

ANLY482 Analytics Practicum

Professor Kam Tin Seong



PROJECT
PROPOSAL

SOCIAL MEDIA
CONTENT ANALYSIS

T(eam)Roll

Gan Sze Huey

Nur Amirah

TABLE OF CONTENTS

Overview	3
1. About SGAG	4
2. Project Motivation	4
3. Project Objective	5
4. Data Collection and Description	5
4.1 Facebook Insights Data Export - SGAG - Page Level	5
4.2 Facebook Insights Data Export - SGAG - Post Level.....	6
5. Scope of Work	7
6. Proposed Methodology	8
6.1 Data Collection.....	8
6.2 Data Preparation.....	8
6.3 Exploratory Data Analysis	8
6.4 Cluster Analysis	8
6.5 Sentiment Analysis	8
6.6 Topic Analysis.....	8
6.7 Content Analysis	9
6.8 Regression Modelling.....	9
7. Gantt Chart of Work Plan.....	10
8. References	10
Appendix	11

OVERVIEW

SGAG, one of Singapore's leading local humour content creators, maintains popular social media sites, an online website and a mobile application. Creating creative content on a daily basis, SGAG has garnered a significant number of followers on the various platforms. With an aim to achieve growth, SGAG hopes to leverage on their rich pool of data and derive valuable insights towards content creation.

However with limited resources, SGAG could not conduct a comprehensive analysis and harness on the big data available to them. This project aims to uncover valuable insights on SGAG's content attributes in order to achieve audience growth. Using data gathered from SGAG's facebook page for the year 2015, the team hopes to firstly, conduct exploratory data analysis so as to identify overall performance trends. Next, the team will be performing cluster analysis followed by sentiment analysis, topic analysis and content analysis. Lastly, the team will be building a regression model, which includes findings derived from the analysis conducted, in order to predict better performing future posts. With the insights gained, the team will be providing recommendations to enable data driven content creation, thus allowing SGAG to achieve their aim of greater growth.

1. ABOUT SGAG

SGAG is one of Singapore's leading local humour content creators. Distributing their creative content through multiple platforms, including popular social media sites, an online website, and a mobile app, the team at SGAG creates quality content daily to engage and entertain Singaporeans. With the goal to make "every Singaporean's day a better one", the company was founded in 2012 by two Singapore Management University undergraduates. As of today, SGAG has since gained a loyal following and have reached out to more than 300 000 Facebook users and 120 000 Twitter users, in addition to at least 200 000 users through other social media platforms, mobile apps and websites. Looking forward, SGAG aims to achieve greater growth and reach among their customers, especially for their target customers of Singaporean youths, working adults and young families between the ages of 18 to 34 years old.

2. PROJECT MOTIVATION

SGAG's motto is "to make readers laugh at least 5 times a day, 365 days a year". As such, SGAG places emphasis on the quality of their 5 daily posts, to ensure that readers will find their posts humorous with a local twist, a good-natured piece of fun without any intention to hurt. Much of their content focuses on a funny stereotype of everyday Singaporean life which locals are able to identify with.

Although SGAG has been very successful thus far, it also recognises that the online content space is very competitive, with newer players such as "SMRT Feedback" joining in the fray to generate local humour content. As such, SGAG needs to evaluate and improve their content strategy to ensure they stay entertaining and engaging to their audience. However, SGAG faces a few challenges in understanding and thus, leveraging on, their past success factors, which may be summarised as follows:

1. What are the characteristics of a "great" post? SGAG has so far thrived on an intuitive understanding of their customer's content preferences. However, SGAG does not have a concrete or clear picture of the kinds of attributes which they can work on to make a specific post a "great" one.
2. What is audience sentiment on "viral" posts? Are they reacting in a positive or negative manner? SGAG is concerned that "viral" posts become popular because they receive a lot of "hate", which goes against their content philosophy which is to make people "laugh", a positive emotion. Currently, they do not have easy visibility on this aspect.

SGAG hopes this project will be able to utilise a rich pool of historical data to derive insights into the concerns posed above, so that SGAG would be better able to formulate a more relevant content creation strategy.

3. PROJECT OBJECTIVE

The final goal of this project is to offer useful insights for SGAG to formulate a better content creation strategy moving forward. To measure the effectiveness of their content strategy, and at a more granular level, the effectiveness of each individual post, SGAG operationalises effectiveness as "growth" which is defined by an increase in 1) Number of fans, 2) Audience reach, and 3) Engagement with audience members. This last indicator is further measured by the number of times audience members perform actions such as "likes", "comments", "shares", "retweets" or clicking on links to find out more about the content SGAG has to offer.

To do so, we attempt to answer the two main challenges posed by SGAG in a concrete, data-driven manner by performing an in-depth analysis on SGAG's historical data. More specifically, we attempt to address the following analysis requirements:

1. To be able to understand whether a post is popular in a "positive" or "negative" manner
2. To assess the role of content layout and design in improving popularity of posts.
3. To develop a list of common topics and be able to understand the role of topic-selection in affecting the popularity of posts

4. DATA COLLECTION AND DESCRIPTION

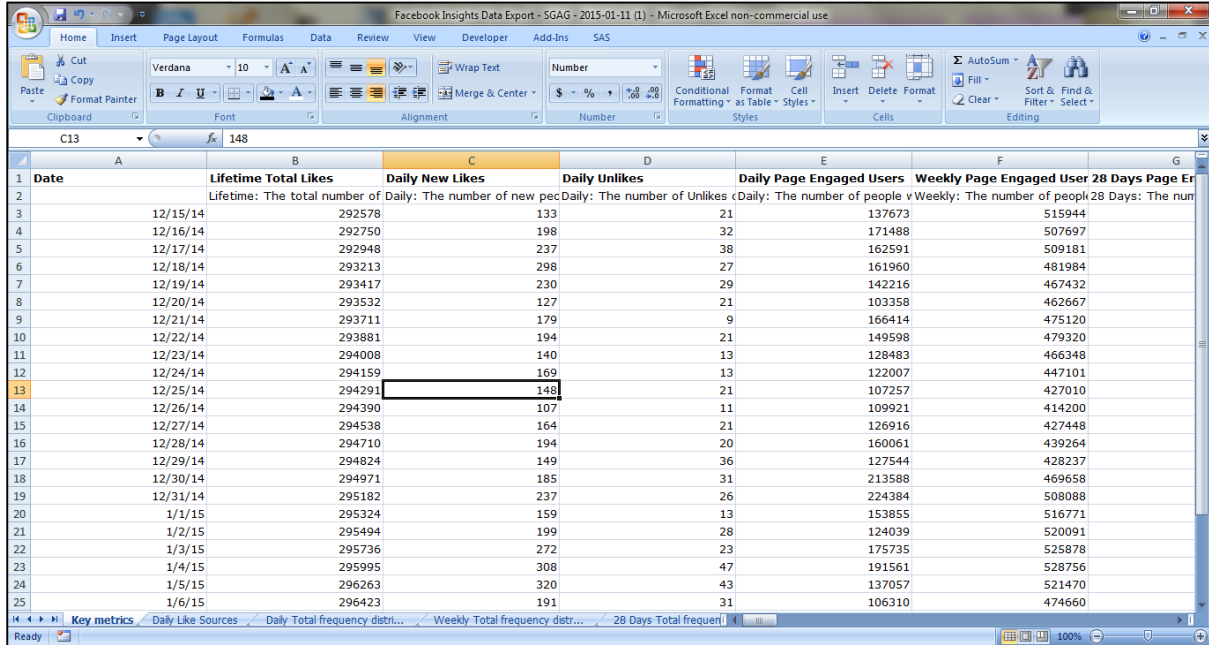
Our two main datasets are: Facebook Insights Data Export - SGAG - Page Level, and Facebook Insights Data Export - SGAG - Post Level. The datasets are sponsored by SGAG and extracted from the Facebook Insights tool. A year's worth of data from 2015 was extracted. Although SGAG also obtained similar data for the same time period from Twitter through Twitter Analytics, this would not be the focus of our project for the present time.

4.1 Facebook Insights Data Export - SGAG - Page Level

This dataset captures key performance indicators of SGAG at the page level. These include variables such as lifetime total likes, new likes, unlikes, number of engaged users, reach, organic reach, number of clicks on content, and number of negative feedback, on the daily level, or aggregated to form weekly and 28 days measures. This dataset also captures information regarding the demographics of SGAG's

customers, their ages and gender, as well as their location in terms of countries and cities.

A snapshot of the data belonging to the "key metrics" tab is shown below:

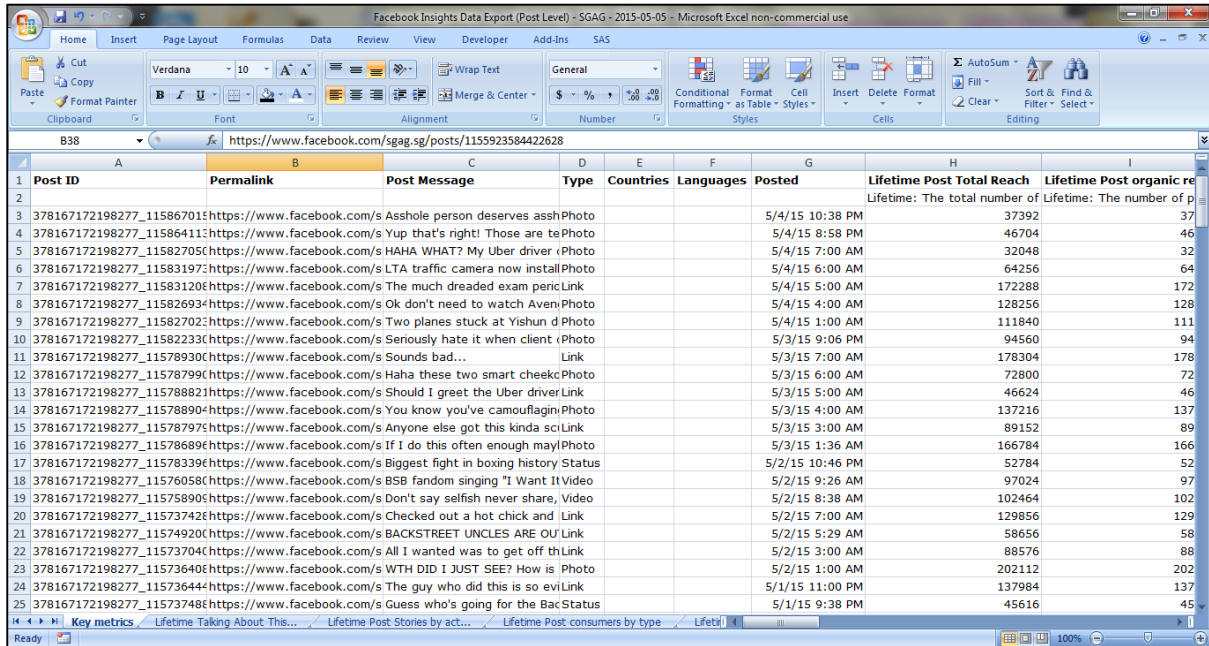


Date	Lifetime Total Likes	Daily New Likes	Daily Unlikes	Daily Page Engaged Users	Weekly Page Engaged User	28 Days Page Engaged User
12/15/14	292578	133	21	137673	515944	
12/16/14	292750	198	32	171488	507697	
12/17/14	292948	237	38	162591	509181	
12/18/14	293213	298	27	161960	481984	
12/19/14	293417	230	29	142216	467432	
12/20/14	293532	127	21	103358	462667	
12/21/14	293711	179	9	166414	475120	
12/22/14	293881	194	21	149598	479320	
12/23/14	294008	140	13	128483	466348	
12/24/14	294159	169	13	122007	447101	
12/25/14	294291	148	21	107257	427010	
12/26/14	294390	107	11	109921	414200	
12/27/14	294538	164	21	126916	427448	
12/28/14	294710	194	20	160061	439264	
12/29/14	294824	149	36	127544	428237	
12/30/14	294971	185	31	213588	469658	
12/31/14	295182	237	26	224384	508088	
1/1/15	295324	159	13	153855	516771	
1/2/15	295494	199	28	124039	520091	
1/3/15	295736	272	23	175735	525878	
1/4/15	295995	308	47	191561	528756	
1/5/15	296263	320	43	137057	521470	
1/6/15	296423	191	31	106310	474660	

4.2 Facebook Insights Data Export - SGAG - Post Level

This dataset similarly captures key metrics of SGAG, but at the post level. Many variables found in the earlier dataset are also reflected in this dataset, but at the post level. We propose that this dataset be our main point of analysis for this project, with the earlier dataset utilised as a supporting analysis.

A snapshot of the data belonging to the "key metrics" tab is also shown below:



Post ID	Permalink	Post Message	Type	Countries	Languages	Posted	Lifetime Post Total Reach	Lifetime Post organic reach
378167172198277_115867015	https://www.facebook.com/s	Asshole person deserves ash	Photo			5/4/15 10:38 PM	37392	37
378167172198277_115864111	https://www.facebook.com/s	Yup that's right! Those are te	Photo			5/4/15 8:58 PM	46704	46
378167172198277_11582705c	https://www.facebook.com/s	HAHA WHAT? My Uber driver «	Photo			5/4/15 7:00 AM	32048	32
378167172198277_115831977	https://www.facebook.com/s	LTA traffic camera now install	Photo			5/4/15 6:00 AM	64256	64
378167172198277_115831206	https://www.facebook.com/s	The much dreaded exam peric	Link			5/4/15 5:00 AM	172288	172
378167172198277_115826934	https://www.facebook.com/s	Ok don't need to watch Aven	Photo			5/4/15 4:00 AM	128256	128
378167172198277_11582702c	https://www.facebook.com/s	Two planes stuck at Yishun d	Photo			5/4/15 1:00 AM	111840	111
378167172198277_11582233c	https://www.facebook.com/s	Seriously hate it when client	Photo			5/3/15 9:06 PM	94560	94
378167172198277_11578930c	https://www.facebook.com/s	Sounds bad...	Link			5/3/15 7:00 AM	178304	178
378167172198277_11578799c	https://www.facebook.com/s	Haha these two smart cheek	Photo			5/3/15 6:00 AM	72800	72
378167172198277_115788821	https://www.facebook.com/s	Should I greet the Uber driver	Link			5/3/15 5:00 AM	46624	46
378167172198277_115788904	https://www.facebook.com/s	You know you've camouflagin	Photo			5/3/15 4:00 AM	137216	137
378167172198277_115787975	https://www.facebook.com/s	Anyone else got this kinda sci	Link			5/3/15 3:00 AM	89152	89
378167172198277_11578689c	https://www.facebook.com/s	If I do this often enough may	Photo			5/3/15 1:36 AM	166784	166
378167172198277_11578339c	https://www.facebook.com/s	Biggest fight in boxing history	Status			5/2/15 10:46 PM	52784	52
378167172198277_11576058c	https://www.facebook.com/s	BSB Fandom singing "I Want It	Video			5/2/15 9:26 AM	97024	97
378167172198277_115758905	https://www.facebook.com/s	Don't say selfish never share,	Video			5/2/15 8:38 AM	102464	102
378167172198277_11573742c	https://www.facebook.com/s	Checked out a hot chick and	Link			5/2/15 7:00 AM	129856	129
378167172198277_11574920c	https://www.facebook.com/s	BACKSTREET UNCLES ARE OU	Link			5/2/15 5:29 AM	58656	58
378167172198277_11573704c	https://www.facebook.com/s	All I wanted was to get off th	Link			5/2/15 3:00 AM	88576	88
378167172198277_11573640c	https://www.facebook.com/s	WTH DID I JUST SEE? How is	Photo			5/2/15 1:00 AM	202112	202
378167172198277_115736444	https://www.facebook.com/s	The guy who did this is so evi	Link			5/1/15 11:00 PM	137984	137
378167172198277_11573748c	https://www.facebook.com/s	Guess who's going for the Bac	Status			5/1/15 9:38 PM	45616	45

5. SCOPE OF WORK

Our proposed work scope will focus on the main content distribution channel SGAG currently uses, which is Facebook. This would be where SGAG garners the most reach and engagement from their target audience. We will also be conducting our analysis based on historical Facebook data for the year 2015, which is suitable due to it being relatively recent.

A step-by-step breakdown of our proposed scope of analysis is as follows:

1. Data Collection – Collect Facebook data for the year 2015 to be analysed, from SGAG
2. Data Preparation – Clean and transform data into a readable CSV for upload
3. Exploratory Data Analysis - Identify overall performance trends
4. Cluster Analysis – Perform segmentation of Facebook posts based on their performance in terms of total reach and engagement level (likes, shares, comments)
5. Sentiment Analysis – Identify differing sentiments based on posts and clusters
6. Topic Analysis - Generate and identify topics based on posts and clusters
7. Content Analysis - Identify key design attributes based on posts and clusters

8. Regression Modelling – Build a regression model that includes success factors derived from analysis, to aid in predicting better performing future posts

6. PROPOSED METHODOLOGY

6.1 Data Collection

Download performance metrics, for the year Jan - Dec 2015, at both Page and Post level, of SGAG's Facebook page from Facebook insights. Conduct data crawling to retrieve comments responses for the various content posted within the same time period.

6.2 Data Preparation

Combine the monthly performance datasets and "response" dataset crawled from the SGAG Facebook page, then select the relevant variables for analysis. The final working dataset will then be transformed into a readable CSV for upload.

6.3 Exploratory Data Analysis

Conduct overall performance analysis on the dataset to identify general trends. For instance, some trends to discover could include seasonality trends in customer engagement across the year, differing engagement levels across different age groups, average number of likes, shares and comments across all posts, and how such indicators are distributed across all posts.

6.4 Cluster Analysis

Based on the performance metrics of reach and engagement level (likes, shares, comments) we will conduct cluster/segmentation analysis on the dataset to identify different clusters of posts with different effects on the performance metrics. For instance, top performing posts, debatable posts, etc. We propose using software tools such as SAS Enterprise Guide to aid in the analysis.

6.5 Sentiment Analysis

Based on the performance clusters derived above, we use Sentiment Analysis to dig deeper and uncover how customers' sentiment can affect the performance ratings of different posts. We will conduct text mining and sentiment topic analysis to discover "happiness" or "hate" levels on different types of posts, taking reference from previous studies on sentiment analysis on social media. We propose using the Text Mining module on SAS Enterprise Miner to aid in the analysis

6.6 Topic Analysis

Based on the performance clusters derived above, we use Topic Analysis to uncover popular themes and topics that customers' are interested in, and their impact on

performance ratings. Some example of themes/topics include, National Service Stories, Government Policies Stories, Funny Viral Stories, Working Life Stories, etc. SGAG is also interested in understanding how different topics appeal to different age groups, and if there are any overarching topics that appeal greatly across all age groups. We will use a mixed method comprising of text and topic mining on "responses" to generate possible topics, supplemented by sampling and manual theme coding to discover any potential lesser-known topic. We propose using the Text Mining module on SAS Enterprise Miner to aid in the analysis.

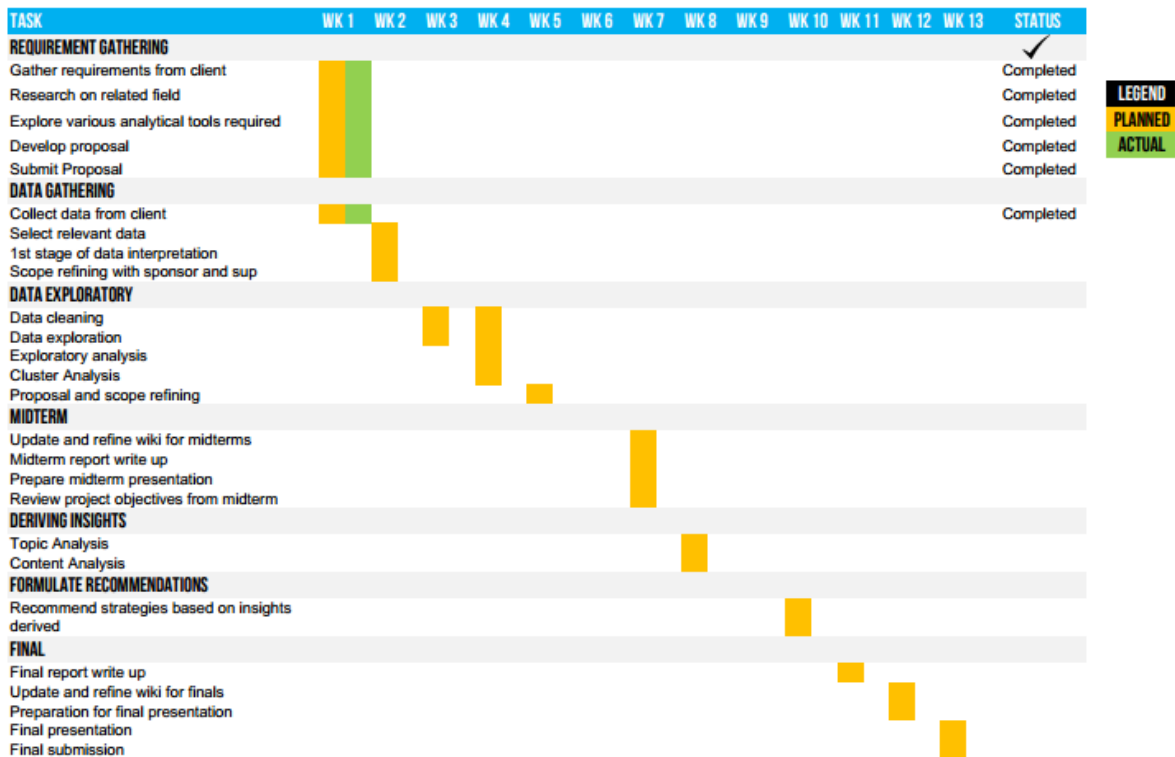
6.7 Content Analysis

Based on the performance cluster derived above, we use Content Analysis to uncover the effects of key design attributes in affecting performance ratings. Some key design attributes include: 1) the number of picture frames used, 2) the kinds of characters used (for instance, common SGAG characters, foreign celebrities, local celebrities, political figures, etc.), and 3) the number of words used. We propose using sampling techniques to identify a representative sample to perform content analysis on, since most of SGAG's content is pictorial and likely to require manual observation and recording of design attributes to be analysed.

6.8 Regression Modelling

Lastly, based on all the performance insights derived from the various analysis above, we propose to use a multi-linear-regression model to assess the overall effect of the success factors derived above, on performance. This model would enable SGAG and the team to understand if all of these factors are sufficient in answering the question "what makes a great post?", or if further studies are required to uncover more factors to improving performance. The model could also serve as a useful scoring tool to gauge future content generated by the creative team, if they should meet SGAG's target performance levels. We propose using SPSS or SAS Enterprise Guide to aid in this analysis.

7. GANTT CHART OF WORK PLAN



(Refer to appendix for clearer image)

As there are only two members in the team, we expect both members to be equally engaged and contribute to each stage of the project together in a collaborative manner.

8. REFERENCES

- (i) SGAG (2015). SGAG Deck for SMU.

Appendix

