. These factors may potentially help social media content creators such as SGAG in improving its contents and thereby, increasing the performance of its business. Therefore, a study on factors affecting the Key Performance Indicator (KPI) of social media content is conducted through Multiple Linear Regression Analysis, as well as Topic Modelling Analysis in this paper.

The paper mainly focuses on the analysis of Facebook post's performance and the KPIs that have been examined comprises of Facebook Post's Reach and Number of Video Viewers. In addition to this, the examination of topics used in Facebook post message and its performance have also been carried out in this paper. For these purposes, two multiple linear regression models and a topic modelling analysis have been developed and results have shown that various factors such as Main Engagement, Negative Feedbacks, Hide All Clicks Action, Unlike Page Action and Type are significant in explaining the variability of post reach, while components like Number of Shares, Unlike Page Clicks, as well as Engagement of Fans significantly affect the variability of the video viewers. Topic Modelling Analysis has also shown that there are eight commonly used topics in the Facebook posts and among these, the top three popular topics were "Events", "Transportation" and "Relationships".

Introduction

In this information age, the use of social media platforms such as Facebook, Twitter and Instagram is ubiquitous and users of these platforms have been increasing in the recent years. This is seen in a research published by Smart Insights (Marketing Intelligence) Ltd, where the number of global active social media users rise 10% to 2.307 Billion from January 2015 to January 2016. The growing adoption of social media has also led to the increase usage of social media marketing and this trend is particularly true for Facebook, a social media platform that is leveraged by 92% of social marketers (Saleh, 2015).

According to an article published by Invesp in 2015, it was estimated that there was 1.39 Billion of active Facebook users per month and there was more than 30 Million Businesses that have a Facebook fan page. Despite the substantial number of business Facebook fan pages, many companies choose to tap on the enormous consumer base of social media content creators for advertising purposes instead of building their own fan page. This has resulted in the increment in social media content creators such as SGAG, TheSmartLocal and SMRT Feedback by The

Vigilanteh which specialize in generating interesting content, as well as paid content for its clients.

Even though the success of these companies have been witnessed by many, the competitiveness of the industry has made it a necessity for them to constantly improve their strategies in order to maintain their competitive advantages. Many times, SGAG faces difficulty in identifying factors affecting their own performance and in continually generating interesting topics to engage its consumers. As such, this paper will aim to discover the important components affecting the performance of their social media posts using multiple linear regressions models, as well as examining the various subjects which help in maintaining their consumer base through the use of topic modelling technique.

This paper focuses on five different sections. Firstly, it begins with the Literature Review which describes the challenges faced by the company and the effectiveness of both abovementioned models in tackling such issues. Secondly, the Methodology section comprises of the description of data to be used for the development of the methods, the derivation process of such models, as well as results uncovered through this process. Next, the Discussion section will highlight the benefits, limitations and assumptions of the methods. Then, the Possible Future Work section will attempt to discuss about the possible extension of the works done in this paper and lastly, the paper will end with a conclusion.

Literature review

Multiple Linear Regression Analysis

Key Performance Indicators (KPI) is often used in the business world to measure the performance of the business to achieve its strategic goals and objectives. SGAG identified Reach as its Facebook posts' KPI and number of Video Viewers as the KPI for video posts. Even though the company is aware of the importance of these KPIs, SGAG faces difficulty in identifying the underlying factors affecting the KPIs.

SGAG provides organic content to large audiences and also provide paid content that is advertisements of its clients. Reach is defined as the number of audience who saw a particular post and it is an important measure to marketers when it comes to paid advertising (i-SCOOP, n.d.). Audience engagement happens when your post attracts your followers to hit a button on like, comment or share. Each click of like, comment, or share increases the possibility of the post showing up in their follower's newsfeed thus increases the reach.

The engagement is highly affected by content of the post (Ramachandran & Balakrishan, 2015). An engaging content strategy is not about proving content you think is good, but to understand audience preferences and tailor the content of the posts to meet their interests (i-SCOOP, n.d.). Hence, to remain competitive in the industry and to gain potential clients, it is important for SGAG to understand the factors affecting post reach and engagement in order to increase its KPIs. The following section will attempt to describe the Multiple Linear Regression method that will help in identifying these factors.

Method

Regression analysis is used to measure the relationship of between response or dependent variable and explanatory or independent variable(s) (Boston University, 2013). In regression analysis, the response variable is denoted by “y” and explanatory variable is denoted by “x”. The correlation coefficient (ranges between -1 to +1) quantifies the strength and direction of relationship between the explanatory and response variables. The variables can be positively or negatively correlated and high correlation coefficient indicates strong relationship between the variable.

A highly engaged content increases post reach, thus factors affecting the KPIs include but not limited to likes, comments and shares. A simple linear regression identifies the relationship between two variables. With multiple factors identified, Multiple Linear Regression will be used to model the relationship between the a continuous dependant variable and the independent variables (Statistics Solution, 2015). The independent variables can be continuous or categorical. To predict the relationship between the response variable and multiple explanatory variables, the multiple regression formula is formed. The formed formula is as follows:

$$\mathbf{Y = b_0 + b_1X_1 + b_2X_2 + \dots\ b_kX_k}$$

Y – Value of the response/dependent variable
b_0 – Intercept
$B_{1..k}$ – Coefficients of the explanatory/independent variable
$X_{1..k}$ – Value of the explanatory/independent variable

In addition, R^2/R-Squared will be used to quantify the performance of the regression model. The value of R^2 ranges between 0 to 1. As R-Squared of the regression explains the variation in the dependent variable that is able to be explained by the independent variables, high R-Squared indicates high accuracy of the model.

Topic Modelling

Topic Modelling is often used to analyse unstructured text, such as social media status or comment fields and it allows businesses to transform these unstructured data into meaningful insights and they are able understand the underlying patterns of the texts.

As SGAG provides social media content, the number of audiences’ reach is important to them and topics used in these contents may potentially affect the reach of the post as well as the engagement of the post. The company generates a few posts per day and thousands of posts per year on different topics of the current happenings, as well as on advertised contents for its clients from various industries. Hence, many different kind of topics can be seen in its posts and overtime, the company has lost track of their post topics and would like to find out the performance of different topics. Therefore, Topic Modelling Analysis would help in discovering the different topic available in their posts, as well as examining the popularity of these topics in increasing its KPIs. The following section describes how Topic Modelling can be performed in order to achieve these goals.

Method

Topic Modelling across different platforms often involves two tasks, namely curating and analysing. Firstly, Topic Modelling presents different terms contained in the data which is similar to a word in a sentence. Secondly, curating the list of terms includes recoding terms, adding terms as a stop word and adding stemming option which essentially groups terms such as stop, stopping or stopped under the same term called "stop". While curating the data, a decision of whether a term is under stop words, have to be recoded, is considered as a phrase or should remain unchanged have to be made. In other words, with curating, the number of relevant terms in the dataset for text analysis will decrease. Afterwhich, analysis of the list of terms based on the Document Term Matrix have to be performed. Each row in the data table corresponds to a row of the Document Term Matrix. The matrix entries contain the frequency of the term occurrence across different row. Lastly, frequency-inverse document frequency (TF IDF) weighting will be used to show the importance of a word to the set of documents.

After completing these two tasks, topic words report that shows the categorization of different terms will be generated. With the help of insights generated from the report, analysts will then decide whether there is a need of having another iteration of curating and analysing. In order to obtain a good model for text analysis, multiple iterations are needed. Once the desirable outcome is achieved from the numerous iterations, the naming of various topic categories generated will have to be decided.

Methodology

Multiple Linear Regression Analysis

Multiple linear regression analysis allows the use of more than one independent variable in examining the relationship between these independent variables and the dependent variable (Uyanık & Güler, 2013). There are various platforms available to assist in constructing multiple linear regression model and in this paper, JMP Pro will be the main software used in assessing the association between the KPI and independent factors affecting it. Fit Model platform of JMP Pro provides various reports of model's performance to assist in the development of multiple linear regression model and it will be used as the platform to specify the different models that will be discussed subsequently in this paper.

In developing these models, variables that may potentially influence the KPI will firstly be identified. Then, the continuous independent variables and dependent variable will be indicated through the Fit Model platform and the platform will then presents the variance inflation factor (VIF). VIF is widely used to measure the degree of multicollinearity between the explanatory variables (O'Brien, 2007). In a multiple regression model, multicollinearity occurs when multiple independent variables are highly correlated with each other and it may increase the inaccuracy of the regression model due to bias within regression coefficient and high standard error (British Dental Journal, 2005). In this paper, VIF of higher than eight will be considered as high multicollinearity and variables with VIF of more than eight will be further re-considered through the following analytical methods.

Possible methods to be used in eliminating factors that causes multicollinearity includes Principal Component Analysis (PCA) and Variables Clustering. Traditionally, PCA have been used for variable reduction by creating a new set of components that is a weighted linear combinations of the original variables (Sanche & Lonergan, 2006). However, original values of variables are lost after the PCA process and it is difficult to interpret the new components created. Hence, Variable Clustering method is prefered for reducing high multicollinearity variables in this paper. The procedure starts with all variables in one cluster, then, a cluster is chosen for splitting based on the smallest percentage of variation explained by its cluster component, the procedure will then repeats until the maximum number of clusters is attained or a certain percentage of variation is achieved (Sanche & Lonergan, 2006). Variable Clustering is available on JMP Pro, and variables with VIF of more than eight obtained through earlier steps will be eliminated through this method and one most representative variable of each cluster will be presented at the end of the clustering procedure. These variables will then be used in the multiple linear regression model.

After re-specifying the continuous variables chosen for the model through variable clustering, variables with high p-value (Prob > |t|) as indicated in the Parameter Estimates report of Fit Model Platform will then be eliminated. P-value determine the significance of the model results and strength of evidence (Rumsey, 2009). The p-value is a number between 0 and 1, where small p-value indicates strong evidence and large p-value indicates weak evidence. Therefore, variables with p-value that is larger than 0.05 will not be used in the models of this paper as it is insignificant in determining the relationship with the dependent variable. Subsequently, categorical independent variables will be added to the model and similarly, negligible variables with high p-value will be eliminated. With these steps, a representing equation of the model will finally be obtained. Lastly, to visualize the impact of each independent variable on the dependent variable, Profiler of JMP Pro is used as it provides the ability to view the degree of sensitivity of the model in response to the modifications made to individual factors.

Evaluation of Factors Affecting Facebook Post KPI using Multiple Linear Regression

Problem Definition

In measuring the performance of SGAG's Facebook posts, the reach of the post which is defined as the total number of users who have seen a particular post is commonly used among the social media content creators. Nevertheless, it is challenging for them to determine the various components that may possibly affect this KPI with merely human judgement. Thus, an attempt to utilize the multiple linear regression analysis in examining these factors influencing the KPI will be discussed in the following section.

Dataset

In the preparation of variables to be used for the model, several potential factors available in SGAG's Facebook post's dataset are firstly identified. In addition to this, other probable components such as time category, weekday or weekend, public holiday, as well as post message length have been derived in order to enhance the accuracy and comprehensiveness of the model. As multiple linear regression analysis requires values to be normally distributed, log transformation has been performed on continuous variables to reduce skewness in the dataset and to prevent some factors of higher values to be dominant in the model. The complete dataset resulted from this selection and transformation process and its accompanying definition can be seen in Figure 1 below.

Factors	**Definition**
Post Message Length	Post message is the description of a Facebook post and this variable specifies the total number of characters used in the post message
Lifetime Engaged Users	Number of unique users who have interacted with a Facebook post by liking, commenting or sharing the post, as well as by liking comments, replying comments of the posts and interacting with the Facebook page
Lifetime People Who Have Liked Your Page and Engaged with Your Post	Number of SGAG's followers who engaged with a particular SGAG Facebook post. Engagement include actions such as liking page and post, commenting and sharing of the post
Lifetime Post Stories by Comment	Number of comments generated by a specific Facebook post
Lifetime Post Stories by Like	Number of likes generated by a specific Facebook post
Lifetime Post Stories by Share	Number of shares generated by a specific Facebook post
Main Engagement	The total number of interactions generated by a Facebook post. Main Engagement includes the number of times a Facebook post is commented on, liked or shared by Facebook users
No of Negative Feedback Per Thousand User	Number of negative feedbacks received for every thousand users reached. Negative feedbacks include actions such as hide post, hide all posts, report spam and unlike page click
Hide All Clicks Per Thousand User	Number of hide all clicks count for every thousand users reached. Hide all clicks results in all SGAG's Facebook posts to be hidden from a user's timeline
Hide Clicks Per Thousand User	Number of hide clicks count for every thousand users reached. Hide click results in a particular post being hidden from a user's timeline

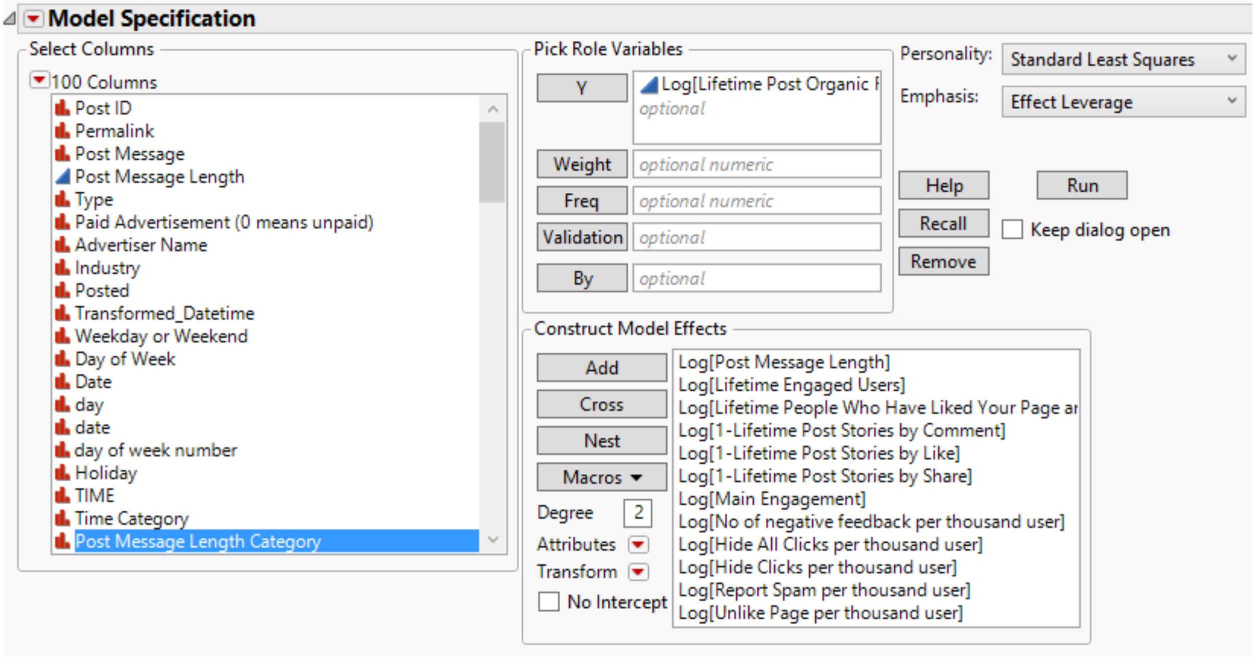
Report Spam Per Thousand User	Number of report spam count for a specific post for every thousand user reached. A report spam may results in a specific post being removed from Facebook
Unlike Page Per Thousand User	Number of unlike page count for every thousand user reached. Unlike page results in the decrease in number of SGAG’s followers
Type	The type of media in a Facebook post. The different types of media include Photo, Video and Link Share
Weekday or Weekend	Weekday or Weekend specifies whether a Facebook post is generated during weekday (Monday to Friday) or weekend (Saturday to Sunday)
Public Holiday	Public Holiday specifies whether a Facebook post is posted on a Public Holiday. Users of Facebook are assumed to be free from work or school during public holiday. Hence, may affect the reach of a Facebook Post
Time	Time specifies the time period where a Facebook post is generated. Time in this paper is split into 9 different categories as follow: * 00:00-04:59 - Midnight * 05:00-06:59 - Dawn * 07:00-09:59 - Morning * 10:00-11:59 - Late Morning * 12:00-13:59 - Noon * 14:00-16:59 - Afternoon * 17:00-18:59 - Evening * 19:00-21:59 - Night * 22:00-23:59 - Late Night
Post Message Length Category	Post message is the description of a Facebook post and Post Message Length Category specifies whether the length (number of characters) of a Facebook post is short or long. For the purpose of this paper, post message with 85 characters or lesser is defined as short and post message with more than 85 characters is considered as long

Figure 1-Post Dataset Description

Other factors such as sharing of Facebook post’s link through Facebook Messenger, Facebook search and Google search have also been considered in the variable selection process. However, there variables are not recorded in the available dataset and could not be derived from the existing data. Hence, will not be used in developing the model.

Analysis Methodology

As mentioned previously, Multiple Linear Regressions can be developed using the Fit Model platform of JMP Pro by selecting the response variable (Y axis) which is Post's reach (Lifetime Organic Post Reach) and the explanatory variables (X axis) which comprises of various continuous variables such as Main Engagement and Negative Feedbacks. Categorical variables will be added to the model once multicollinearity is eliminated from the model.



Model Specification

Select Columns

100 Columns

- Post ID
- Permalink
- Post Message
- Post Message Length
- Type
- Paid Advertisement (0 means unpaid)
- Advertiser Name
- Industry
- Posted
- Transformed_Datetime
- Weekday or Weekend
- Day of Week
- Date
- day
- date
- day of week number
- Holiday
- TIME
- Time Category
- Post Message Length Category

Pick Role Variables

Y: Log[Lifetime Post Organic F

optional

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Personality: Standard Least Squares

Emphasis: Effect Leverage

Construct Model Effects

- Log[Post Message Length]
- Log[Lifetime Engaged Users]
- Log[Lifetime People Who Have Liked Your Page ar
- Log[1-Lifetime Post Stories by Comment]
- Log[1-Lifetime Post Stories by Like]
- Log[1-Lifetime Post Stories by Share]
- Log[Main Engagement]
- Log[No of negative feedback per thousand user]
- Log[Hide All Clicks per thousand user]
- Log[Hide Clicks per thousand user]
- Log[Report Spam per thousand user]
- Log[Unlike Page per thousand user]

Degree: 2

□ No Intercept

□ Keep dialog open

Figure 2 - Initial Post KPI model variables selection

RSquare	0.9738288247
RSquare Adj	0.9736652549
Root Mean Square Error	0.1432921804
Mean of Response	12.170766568
Observations (or Sum Wgts)	1933

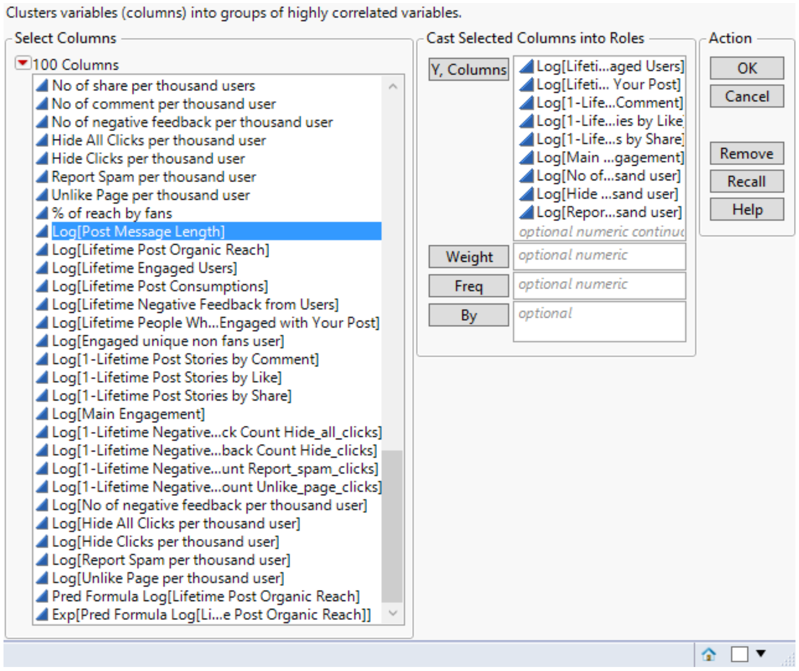
Figure 3 - Initial Fit Model Result

As seen from Figure 3, 97% of variability of the response variable can be explained by the explanatory variables. High R-Square indicates high accuracy of the model. Nonetheless, multicollinearity will need to firstly be examined in order to attain the true factors affecting the KPI.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.6745345819	0.0788112305	84.69	<.0001	-
Log[Post Message Length]	-0.012346746	0.0050212799	-2.46	0.0140	1.2107999609
Log[Lifetime Engaged Users]	0.2605620354	0.0216906484	12.01	<.0001	33.353948038
Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	-0.207046063	0.0225550534	-9.18	<.0001	25.430528502
Log[1-Lifetime Post Stories by Comment]	-0.003728816	0.0084270047	-0.44	0.6582	11.743411805
Log[1-Lifetime Post Stories by Like]	-0.148498706	0.0564858917	-2.63	0.0086	316.8422718
Log[1-Lifetime Post Stories by Share]	-0.008990219	0.0085164139	-1.06	0.2913	13.399878139
Log[Main Engagement]	0.1728637159	0.068333113	2.53	0.0115	503.31054979
Log[No of negative feedback per thousand user]	0.4520888611	0.0227909579	19.84	<.0001	22.443169858
Log[Hide All Clicks per thousand user]	-0.127644396	0.0076284865	-16.73	<.0001	2.9465657698
Log[Hide Clicks per thousand user]	-0.434033703	0.0224787126	-19.31	<.0001	20.138751107
Log[Report Spam per thousand user]	-0.625354801	0.0115775278	-54.01	<.0001	9.5185486791
Log[Unlike Page per thousand user]	-0.188574361	0.0088574765	-21.29	<.0001	5.5668332132

Figure 4 - Initial Parameter Estimates Report

From Figure 4 above, the highlighted columns such as Lifetime Engaged Users and Hide Clicks per thousand user indicate variables with VIF of larger than 8. These variables will be re-selected through the use of variables clustering analysis to eliminate the multicollinearity.



Clusters variables (columns) into groups of highly correlated variables.

Select Columns

100 Columns

- No of share per thousand users
- No of comment per thousand user
- No of negative feedback per thousand user
- Hide All Clicks per thousand user
- Hide Clicks per thousand user
- Report Spam per thousand user
- Unlike Page per thousand user
- % of reach by fans
- Log[Post Message Length]
- Log[Lifetime Post Organic Reach]
- Log[Lifetime Engaged Users]
- Log[Lifetime Post Consumptions]
- Log[Lifetime Negative Feedback from Users]
- Log[Lifetime People Wh...Engaged with Your Post]
- Log[Engaged unique non fans user]
- Log[1-Lifetime Post Stories by Comment]
- Log[1-Lifetime Post Stories by Like]
- Log[1-Lifetime Post Stories by Share]
- Log[Main Engagement]
- Log[1-Lifetime Negative...ck Count Hide_all_clicks]
- Log[1-Lifetime Negative...back Count Hide_clicks]
- Log[1-Lifetime Negative...unt Report_spam_clicks]
- Log[1-Lifetime Negative...ount Unlike_page_clicks]
- Log[No of negative feedback per thousand user]
- Log[Hide All Clicks per thousand user]
- Log[Hide Clicks per thousand user]
- Log[Report Spam per thousand user]
- Log[Unlike Page per thousand user]
- Pred Formula Log[Lifetime Post Organic Reach]
- Exp[Pred Formula Log[Li...e Post Organic Reach]]

Cast Selected Columns into Roles

Y, Columns
- Log[Lifeti...aged Users]
- Log[Lifeti... Your Post]
- Log[1-Life...Comment]
- Log[1-Life...ies by Like
- Log[1-Life...s by Share]
- Log[Main ...gagement]
- Log[No of...sand user]
- Log[Hide ...sand user]
- Log[Repor...sand user]

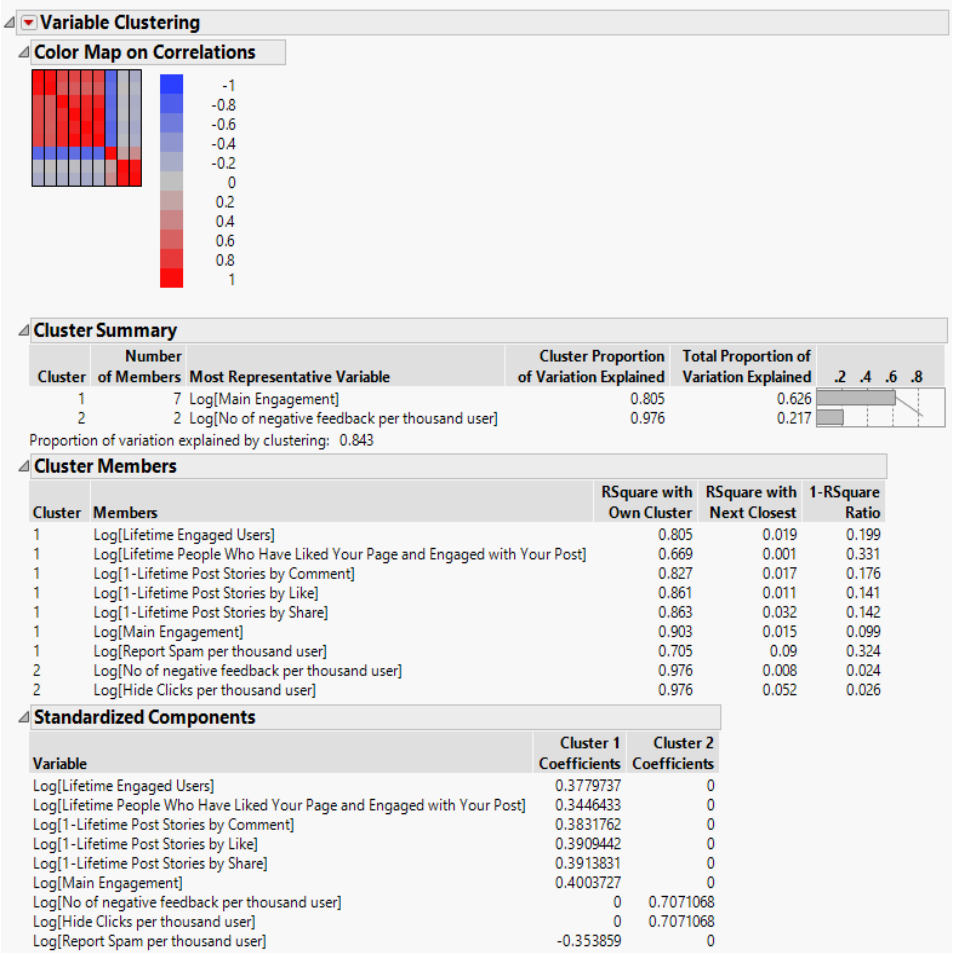
optional numeric continu

Weight: optional numeric

Freq: optional numeric

By: optional

Figure 5 - Variable Clustering-Variables Selection



Variable Clustering

Color Map on Correlations

Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained	.2 .4 .6 .8
1	7	Log[Main Engagement]	0.805	0.626	
2	2	Log[No of negative feedback per thousand user]	0.976	0.217	

Proportion of variation explained by clustering: 0.843

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	Log[Lifetime Engaged Users]	0.805	0.019	0.199
1	Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	0.669	0.001	0.331
1	Log[1-Lifetime Post Stories by Comment]	0.827	0.017	0.176
1	Log[1-Lifetime Post Stories by Like]	0.861	0.011	0.141
1	Log[1-Lifetime Post Stories by Share]	0.863	0.032	0.142
1	Log[Main Engagement]	0.903	0.015	0.099
1	Log[Report Spam per thousand user]	0.705	0.09	0.324
2	Log[No of negative feedback per thousand user]	0.976	0.008	0.024
2	Log[Hide Clicks per thousand user]	0.976	0.052	0.026

Standardized Components

Variable	Cluster 1 Coefficients	Cluster 2 Coefficients
Log[Lifetime Engaged Users]	0.3779737	0
Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	0.3446433	0
Log[1-Lifetime Post Stories by Comment]	0.3831762	0
Log[1-Lifetime Post Stories by Like]	0.3909442	0
Log[1-Lifetime Post Stories by Share]	0.3913831	0
Log[Main Engagement]	0.4003727	0
Log[No of negative feedback per thousand user]	0	0.7071068
Log[Hide Clicks per thousand user]	0	0.7071068
Log[Report Spam per thousand user]	-0.353859	0

Figure 6 - Result of Variable Clustering

As shown in Figure 6, there were two clusters formed and the representative variables are Main Engagement and No of negative feedback per thousand users. Hence, other variables will be removed from the model and the model will be re-run and further evaluated.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.4834965707	0.0729933085	88.82	<.0001	-
Log[Post Message Length]	-0.012606496	0.0094639455	-1.33	0.1830	1.1053233317
Log[Main Engagement]	0.2108027014	0.008578625	24.57	<.0001	1.9170512422
Log[No of negative feedback per thousand user]	0.0448740245	0.0116652919	3.85	0.0001	1.3945330549
Log[Hide All Clicks per thousand user]	-0.194843142	0.0124304264	-15.67	<.0001	1.8606912617
Log[Unlike Page per thousand user]	-0.658934268	0.0123062424	-53.54	<.0001	2.5668115439

Figure 7- Parameter Estimates Report After Removal of Variables with VIF >8

From Figure 7, the remaining independent variables are those with VIF < 8. After eliminating multicollinearity from the model, we will now attempt to filter insignificant factors from this model. As explained previously, variables with p-value (Prob>|t|) of larger than 0.05 are considered as insignificant and the highlighted factor (Post Message Length) have a p-value of 0.18. Therefore, will be removed from the model to improve the true accuracy of the model. Afterwhich, categorical variables such as type and public holiday will also be included to the model.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.5615831279	0.0590705157	111.08	<.0001	-
Log[Main Engagement]	0.2725132304	0.0090431752	30.13	<.0001	2.4240129257
Log[No of negative feedback per thousand user]	0.0491550117	0.0114870526	4.28	<.0001	1.5386842122
Log[Hide All Clicks per thousand user]	-0.167424269	0.0118979289	-14.07	<.0001	1.9397228905

Log[Unlike Page per thousand user]	-0.576924494	0.0126368175	-45.65	<.0001	3.0797337569
Type[Link]	-0.051257923	0.0133314366	-3.84	0.0001	1.2925389457
Type[Photo]	-0.16277523	0.0110410311	-14.74	<.0001	1.5044253862
Weekday or Weekend[Weekday]	-0.005806343	0.0071787778	-0.81	0.4187	1.0269336703
Holiday[Holiday]	-0.027639689	0.0161276649	-1.71	0.0867	1.0111475721
Time Category[Afternoon]	-0.017129676	0.0193873747	-0.88	0.3770	2.7835612329
Time Category[Evening]	0.0094737846	0.0181923372	0.52	0.6026	3.0065263741
Time Category[Late Morning]	-0.016331828	0.026470319	-0.62	0.5373	3.0119335303
Time Category[Late Night]	0.0088809689	0.0294894151	0.30	0.7633	3.352201668
Time Category[Midnight]	0.1129901689	0.0884663594	1.28	0.2017	20.473462759
Time Category[Morning]	-0.05599346	0.0366253397	-1.53	0.1265	4.4228850158
Time Category[Night]	0.0132214261	0.017908259	0.74	0.4604	3.1088663485
Post Message Length Category[Long]	-0.0066727	0.0065533504	-1.02	0.3087	1.1126584169

Figure 8 - Parameter Estimates Report with Categorical Variables

As seen above in Figure 8, categorical variables except for Type have high p-value and are insignificant in explaining the response variable. Hence, these will be eliminated from this model as well.

Result

RSquare	0.8954256411
RSquare Adj	0.8951222339
Root Mean Square Error	0.2835935384
Mean of Response	12.131227625
Observations (or Sum Wgts)	2075

Figure 9 - Final Post KPI Fit Model Result

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.5956308923	0.05501587	119.89	<.0001	-
Log[Main Engagement]	0.2691110379	0.008926188	30.15	<.0001	2.3518679998
Log[No of negative feedback per thousand user]	0.0535310367	0.0113320038	4.72	<.0001	1.4911919211
Log[Hide All Clicks per thousand user]	-0.169625705	0.0118500108	-14.31	<.0001	1.9161181524
Log[Unlike Page per thousand user]	-0.579683556	0.0126012574	-46.00	<.0001	3.0496735143
Type[Link]	-0.051800677	0.0130757486	-3.96	<.0001	1.2382567402
Type[Photo]	-0.162434675	0.0110281163	-14.73	<.0001	1.4946582486

Figure 10 - Final Parameter Estimates Report

The final model and its reports can be seen from Figure 9 and 10. In this model, 89.5% of the variability of the response variable (Post Reach) can be explained by the various independent variables (Main Engagement, Number of Negative Feedbacks per thousand user, Hide All Clicks per thousand user, Unlike Page per thousand user, Type). The VIF of the explanatory variables are lesser than 8, indicating no multicollinearity exists and p-value of all variables are lesser than 0.0001 indicating strong it is a strong model.

Equation representing the relationship between the dependent variable and the independent variables can be written as such:

Post Reach = **6.60** (intercept) + **0.27** (Log[Main Engagement]) + **0.05** (Log[No of negative feedback per thousand user]) - **0.17** (Log[Hide All Clicks per thousand user]) - **0.58** (Log[Unlike Page per thousand user])+ Match(Type)("Link" → **-0.05**, "Photo" → **- 0.16**, "Video" → **0.21**)

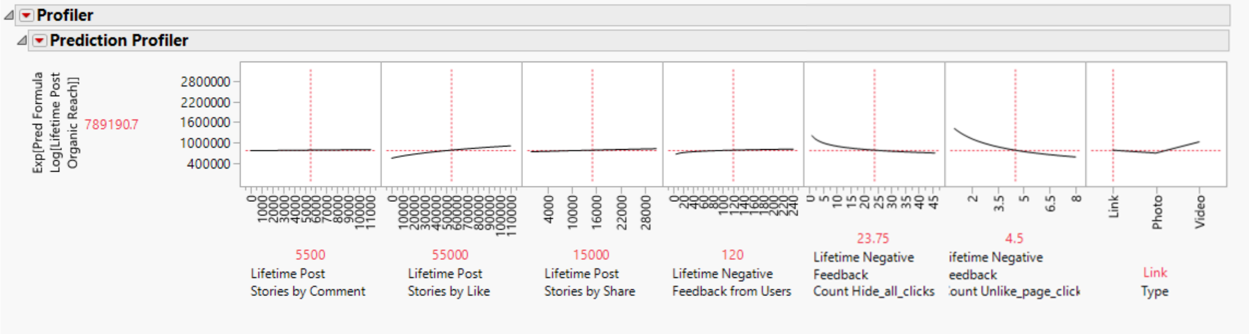


Figure 11 - Profiler of Final Model

To further examine the effects of the independent variables on the dependent variable, the equation which consists of values which were previously transformed by logarithm function is converted back to the original values through the use of exponential function. Through the profiler seen in Figure 11, we will be able to see the degree of linearity of each independent variable towards the dependent variable, as well as the degree of sensitivity of the response variable to the adjustments of the explanatory variables.

Evaluating and Establishing Facebook Video Post KPI using Multiple Linear Regression

Problem Definition

In measuring the performance of SGAG's video posts, number of unique Facebook users who view the video is used. However, SGAG is unsure of which factors to focus on to increase the number of video viewers. With the number of unique video viewers on Facebook as the KPI, several potential factors that may influence the number of video viewers have been identified.

Dataset

Similarly, probable factors available in SGAG's Facebook video post's dataset are firstly selected. Then, other components such as time category, day of week, public holiday, as well as video post length category have been derived in order to improve the accuracy of the model. As mentioned previously, multiple linear regression analysis requires values to be normally distributed. Hence, log transformation has also been performed on the continuous variables. The final dataset derived from this selection and transformation process and its accompanying definition can be seen in Figure 12 below.

Factors	**Definition**
Lifetime Post Consumers by Clicks to Play	Number of Facebook users who plays a video post by clicking on the play button
No of Negative Feedback Per Thousand User	Number of negative feedbacks received for every thousand users. Negative feedbacks include actions such as hide post, hide all posts, report spam and unlike page click

Hide All Clicks Per Thousand User	Number of hide all clicks count for every thousand users reached. Hide all clicks results in all SGAG’s Facebook posts to be hidden from a user’s timeline
Hide Clicks Per Thousand User	Number of hide clicks count for every thousand users reached. Hide click results in a particular post being hidden from a user’s timeline
Report Spam Per Thousand User	Number of report spam count for a specific post for every thousand user reached. A report spam may results in a specific post being removed from Facebook
Unlike Page Per Thousand User	Number of unlike page count for every thousand user reached. Unlike page results in the decrease in number of SGAG’s followers
Lifetime Post Stories by Comment	Number of comments generated by a specific Facebook post
Lifetime Post Stories by Like	Number of likes generated by a specific Facebook post
Lifetime Post Stories by Share	Number of shares generated by a specific Facebook post
Lifetime People Who Have Liked Your Page and Engaged with Your Post	Number of SGAG’s followers who engaged with a particular SGAG Facebook post. Engagement include actions such as liking page and post, commenting and sharing of the post
Video Length (Category)	Video Length is total time length of a Facebook video in seconds and Video Length Category specifies whether the length of a Facebook video is short or long. For the purpose of this paper, video that is lesser than 30 seconds is considered as short, video between 30 seconds and 60 seconds is defined as medium, whereas video that is longer than 60 seconds is considered as long.
Time	Time specifies the time period where a Facebook post is generated. Time in this paper is split into 9 different categories as follow: * 00:00-04:59 - Midnight * 05:00-06:59 - Dawn * 07:00-09:59 - Morning * 10:00-11:59 - Late Morning * 12:00-13:59 - Noon * 14:00-16:59 - Afternoon * 17:00-18:59 - Evening * 19:00-21:59 - Night * 22:00-23:59 - Late Night
Day of Week	Day of Week specifies the day which a Facebook post is

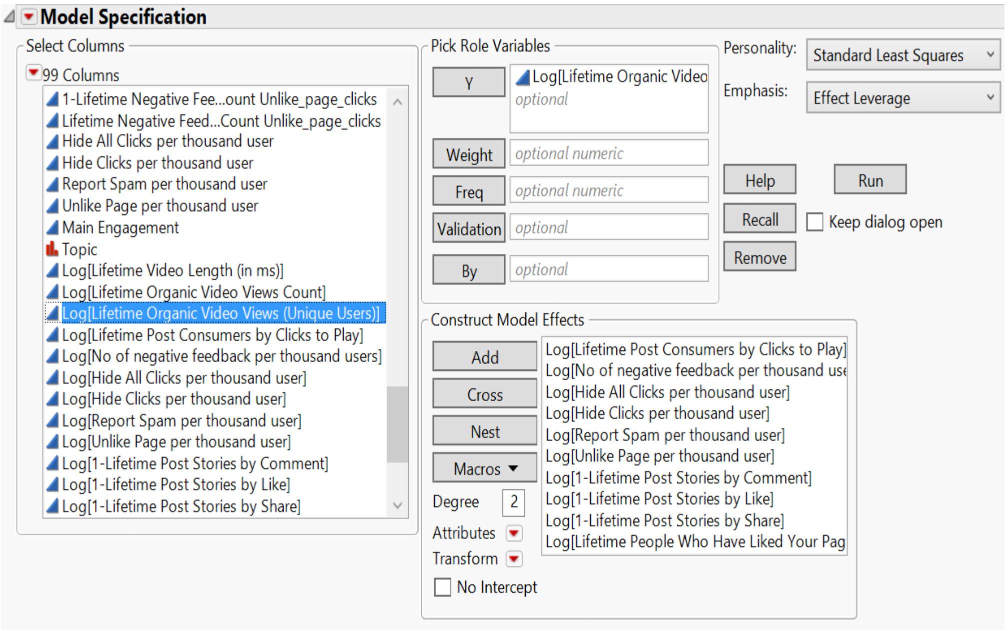
	generated. Day of Week includes Monday to Sunday
Public Holiday	Public Holiday specifies whether a Facebook post is posted on a Public Holiday. Users of Facebook are assumed to be free from work or school during public holiday. Hence, may affect the reach of a Facebook Post

Figure 12 - Video Dataset Descriptions

Other factors: 1) Post Tagging (i.e. SGAG post with User A), 2) Facebook Search, 3) Google Search (external search which lead user to the post), 4) Link Shared (video link copied by Facebook user and circulate via other platforms) are also taken into considerations. However, it could not be included as these factors were not captured in the video dataset nor can it be derived from any existing variable.

Analysis Methodology

Multiple Linear Regressions in JMP can be achieved by using the Fit Model function by selecting the response variable (Y axis) and the explanatory variables (X axis) which explain the changes in response variable. In this case, we would like to see how number of video viewers changes as a result of change in the potential factors, thus video viewers is the response variable on Y axis and other factors in the X axis. As video length, time, day of week and public holidays are categorical variables, it will be added into the model only when the multicollinearity has been eliminated from the model.



Model Specification

Select Columns

99 Columns

- 1-Lifetime Negative Fee...ount Unlike_page_clicks
- Lifetime Negative Feed...Count Unlike_page_clicks
- Hide All Clicks per thousand user
- Hide Clicks per thousand user
- Report Spam per thousand user
- Unlike Page per thousand user
- Main Engagement
- Topic
- Log[Lifetime Video Length (in ms)]
- Log[Lifetime Organic Video Views Count]
- Log[Lifetime Organic Video Views (Unique Users)]
- Log[Lifetime Post Consumers by Clicks to Play]
- Log[No of negative feedback per thousand users]
- Log[Hide All Clicks per thousand user]
- Log[Hide Clicks per thousand user]
- Log[Report Spam per thousand user]
- Log[Unlike Page per thousand user]
- Log[1-Lifetime Post Stories by Comment]
- Log[1-Lifetime Post Stories by Like]
- Log[1-Lifetime Post Stories by Share]

Pick Role Variables

Y: Log[Lifetime Organic Video

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Personality: Standard Least Squares

Emphasis: Effect Leverage

□ Keep dialog open

Construct Model Effects

- Log[Lifetime Post Consumers by Clicks to Play]
- Log[No of negative feedback per thousand use
- Log[Hide All Clicks per thousand user]
- Log[Hide Clicks per thousand user]
- Log[Report Spam per thousand user]
- Log[Unlike Page per thousand user]
- Log[1-Lifetime Post Stories by Comment]
- Log[1-Lifetime Post Stories by Like]
- Log[1-Lifetime Post Stories by Share]
- Log[Lifetime People Who Have Liked Your Pag

Degree: 2

□ No Intercept

Figure 13 - Initial Video KPI model variables selection

RSquare	0.939804
RSquare Adj	0.937532
Root Mean Square Error	0.252624
Mean of Response	11.13354
Observations (or Sum Wgts)	276

Figure 14 - Initial Fit Model Result

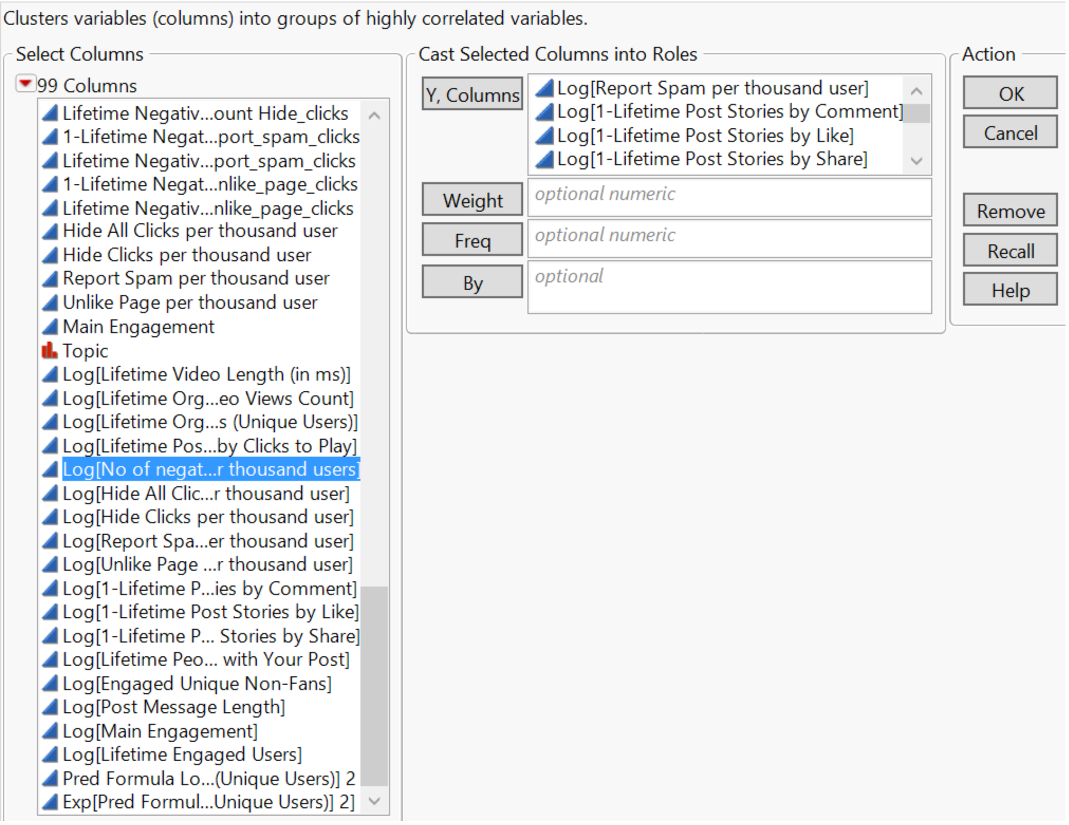
From figure 14, we can see that 93% of variability of the response variable can be explained by the explanatory variables. Although the higher the R-Square value, the more accurate a model is. We will attempt to examine the multicollinearity issues among these factors before finalizing the model.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	2.967288	0.293845	10.1	<.0001	.
Log[Lifetime Post Consumers by Clicks to Play]	0.026804	0.017745	1.51	0.1321	12.09757
Log[No of negative feedback per thousand users]	-0.02056	0.035242	-0.58	0.56	3.317779
Log[Hide All Clicks per thousand user]	0.019215	0.029222	0.66	0.5114	2.429442
Log[Hide Clicks per thousand user]	0.066048	0.041514	1.59	0.1128	3.714407
Log[Report Spam per thousand user]	-0.31354	0.047437	-6.61	<.0001	9.655638
Log[Unlike Page per thousand user]	-0.03613	0.03774	-0.96	0.3393	6.146838
Log[1-Lifetime Post Stories by Comment]	-0.01835	0.027589	-0.67	0.5066	9.395337
Log[1-Lifetime Post Stories by Like]	0.209645	0.034614	6.06	<.0001	9.006644

Log[1-Lifetime Post Stories by Share]	0.013391	0.026715	0.5	0.6166	23.82651
Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	0.481781	0.047202	10.21	<.0001	6.56383

Figure 15 - Parameter Estimates

From Figure 15, we can see that variable Lifetime Post Consumers by Clicks to Play, Report Spam per thousand user and the Lifetime Post Stories by Like, Comment and Share have a VIF over 8. Next, we will perform Clustering Analysis to eliminate multicollinearity.



Clusters variables (columns) into groups of highly correlated variables.

Y, Columns: Log[Report Spam per thousand user], Log[1-Lifetime Post Stories by Comment], Log[1-Lifetime Post Stories by Like], Log[1-Lifetime Post Stories by Share]

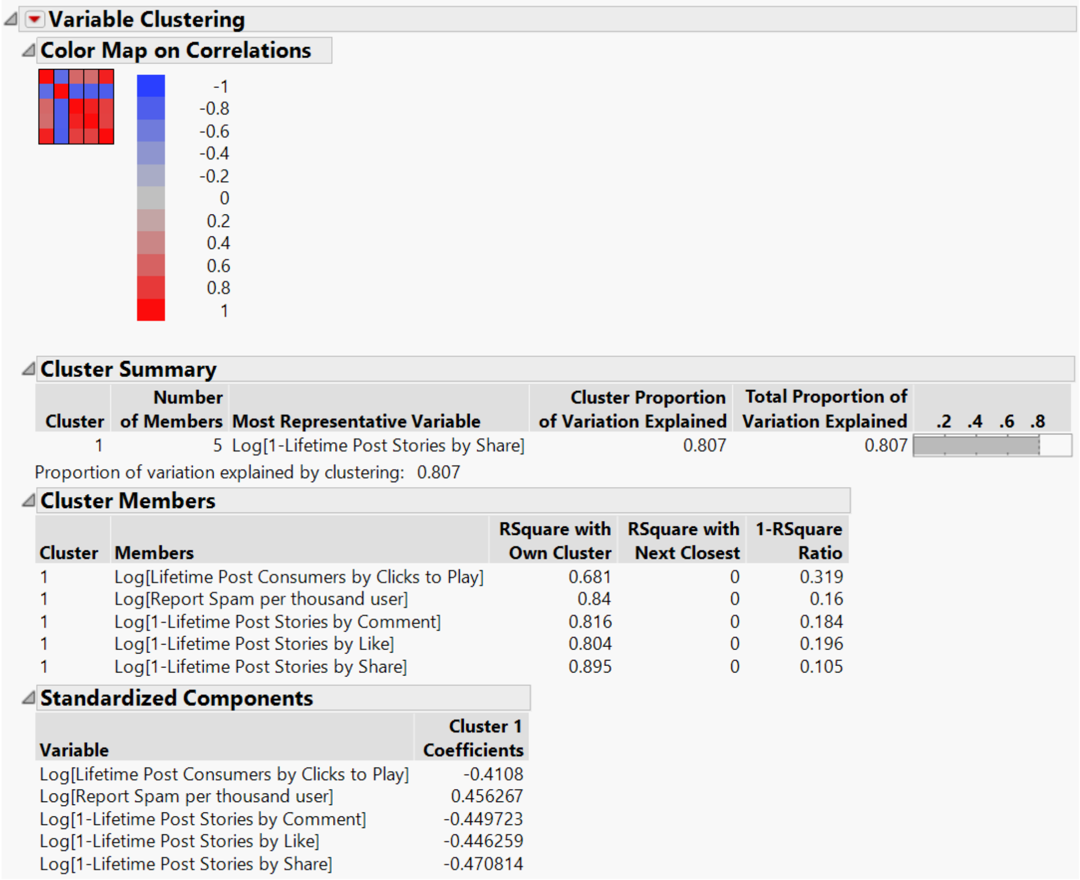
Weight: optional numeric

Freq: optional numeric

By: optional

Select Columns
99 Columns
Lifetime Negativ...ount Hide_clicks
1-Lifetime Negat...port_spam_clicks
Lifetime Negativ...port_spam_clicks
1-Lifetime Negat...nlike_page_clicks
Lifetime Negativ...nlike_page_clicks
Hide All Clicks per thousand user
Hide Clicks per thousand user
Report Spam per thousand user
Unlike Page per thousand user
Main Engagement
Topic
Log[Lifetime Video Length (in ms)]
Log[Lifetime Org...eo Views Count]
Log[Lifetime Org...s (Unique Users)]
Log[Lifetime Pos...by Clicks to Play]
Log[No of negat...r thousand users]
Log[Hide All Clic...r thousand user]
Log[Hide Clicks per thousand user]
Log[Report Spa...er thousand user]
Log[Unlike Page ...r thousand user]
Log[1-Lifetime P...ies by Comment]
Log[1-Lifetime Post Stories by Like]
Log[1-Lifetime P... Stories by Share]
Log[Lifetime Peo... with Your Post]
Log[Engaged Unique Non-Fans]
Log[Post Message Length]
Log[Main Engagement]
Log[Lifetime Engaged Users]
Pred Formula Lo...(Unique Users)] 2
Exp[Pred Formul...Unique Users)] 2]
Cast Selected Columns into Roles
Action
OK
Cancel
Remove
Recall
Help

Figure 16 - Variable Clustering Variables Selection



Variable Clustering

Color Map on Correlations

Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained	.2 .4 .6 .8
1	5	Log[1-Lifetime Post Stories by Share]	0.807	0.807	

Proportion of variation explained by clustering: 0.807

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	Log[Lifetime Post Consumers by Clicks to Play]	0.681	0	0.319
1	Log[Report Spam per thousand user]	0.84	0	0.16
1	Log[1-Lifetime Post Stories by Comment]	0.816	0	0.184
1	Log[1-Lifetime Post Stories by Like]	0.804	0	0.196
1	Log[1-Lifetime Post Stories by Share]	0.895	0	0.105

Standardized Components

Variable	Cluster 1 Coefficients
Log[Lifetime Post Consumers by Clicks to Play]	-0.4108
Log[Report Spam per thousand user]	0.456267
Log[1-Lifetime Post Stories by Comment]	-0.449723
Log[1-Lifetime Post Stories by Like]	-0.446259
Log[1-Lifetime Post Stories by Share]	-0.470814

Figure 17- Result of Clustering Variables

In Figure 17, it is shown that Lifetime Post Stories by Share is the chosen explanatory variable which have the lowest intra distance among the variables in cluster 1 given only one cluster in this case. Hence, we will remove variable Lifetime Post Consumers by Clicks to Play, Report Spam per thousand user and the Lifetime Post Stories by Like and Comment and rerun the model again:

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	2.258531	0.270359	8.35	<.0001	.
Log[No of negative feedback per thousand users]	-0.01444	0.040085	-0.36	0.7189	3.235836
Log[Hide All Clicks per thousand user]	0.025645	0.032556	0.79	0.4316	2.273203
Log[Hide Clicks per thousand user]	-0.02922	0.046045	-0.63	0.5263	3.444875

Log[Unlike Page per thousand user]	-0.1858	0.038252	-4.86	<.0001	4.760512
Log[1-Lifetime Post Stories by Share]	0.103495	0.010309	10.04	<.0001	2.674663
Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	0.747841	0.0328	22.8	<.0001	2.38937

Figure 18 - Parameter Estimates Report After Removal of Variables with VIF >8

As seen in Figure 18, there are no more explanatory variables with VIF over 8. As we are using 95% confidence interval, explanatory variables with Prob>|t| more than 0.05 are considered insignificant to explain the response variable. Variable No of negative feedback per thousand users, Hide All Clicks per thousand user and Hide Clicks per thousand user will be removed from the model as they are least significant in explaining the changes in Video Viewers.

At the same time, categorical variables such as video length, video originality, time category, day of week, and public holiday will be included in the model to evaluate its significance to explain changes in Video Viewers.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	2.422562	0.247437	9.79	<.0001	.
Log[Unlike Page per thousand user]	-0.15096	0.032378	-4.66	<.0001	3.786427
Log[1-Lifetime Post Stories by Share]	0.107533	0.010099	10.65	<.0001	2.849313
Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	0.766859	0.029952	25.6	<.0001	2.21197
Shared Video?[Others]	-0.10114	0.069042	-1.46	0.1442	4.385703
Shared Video?[SGAG]	-0.04791	0.070467	-0.68	0.4972	4.480911
Day of Week[Sunday]	-0.05016	0.044443	-1.13	0.2601	1.890033
Day of Week[Monday]	0.016845	0.037422	0.45	0.653	1.639804
Day of Week[Tuesday]	0.0391	0.04543	0.86	0.3902	1.919209

Day of Week[Wednesday]	0.011373	0.039168	0.29	0.7718	1.666241
Day of Week[Thursday]	-0.0572	0.041467	-1.38	0.169	1.802608
Day of Week[Friday]	0.048496	0.046113	1.05	0.2939	1.919154
Holiday[Holiday]	0.028394	0.044269	0.64	0.5218	1.085721
Time category[Afternoon]	-0.10162	0.060625	-1.68	0.0949	2.717747
Time category[Evening]	-0.06349	0.051075	-1.24	0.215	2.883888
Time category[Late Morning]	-0.02415	0.074019	-0.33	0.7445	2.79555
Time category[Late Night]	-0.14022	0.072663	-1.93	0.0548	2.694071
Time category[Midnight]	0.398871	0.252738	1.58	0.1158	17.92777
Time category[Morning]	0.06428	0.116345	0.55	0.5811	4.606902
Time category[Night]	-0.01839	0.048357	-0.38	0.7041	3.164786
Length Category[Long]	0.032641	0.023931	1.36	0.1738	1.446147
Length Category[Medium]	0.091198	0.025999	3.51	0.0005	1.54032

Figure 19 - Parameter Estimates Report with Categorical Variables

Additional categorical variables are not significant in explaining variability in Video Viewers due to its high p-value. Therefore, these variables will be removed from the model.

Result

RSquare	0.918421
RSquare Adj	0.917521
Root Mean Square Error	0.290281
Mean of Response	11.13354
Observations (or Sum Wgts)	276

Figure 20 - Final Video Fit Model Result

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	2.376078	0.232272	10.23	<.0001	.
Log[Unlike Page per thousand user]	-0.19896	0.032192	-6.18	<.0001	3.387392
Log[1-Lifetime Post Stories by Share]	0.101933	0.010153	10.04	<.0001	2.60647
Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post]	0.731101	0.029518	24.77	<.0001	1.944137

Figure 21 - Final Parameter Estimates

This model can explain about 92% of the variability of the response variable (Video Viewers) can be explained by the various independent variables (Figure 20). All the explanatory variables in the final model have VIF value less than 8 and have a p-value (Prob>|t|) of less than 0.0001 which indicates a strong model (Figure 21).

The equation below illustrates the relationship of the explanatory variables to Video Viewers:
Video Viewers = **2.3761** (intercept) - **0.1920** (Log[Unlike page per thousand user) + **0.1019** (Log[1-Lifetime Post Stories by Share]) + **0.7311** (Log[Lifetime People Who Have Liked Your Page and Engaged with Your Post])

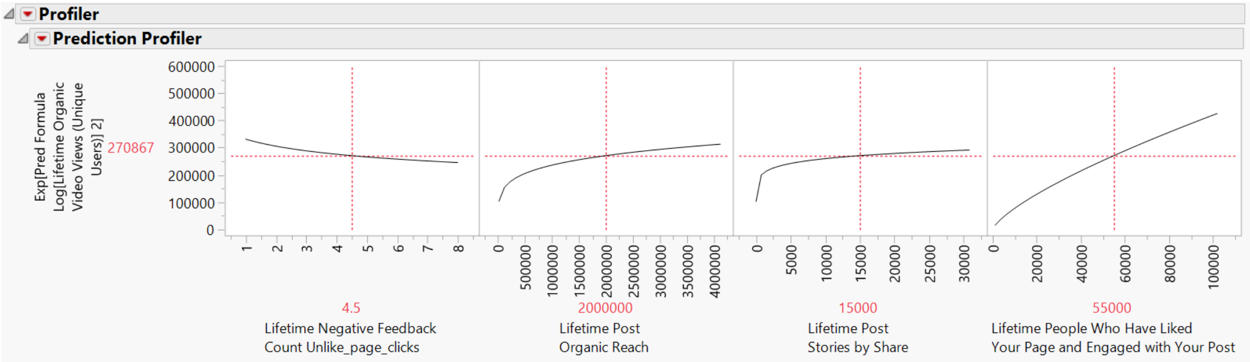


Figure 22 Profiler of Final Model

To further examine the effects of the independent variables on the dependent variable, the equation which consists of values which were previously transformed by logarithm function is converted back to the original values through the use of exponential function. Through the profiler in Figure 22, we will be able to see the degree of linearity of each independent variable towards the dependent variable, as well as the degree of sensitivity of the response variable to the adjustments of the explanatory variables.

Analysis of Social Media Content through Topic Modelling

Development of Topic Modelling in this paper will be done through Text Explorer platform of JMP. The text explorer platform helps to create structure for unstructured text (JMP, n.d.) and transform the text into meaningful data. Text Explorer provide the command to allow user to edit the term as well as visualize the terms in the data using word cloud and many other useful features in the analysis of this paper.

Comparison of Topic Modelling across different software

For JMP Topic Modelling, it allows user to rename the individual terms so as to transform into a more meaningful word. Next, it also allows user to recode and group multiple similar term into one term, for example, terms ‘mum’ and ‘mother’ will be recoded as ‘mother’. Lastly, useful phrases can be added into the list of terms.

For SAS Text Miner, user just need to click on one feature and the category will be automated generated for the different row in the dataset. Lastly, it also displays the number of documents that is contained for each topic ID.

Lastly, for Python, the necessity to develop out the code for the program as well as the time involves. Next, the performance issues for Python, it may not be handle a large set of data compared to JMP and SAD. Lastly, with the knowledge to code out the program, user also have the control of using his own algorithm as well as setting the conditions.

Problem Definition

In the dataset exported from SGAG’s Facebook posts, there is no data column mentioning about the topic of their posts available. Hence, it is challenging for the company to understand the interests generated for the different topics. For example, a topic on celebrities may be able gain more reach as compared to a topic on food, but the company is unable to derive this information. Thus, an attempt to generate the post topics to understand how each topic will affect the performance of the posts will be made in this paper.

Dataset

For text explorer analysis, the main aim is to find the list of suitable topics from the list of terms. Hence, only the unstructured text which is the post message will be used for this model.

Factors	Definition
Post Message	Description of a Facebook Post

Figure 23 - Topic Modelling Dataset description

In this paper, Post Message variable is used as the main variable in the JMP Text Explorer platform in order to transform the unstructured text data and convert them into meaningful data consisting of post topics.

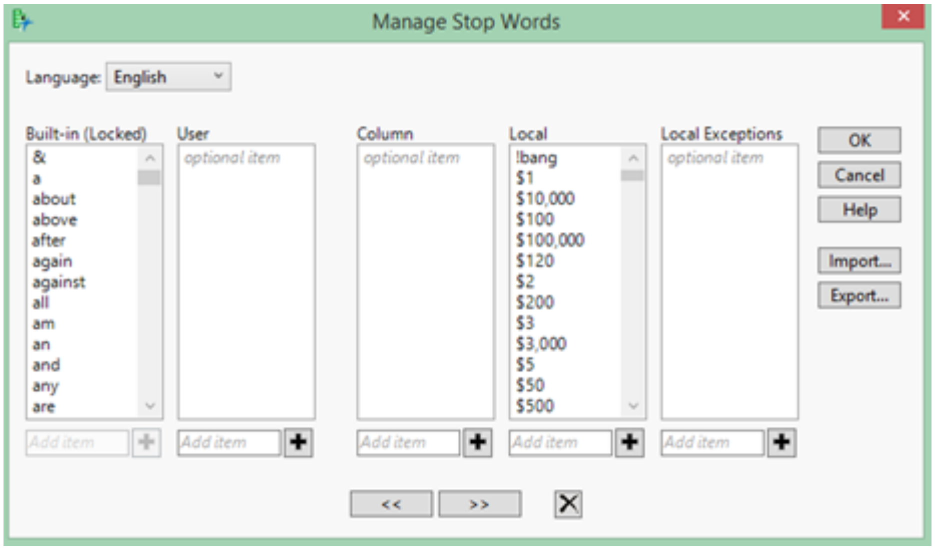
Analysis Methodology

Text analysis is often an iterative process as it is difficult clean up all the terms in just one round and produce accurate insights. Hence, to obtain a more accurate analysis, there is a need to perform curating and analysing repeatedly on the terms found.

Curating the List of Terms

In order to generate the final list of terms, stop words, re-coded words as well as phrases will need to be defined.

Managing Stop Words



Language: English

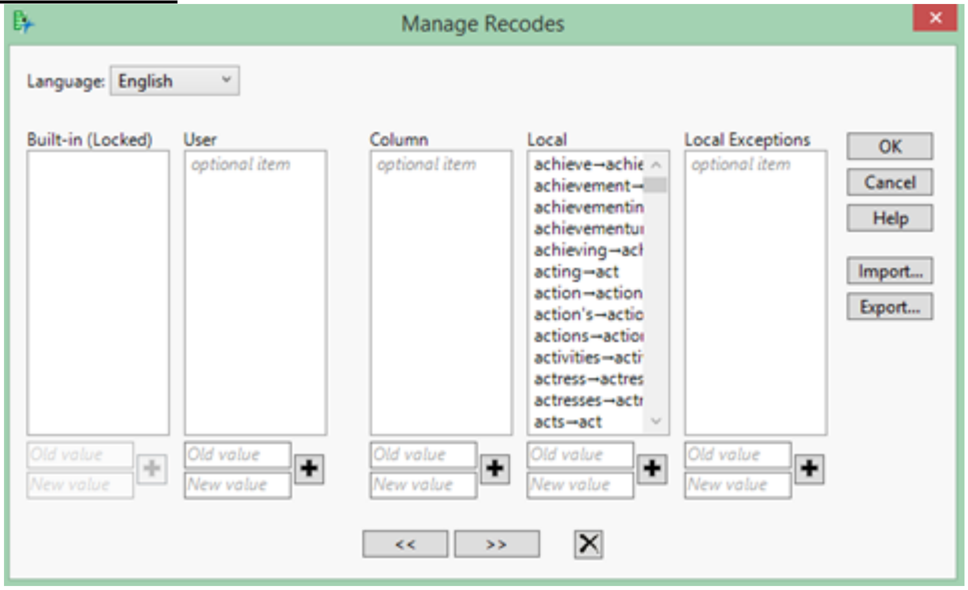
Built-in (Locked): &, a, about, above, after, again, against, all, am, an, and, any, are

Local: !bang, $1, $10,000, $100, $100,000, $120, $2, $200, $3, $3,000, $5, $50, $500

Figure 24 – List of Stop Words

Stop words refers to common words as well as irrelevant words such as “face”, “haha” and “hello”. Figure 24 shows a list of words that is defined as stop words by default, as well as specific words that are added as stop words is shown.

Managing Recoded Terms



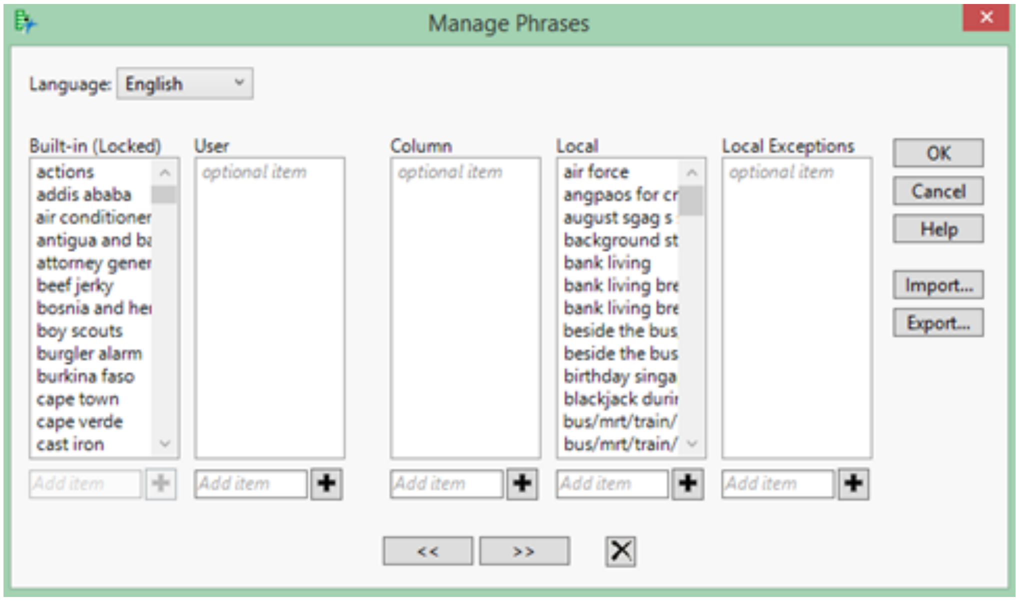
Language: English

Local: achieve→achie, achievement→, achievementin, achievementu, achieving→acl, acting→act, action→action, action's→actio, actions→actio, activities→acti, actress→actres, actresses→actr, acts→act

Figure 25 – List of Recoded Terms

Recoded terms refer to grouping of relevant words into one single term. For instance, words like bus, mrt, train are being grouped together to form a common term. It also refers to words that are being renamed for clearer understanding, such as the word army is renamed as sgag army. Figure 25 shows a list of terms that are recoded in order to provide a meaningful analysis.

Managing Phrases



Language: English

Built-in (Locked)
actions
addis ababa
air conditioner
antigua and ba
attorney gener
beef jerky
bosnia and her
boy scouts
burgler alarm
burkina faso
cape town
cape verde
cast iron

User

Column

Local
air force
angpaos for cr
august sgag s
background st
bank living
bank living bre
bank living bre
beside the bus
beside the bus
birthday singa
blackjack durir
bus/mrt/train/
bus/mrt/train/

Local Exceptions

Figure 26 – List of Phrases

An example of a phrase can be ‘joseph Isaac schooling’ or ‘lee hsien loong’ and phrases that are generated by Text Explorer can either be added to the list of stop words or recoded terms. Phrases that can potentially provide meaningful analysis are added to the list of recoded terms in this paper.

Analysing the List of Terms

After concluding the curation process through stop words, recoding and managing phrases, analysis on the curated list of terms based on document term matrix will then be executed. Document term matrix uses inverse document frequency where each value in the matrix represents the number of occurrences of term i in document j. In this paper, Latent Semantic Analysis and Topic Modelling will be used for the analysis of list of terms.

Model

Latent Semantic Analysis

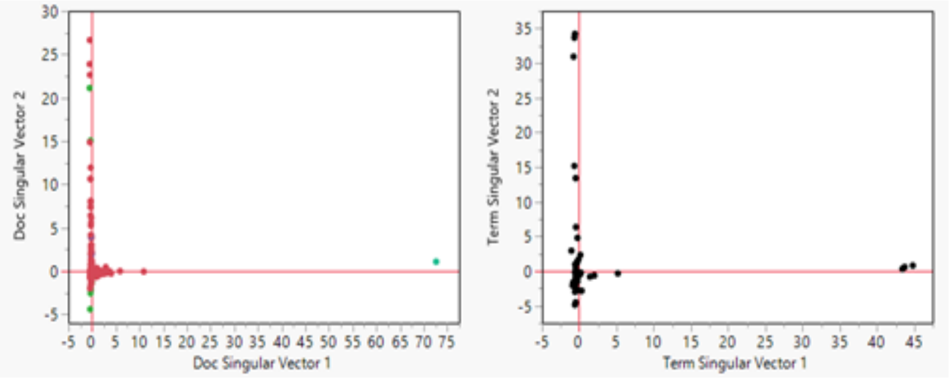
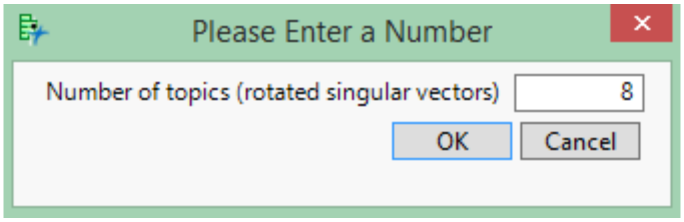


Figure 27 – Latent Semantic Analysis

The concept of Latent Semantic Analysis (LSA) helps to find the relevance between the documents (post message) and the words. LSA adopts the logic that terms that are of relevant will appear near to each other. The vectors capture the relationship among diverse words with similar meanings or topic areas.

Latent Semantic Analysis (LSA) shows the relevance between documents (post message) and words in the respective document. In addition, LSA believes that terms that are relevant will appear near to each other. Hence, the vectors capture the relationship among different words with similar meanings or similar topic areas. For example, terms such as ‘army’, ‘scoot’ and ‘travel’, are clustered near to each other because the three words tend to appear in the same documents. These words will most likely be categorized into the same topic.

Topic Analysis



Please Enter a Number

Number of topics (rotated singular vectors) 8

Figure 28 – Number of rotated singular vectors for Topic Analysis

The Topic Analysis performs a rotated singular value decomposition (SVD) of the document term matrix (DTM). The inputted number of rotated singular vectors will be equivalent to the number of topics. JMP will generate the list of Topic with the respective terms as well as the Topic Scores report.

Results

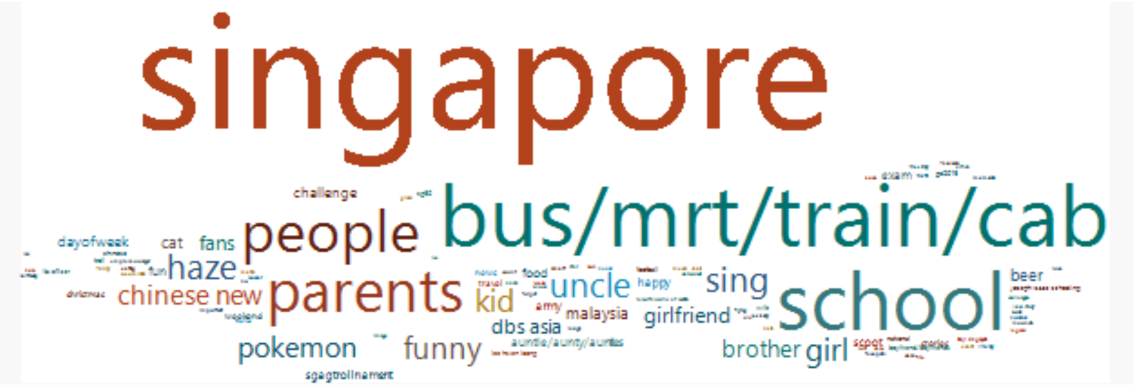
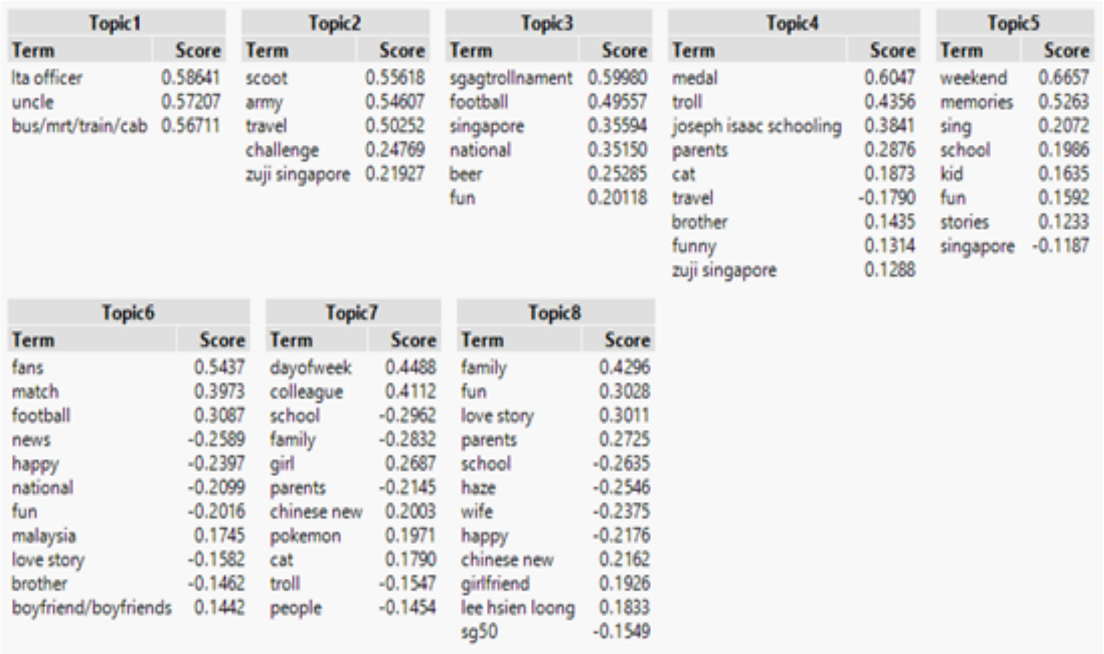


Figure 29 – Word Cloud



Topic1		Topic2		Topic3		Topic4		Topic5	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
lta officer	0.58641	scoot	0.55618	sgagtrollnament	0.59980	medal	0.6047	weekend	0.6657
uncle	0.57207	army	0.54607	football	0.49557	troll	0.4356	memories	0.5263
bus/mrt/train/cab	0.56711	travel	0.50252	singapore	0.35594	joseph isaac schooling	0.3841	sing	0.2072
		challenge	0.24769	national	0.35150	parents	0.2876	school	0.1986
		zuji singapore	0.21927	beer	0.25285	cat	0.1873	kid	0.1635
				fun	0.20118	travel	-0.1790	fun	0.1592
						brother	0.1435	stories	0.1233
						funny	0.1314	singapore	-0.1187
						zuji singapore	0.1288		

Topic6		Topic7		Topic8	
Term	Score	Term	Score	Term	Score
fans	0.5437	dayofweek	0.4488	family	0.4296
match	0.3973	colleague	0.4112	fun	0.3028
football	0.3087	school	-0.2962	love story	0.3011
news	-0.2589	family	-0.2832	parents	0.2725
happy	-0.2397	girl	0.2687	school	-0.2635
national	-0.2099	parents	-0.2145	haze	-0.2546
fun	-0.2016	chinese new	0.2003	wife	-0.2375
malaysia	0.1745	pokemon	0.1971	happy	-0.2176
love story	-0.1582	cat	0.1790	chinese new	0.2162
brother	-0.1462	troll	-0.1547	girlfriend	0.1926
boyfriend/boyfriends	0.1442	people	-0.1454	lee hsien loong	0.1833
				sg50	-0.1549

Figure 30 – Topics

After iterations of curating and analysing the terms in the dataset, Figure 30 shows the final topic report for this paper. Below is the list of topics that has been identified.

Topic Number	Topic
Topic 1	Transportation
Topic 2	Travel
Topic 3	Events
Topic 4	Olympics
Topic 5	Moments
Topic 6	Hobbies
Topic 7	Daily Lives
Topic 8	Relationships

Figure 31

With this generated set of topics in Figure 31, we refer to the model established earlier on to find out which topic generate the highest engagement, which topic generate the highest shares as well as which topic causes the highest negative feedback value.

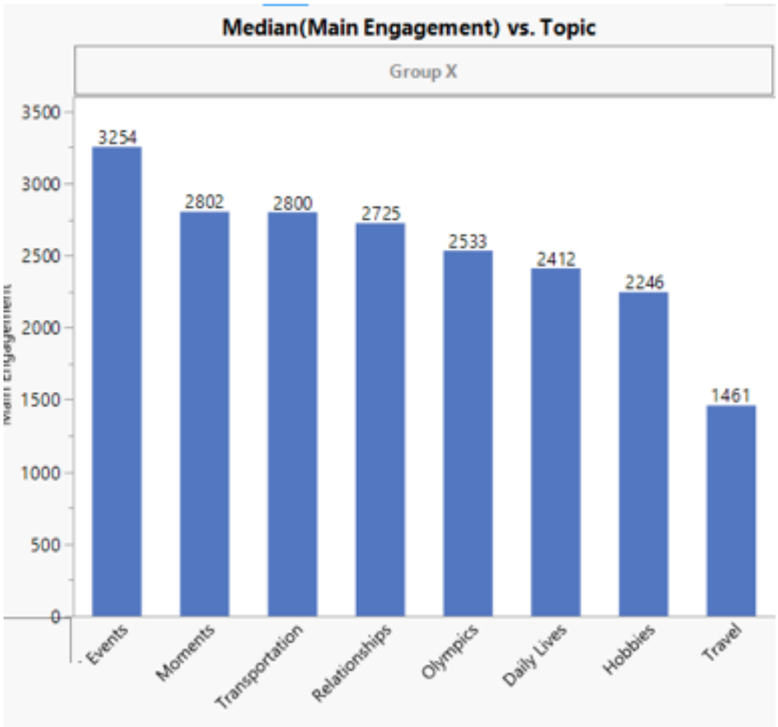


Figure 32 – Number of Main Engagement by Topic

As seen in Figure 32, Events has the highest number of main engagement at 3254 followed by Moments at 2802 main engagement with Transportation at 2800.

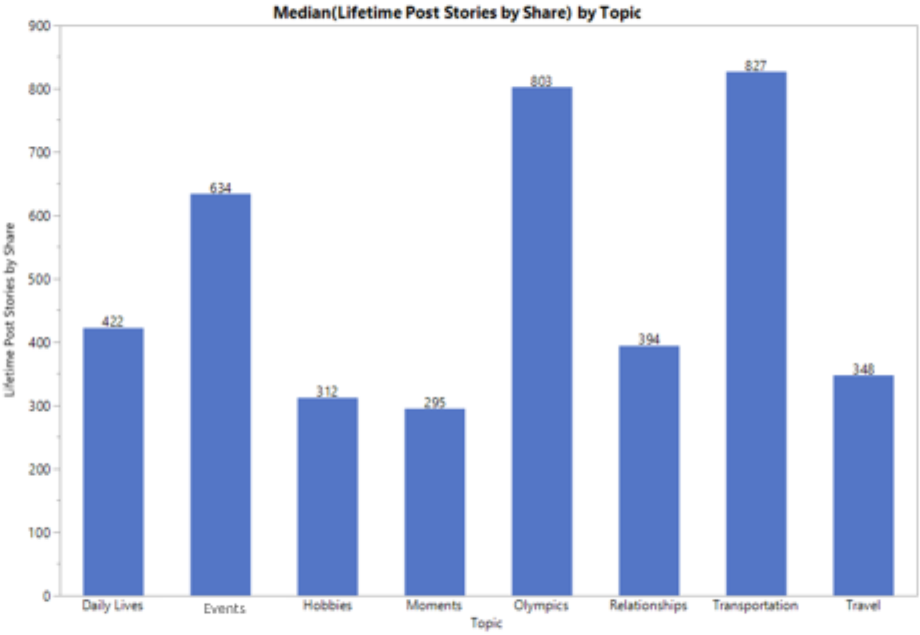


Figure 33 – Number of Shares by Topic

In Figure 33, Transportation has the highest number of shares at 827 followed by Olympics at 803 shares.

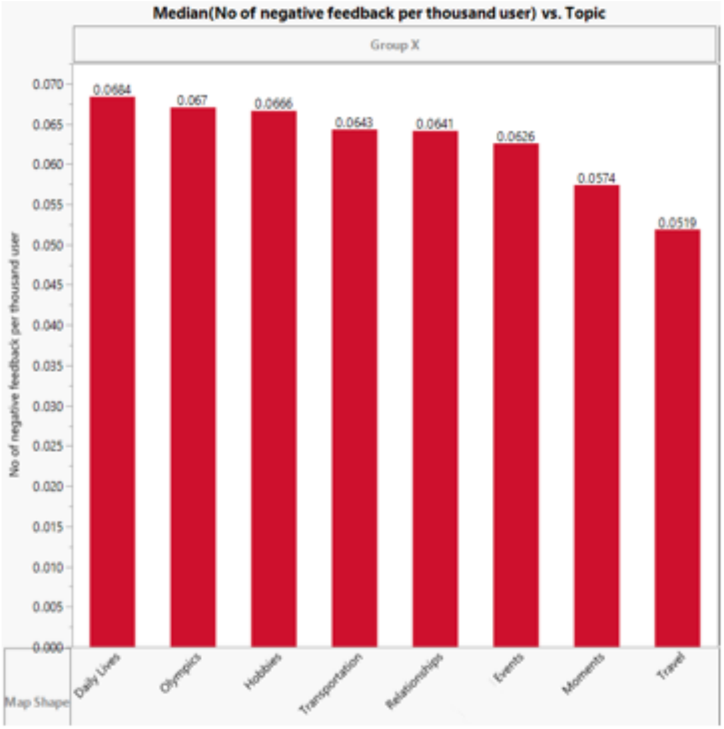


Figure 34 – Number of Negative Feedback by Topic

From Figure 34, Daily Lives has the highest number of negative feedbacks at 0.0684 per thousand user followed by Olympics at 0.067 per thousand user.

Discussion

This paper examined the relationship between explanatory variables affecting post reach and number of video viewers using Multiple Linear Regression. Topic modelling technique is incorporated to better understand audience engagement based on various topics identified.

Multiple Linear Regression assumes a linear relationship between the dependent and independent variables. Hence, any outlier(s) in the dataset may distort the accuracy of the model. In addition, Multiple Linear Regression analysis requires values to be normally distributed. Thus, log transformation has been performed on continuous variables to reduce skewness in the dataset and to prevent some factors of higher values to be dominant in the model. In this paper, independent variables with VIF more than 8 are assumed to have multicollinearity issues and were removed from the model.

Despite both models being able to explain at least 89% of the variances of the independent variables, the model did not take into account of the effect of independent variables on other variables that may decrease the value of the dependent variable. For instance, fans engagement and main engagement help to improve reach or video viewers. However, when more people are seeing the post and engage with the post, it also increases the likelihood of people giving negative feedback (hide click, hide all posts, report spam, unlike page)(Wittman, 2012).

In the topic modelling, it is inevitable to manually regrouping some words due to inconsistency of the post message wording. For instance, "gf" and girlfriend were both used in different post messages. Thus, regrouping is necessary in order to obtain better topic report. In addition, the accuracy of the topics may be affected due to homonyms and the topics formulated were only based on one-year dataset obtained. To better examine the performance of the topics, it would be better to have dataset of a few years' worth.

Possible Future Works

The analysis and findings produced through this paper are constrained by the limitation of time and availability of data. Therefore, there are many other means that may help in improving the comprehensiveness of the analysis and models. For example, dataset used for the analysis in this paper was only one-year worth of the data. Hence, increasing the amount of data to several years' worth of dataset may be able to provide better insights and may result in models with greater accuracy.

Moreover, as mentioned previously, some data such as the number of Facebook post's link share through medium like Facebook messenger were also unavailable. With this additional dataset, as well as other potential data that may help in deriving the models, a more thorough analysis can be obtained to provide greater insights for the social media content producer.

Last but not least, considering the ever changing social media platforms, users, as well as consumer behaviours, the analysis of current performance and factors affecting it may not be applicable in the near future. Hence, a possible initiative that is more future proof for SGAG is to develop a user interface whereby the company would be able to load a new set of data in and new models would be generated based on the data given.

Conclusion

In conclusion, this paper aims to help SGAG to discover important factors affecting the performance of their Facebook posts which help them to remain competitive in the industry and increase their consumer base. Two Multiple Linear Regressions were constructed to identify significant factors affecting the KPIs and a Topic Modelling Technique was used to examine various subjects from the posts. The Post Reach regression model explains 89.5% of the variation in Main Engagement, Number of Negative Feedbacks per thousand user, Hide All Clicks per thousand user, Unlike Page per thousand user, Post Type. On the other hand, the Video Viewers regression model explains around 92% of the variation in Unlike page per thousand user, Fans Engagement and Number of Post Shares. The dependent variables in both model has a p-value of less than 0.0001 which indicates both model are strong models. The text modelling technique formulated 8 topics which give SGAG a better insight of the performance of the posts and help to formulate their content strategy based on audience preferences.

Acknowledgement

We would like extend our gratitude to our sponsor, Mr Karl Mak for entrusting the team with SGAG's social media content analysis of their Facebook posts. We are really thankful and appreciate Mr Karl's support and the prompt response to our enquiries.

We would also like to thank Professor Kam Tin Seong, our project supervisor, for his kind assistance with the analytics topic, the guidance and valuable suggestions to the team during this period.

References

Chaffey, D. (2016). Global Social Media Research Summary 2016. *Smart Insights*. Retrieved from http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Graham, M. (2003). Confronting Multicollinearity in Ecological Multiple Regression. *Ecological Society of America*, 84(11), 2003, pp, 2809-2815. Retrieved from http://www.auburn.edu/~tds0009/Articles/Graham%202003.pdf
Paul, R. (2006). Multicollinearity: Causes, Effects and Remedies. *Indian Agricultural Statistics Research Institute*. Retrieved from http://www.iasri.res.in/seminar/AS-299/ebooks%5C2005-2006%5CMsc%5Ctrim2%5C3.%20Multicollinearity-%20Causes,Effects%20and%20Remedies-Ranjit.pdf

O'Brien, R. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Springer*, 41:673–690. doi: 10.1007/s11135-006-9018-6. Retrieved from http://web.unbc.ca/~michael/courses/stats/lectures/VIF%20articlea.pdf

Ramachandran, S., & Balakrishan, D. (2015). Reach and Engagement: Making the Most of Social Media Marketing. *Tfm Insights.* Retrieved from http://tfmainsights.com/reach-engagement-making-social-media-marketing/

Boston University School of Public Health. (2013). Introduction to Correlation and Regression analysis. Retrieved from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable5.html

Resources ESRI. (n.d.) Regression Analysis Basics. Retrieved from http://resources.esri.com/help/9.3/arcgisengine/java/GP_ToolRef/Spatial_Statistics_toolbox/regression_analysis_basics.htm

Statistics Solutions. (n.d). What is Multiple Linear Regression? Retrieved from http://www.statisticssolutions.com/what-is-multiple-linear-regression/

Wittman, C. (2012). The Simple Reason Facebook Pages are Losing Reach: Negative Feedback. *Social Fresh*. Retrieved from https://www.socialfresh.com/facebook-negative-feedback/

Saleh, K. (2015). Facebook Advertising Statistics and Trends [Infographic]. Invespcro. Retrieved from http://www.invespcro.com/blog/facebook-advertising-statistics/

Uyanık, G., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877042813046429

Sanche, R., & Lonergan, K. (2006). Variable Reduction for Predictive Modeling with Clustering. Retrieved from https://www.casact.org/pubs/forum/06wforum/06w93.pdf

British Dental Journal. (2005). Problems of Correlations Between Explanatory Variables in Multiple Regression Analyses in the Dental Literature. Retrieved from http://www.nature.com/bdj/journal/v199/n7/full/4812743a.html

Rumsey, D. (n.d.). What a Value Tells You About Statistical Data. *Dummies*. Retrieved from http://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/

Wisnowski, J., Castillo, F., Karl, A., & Rushing, H. (2015). Harness the Power of JMP®: Big Data and Social Media for Competitor Analytics. Retrieved November 20, 2016, from http://www.adsurgo.com/wp-content/uploads/2015/09/JMP-for-Competitive-Intelligence-with-Text-Analytics.pdf

Topic Data: Learn What Matters to Your Audience. (2015). Retrieved November 20, 2016, from https://www.facebook.com/business/news/topic-data

Text Explorer Platform Overview. (n.d.). Retrieved November 20, 2016, from http://www.jmp.com/support/help/13/Text_Explorer_Platform_Overview.shtml

Milley, A. (2016). JMP 13 Preview: Now you can “textcavate” your data with the new Text Explorer. Retrieved November 20, 2016, from http://blogs.sas.com/content/jmp/2016/09/02/jmp-13-preview-now-you-can-textcavate-your-data-with-the-new-text-explorer-platform/

JMP. (n.d.). Why JMP®? Retrieved from https://www.jmp.com/content/dam/jmp/documents/en/software/jmp/jmp13/why-jmp-13.pdf

Appendix

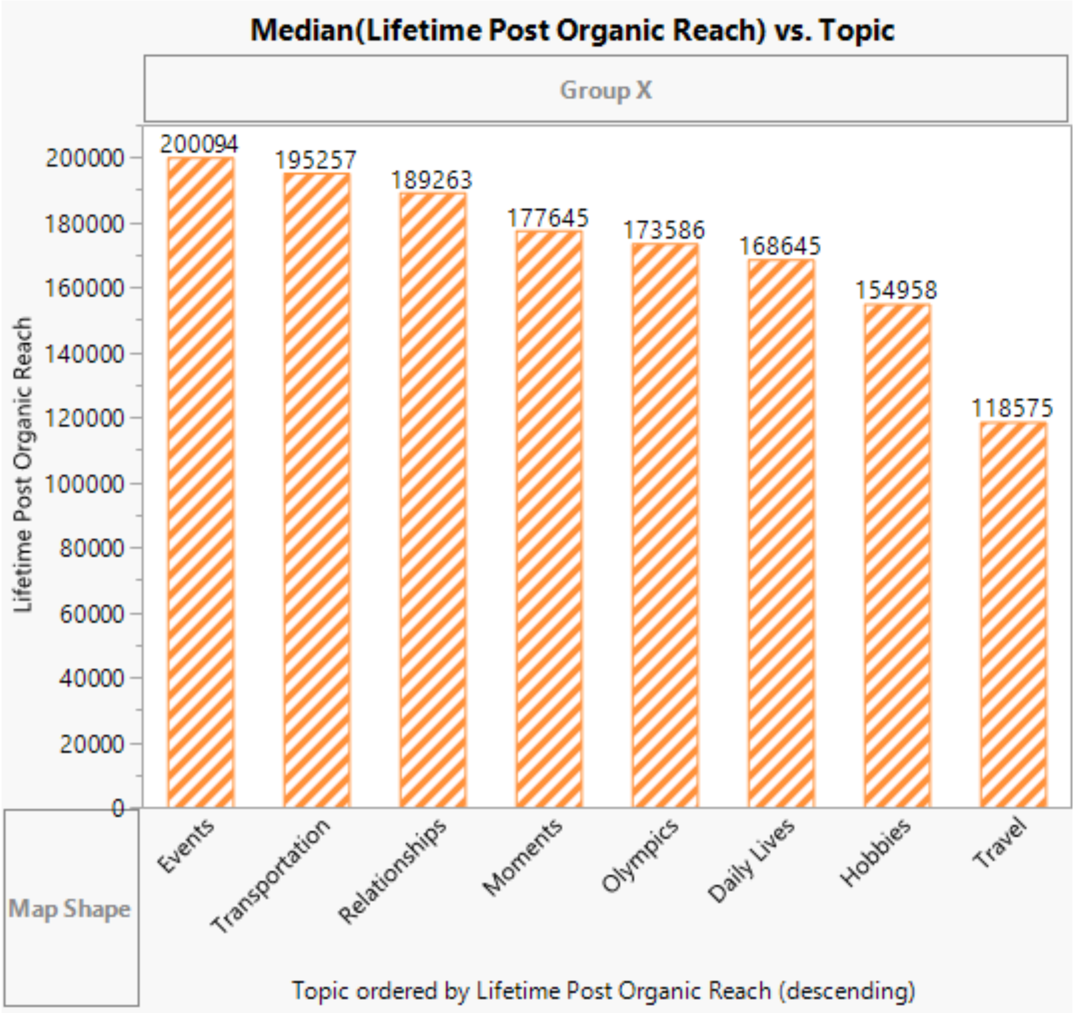


Figure 1 - Reach by Topic