ANLY482 Analytics Practicum

Professor Kam Tin Seong

PROJECT

INTERIM

REPORT

SOCIAL MEDIA
CONTENT ANALYSIS

T(eam)Roll

Gan Sze Huey

Nur Amirah

TABLE OF CONTENTS

1. OVERVIEW & FEEDBACK GAINED

1.1 Proposal Recap

Our client is SGAG, one of Singapore's leading local humour content creators. Their goal is to achieve audience growth with whom they share their creative content across various social media platforms. Through this collaboration with SMU, they hope to answer the business question of "what makes a great post", and gain insights to enable them to create popular content that is data-driven by an understanding of audience preferences rather than raw intuition.

There are two main project objectives proposed to answer the above business question: 1) to assess the role of content layout and design in improving posts' popularity, and 2) to develop a list of common topics and understand the role of topic selection in affecting the popularity of posts. Our data comprises of one year's (Jan-Dec 2015) worth of post and page level performance metrics extracted from the Facebook insights tool. Key analyses methods to be used include: cluster analysis, sentiment analysis, topic analysis, content analysis and regression modelling.

1.2 Feedback Gained

We have obtained feedback regarding our proposed analysis methods from both our project supervisor, Prof. Kam, and our project sponsor, Mr. Karl, on separate discussion meetings. While the details of the feedback session may be found in our meeting minutes documentation, the main areas for improvement were:

1.2.1 Focus on social media Content Analysis

As discussed with Prof. Kam, there are two common analysis methods for social media content: social network analysis and content analysis. While the former focuses on identifying social network through graph theories, the latter would be our main focus in identifying key content types and patterns on the SGAG content site. Thus, after discussion, our team has decided to scope out social media content analysis as our project focus.

1.2.2 Removal of Sentiment Analysis as a key analysis method

As discussed with Prof. Kam, it was highlighted that sentiment analysis is a method used to discover respondent's thoughts and feeling regarding specific content, through their responses. There have however been discrepancies between the face-value of textual responses, as opposed to respondent's actual thoughts and feelings, which causes less accurate sentiment analysis findings. Some examples where such a discrepancy occurred included the 2015 Singapore General Elections, where social media favourable sentiment towards opposition parties did not realise in actual voting results (Yi, 2015).

Further, text mining as a method for sentiment analysis would also require a large set of data for more accurate results. On examining the dataset from SGAG, there were a number of challenges which will make sentiment analysis on post responses difficult: 1) responses tend to be names of friends tagged, and thus do not represent a supportive or opposite opinion to the content. 2) responses also have high tendency to be a meme (picture), where text mining techniques would not be suitable. 3) responses tend to have the purpose of further trolling, or inciting more humour, which in itself does not lend to an agreement or disagreement to the content posted.

FIGURE 1: RESPONSE ANALYSIS OF SGAG POST (SGAG, 2015)

In light of the above, we reconsidered the appropriateness of sentiment analysis for SGAG and decided that the current data available was insufficient and unsuitable for a thorough sentiment analysis to be performed. With Prof. Kam's advise, the team decided to drop this analysis in our methodology.

1.2.3 Need to define clear Analytical Problems

As feedback from our proposal report, we were made aware of the need to define clearer analytical problems and objectives. In particular, Prof. Kam highlighted that we have not considered the limitations of our data and how we planned to overcome these to conduct better analyses. After meeting sessions to examine our dataset, our team has since refined these analytical problems and will elaborate on them further in section 2.

1.2.3 Preliminary proposed Topical Framework needs further refinement

As part of our data preparation, the team planned to identify the topics captured by each post. As discussed with our project sponsor regarding how they decided on topics for content creation, we created a draft list of topics to be used for individual post topic identification. Although our project sponsor gave feedback that the topical framework suggested was good, it was likely that it would require more refinement. Taking this feedback into consideration, the team discussed further how to better capture these topics and themes for better representation, and will elaborate on this method further in section 3.

PROJECT PROGRESS UPDATE

The main progress steps we have taken with our project include: 1) addressing analytical problems, 2) refining topical framework, 3) methodology - data preparation, 4) exploratory data analysis, and 5) revised work scope and plan.

2. ANALYTICAL PROBLEMS AND OBJECTIVE

In re-examining our dataset in more detail, there are two main analytical problems which we identified and want to mitigate:

2.1 Ensuring comparability of posts' performance

SGAG's content is posted 5 times daily, 365 days a year. Similar to most social media content, audience response trickles in over time. For instance, on day 1 of the post being released, the post may garner 89 likes, with the total on day 2 rising to 101 likes, and day 3 ending with 124 likes, etc. For posts to be comparable, it is preferred that posts' performance metrics be measured at the same points of time. For instance, it would be ideal to equally measure posts' net performance on day 10, so that posts are compared on the basis of their performance within the first ten days.

However, we are limited by Facebook insights file export function, where the only data available for export would be the net performance metrics for each post at the point of data export. So a post released on 1 Jan 2016 would record 30 days worth of responses when retrieved on 30 Jan 2016, whereas a post released on 20 Jan would record only 10 days worth of responses when retrieved on 30 Jan 2016. This makes an equal comparison more difficult between posts.

To mitigate this limitation, the team observed growth patterns in responses for a collection of recent posts across January 2016, and found that for the general response types of "comments", "likes" and "shares", net response count tended to stagnate after 14 days of release. For instance, day 10 may end with 700 likes, and day 11 may end with 800 likes. Day 14 may end with 1000 likes, and day 20 may still end with 1000 likes, meaning that there were no additional responses after day 14. Thus, as long as posts' response were harvested after at least 14 days after its release date, posts were typically comparable regardless of differences in their release date. As such, we retrieved data for the twelve-month period 1 Jan 2015- 31 Dec 2015, on 19 Feb 2016, where responses has stabilised and the different posts' performance would be typically equally comparable.

2.2 Attribute selection to represent "growth"

Data extracted from extracted from Facebook Insights records many attributes that could be used to evaluate "growth". Some of these are reflected below:

Performance Metrics
Lifetime Post Total Reach
Lifetime Post organic reach
Lifetime Post Total Impressions
Lifetime Post Organic Impressions
Lifetime Engaged Users
Lifetime Post Consumers
Lifetime Post Consumptions
Lifetime Negative feedback
Lifetime Negative Feedback from Users
Lifetime Post Impressions by people who have liked your Page
Lifetime Post reach by people who like your Page
Lifetime Paid reach of a post by people who like your Page
Lifetime People who have liked your Page and engaged with your post
Comments
Likes
Shares

Most of the various attributes play their part in assessing the extent of SGAG's audience engagement, albeit through different perspectives. For instance, "Total Reach" counts the net number of users the post has reached, while "Organic Reach" counts only the number of new users the post has reached. Such slight differences gives rise the question of which metric would be most suited for our main performance index.

Through consulting with SGAG, the team learnt that they relied most heavily upon the number of "comments", "likes" and "shares" to gauge and measure their own performance. In other words, SGAG typically considered a post to be successful in engaging audience members if it had a large number of likes, comments and shares. The team has agreed to use the same three indicators as our main performance index, so that it is easier to align with SGAG's current needs.

The team also noted that although the three attributes tend to be similarly correlated (posts that are highly liked also tend to be highly commented and shared), there are times when posts may rank higher in number of shares, but lower in number of likes and comments, and vice versa. Thus, using number of "likes" as an overall representation of the response level would not be sufficient. Instead, we propose using a combined score of "likes"+"comments"+"shares" to measure and rank overall performance of the posts. This scoring step would be further analysed and considered before implementation during our regression modeling analysis.

2.3 Analytical Objectives

Taking into account the analytical problems faced, we aim to:

(i) Identify better performing posts and weak performing posts
- Based on clustering engagement level (number of likes, shares and comments)

(ii) Identify the patterns between the characteristics of a picture post and engagement level
- Based on the relationship between engagement level and picture design
  - Picture design is based on the number of description lines, character used (e.g. Animals, Foreign celebrities) and the number of frames

(iii) Identify the pattern between engagement level and the topic discussed
- Based on the relationship between engagement level and picture tags

(iv) Identify the relationship between the performance of SGAG's page and the topics discussed
- Based on the relationship between lifetime total likes (SGAG's page) and frequently mentioned picture tags

(v) Identify the relationship between the composition of SGAG's Facebook fans and the topics discussed
- Based on the relationship between the proportion of Male and Female fans from a range of age group and frequently mentioned picture tags

(vi) Identify time-series pattern (eg. the day and time posted) and cluster posts according to engagement level.
- Based on the date and time posted for every posts and the engagement level

3. REVISED TOPICAL FRAMEWORK

SGAG's original basic topical framework for content creation was segmented based on age and occupation. Based these segments, the team sampled some posts from SGAG on derived a basic list of potential topics for our framework.

Age Group	Gender	Segment Description	Possible Topics	Topic Examples
18 - 20	Male	National Service, Education	National Service	NS jokes, military scenarios
	Female	University, Education, Dating	Education	School life, exams
21-25	Male	University, Education, Dating	Male Problems	Cars, girlfriend problems
	Female	Graduation, Working	Female Problems	Boyfriend problems, shopping, girls night out
26 - 35	Male	Graduation, Working	Working Blues	Colleagues, bosses, deadlines
	Female	Family, Marriage	Family	Husbands, wives
All	All	Festivals, Breaking News	Festivals	Christmas, Deepavali, Hari Raya
			Breaking News	MRT breakdowns, haze, straits times

However, upon testing the above framework on a small collection of SGAG's posts (~approx. 50 posts), we realised that the framework could not sufficiently capture a lot of other prominent themes. Given the richness of SGAG's content in capturing various aspects of Singaporean life, the above topics were insufficient and we expanded the selection to include other prominent themes:

Age Group	Gender	Segment Description	Possible Topics	Topic Examples
18 - 20	Male	National Service, Education	National Service	NS jokes, military scenarios
	Female	University, Education, Dating	Education	School life, exams
21-25	Male	University, Education, Dating	Male Problems	Cars, girlfriend problems
	Female	Graduation, Working	Female Problems	Boyfriend problems, shopping, girls night out
26 - 35	Male	Graduation, Working	Working Blues	Colleagues, bosses, deadlines
	Female	Family, Marriage	Family	Husbands, wives
All	All	Festivals, Breaking News	Festivals	Christmas, Deepavali, Hari Raya
			Breaking News	MRT breakdowns, haze, straits times
		Generic situations in everyday life	Friends	Meetups, heart-to-heart sessions
			Singaporean Life	HDB, SMRT, COE prices
			Media Entertainment	Movies, celebrity gossip
			Politics	GE2015, policies, parliament

We then tested our expanded framework on a much larger collection of SGAG's posts (~approx. 300 posts). Again, the team realised that although the most salient aspects of a post's content was captured, even the expanded framework was unable to capture smaller nuances in content, such as a play on words, or the intent to flag out undesirable behaviour, etc.

To continue to expand the topical framework and dummy code each individual post according to all the possible topics and themes would result in far too many topical categories and possibly overlapping themes. Another severe limitation to the fixed topical framework would be its difficulty to constantly accomodate new emerging themes and phase out older obsolete themes. Such dynamism would be important to SGAG where as a social media content provider, audience preferences change frequently and it is important for SGAG to stay relevant.

After much discussion, the team decided that a fixed topical framework would not appropriate for analysing such a diverse and dynamic content dataset. Instead, a different topic identification method is required, and the team proposes to use "tagging".

3.1 Tagging Posts - Benefits and Limitations

The team turns to the option of "tagging" content as a flexible way to identify and analyse dynamic content. The notion of "tagging" in our project is similar to the technique of metadata tagging. The concept is not uncommon and can be found in various forms of online content. One example is the Korean celebrity gossip content website, allkpop.com which uses article tags to identify celebrities mentioned in the article, or key programs that were featured. Through these tags, readers can easily get an overall idea of the of content represented in the article. Similarly, through tagging of topics to SGAG's posts, the team too is able to identify topics represented in these pictorial content.

3.1.1 Benefits

A key benefit of using tagging in our analysis is its flexibility to be defined by authors. Any amount of tags may be used, added or dropped, which allows for dynamic identification of new and old content, without being limited to a fixed framework. For instance, if a new post reflects a new form of content in the form of humour stories related to a new movie, "Star Wars: The Force Awakens", then a new tag can be created to reflect this - tag "Star wars, TFA". Details and nuances in content can thus be easily captured and readily used for analysis later on.

Another key benefit of tagging is its innate ability to represent the main ideas of a picture post, in textual form. This is important for the analysis of SGAG's content because there are very limited tools available in the market for rapid picture analysis and segmentation. While textual data can more easily afford deep and insightful analysis through the use of text mining techniques, the same cannot be applied to pictorial data. Using unique tags to convert pictorial data into textual data is an

effective way of rendering pictorial data more easily analysed in large quantities. This is important in our study of thousands of pictorial data.

3.1.2 Limitations & Mitigation

One of the main criticisms and limitations of the "tagging" method is its non-standardisability - the fact that there is no controlled vocabulary system. Analysts noted that since tags are defined by users in a flexible manner, similar ideas in a piece of content may be tagged with different words by different users; the use of synonyms. For instance, to identify an article about "Despicable Me", different users may use the terms "movie", "film" or "cartoon" to represent the same idea. Across multiple collaborators it has been noticed that semantics make textual analysis more difficult later.

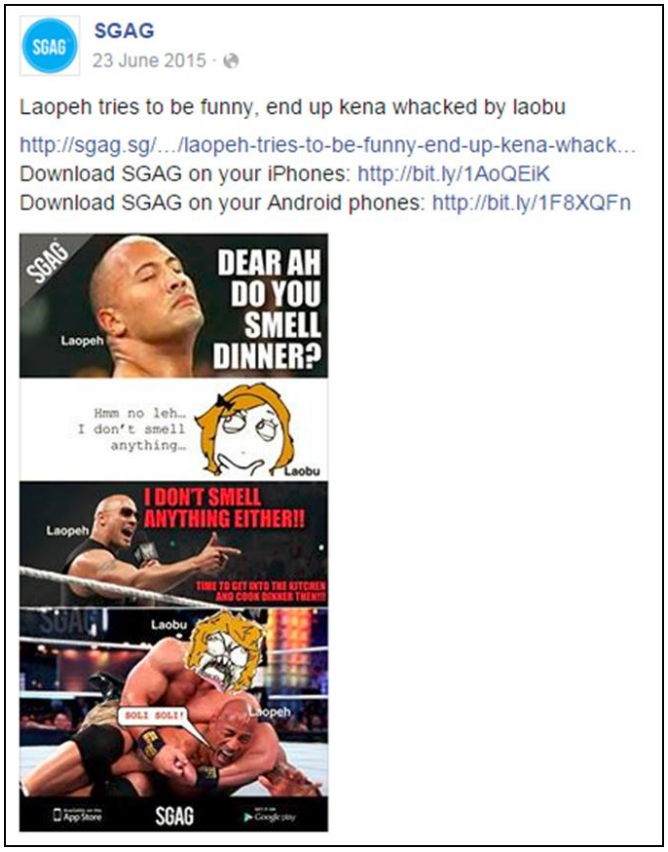
In order to be able to perform segmentation analysis and topic modeling for our project later on, our team has decided to mitigate this limitation with the dual use of a tagging chart, and flexibility in recording details. For each post, we would first tag posts according to the fixed set of 20 generic topics developed. After which, we would include specific tags to capture smaller nuances present in the posts, such as "word puns", "Star Wars", etc. This standard tagging procedure allows team members to operate within an overarching topical structure (parent themes), which reduces the problem of too many synonymns used, while affording a good level of details (children themes) which can be used for detailed analysis later on.

Generic Topical Framework	
Animals	National Service
Breaking News	Politics
Education	Relationships
Female Problems	Sibei Motivation
Festivals	Singaporean Life
Funny Convos	Submissions
Male Problems	Throwback
Media Entertainment	Types of people
MGAG	Weather
Name to shame/honour	Working Blues

3.1.3 Standard Tagging Procedure - Walkthrough Example

This is an example of how the standard tagging procedure is applied. The post may be obtained at (https://www.facebook.com/sgag.sg/posts/1188289691186017:0).

FIG 2. STANDARD TAGGING PROCEDURE EXAMPLE (SGAG, 2015)



Tagging & Design Attributes	Values
Topical framework tags:	relationship, funny convos
Detailed flexible tags:	husband, wife, wife rage, dinner, trying to be funny, john cena
No. of frames:	>3
No. of description lines:	>3
Characters used:	troll faces/memes, foreign celebrity

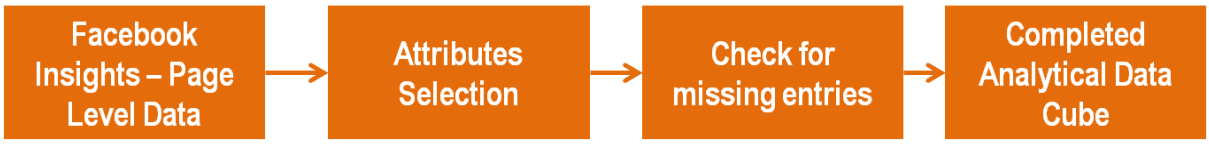
4. METHODOLOGY

4.1 Data Preparation

4.1.1 Facebook Insights - Page Level

Data Preparation for Page Level data was fairly straightforward:

FIG.3 DATA PREPARATION - PAGE LEVEL



Firstly, attributes were selected for analysis, which were

Generic Page Information	Performance Metrics	Audience Demographics (Gender - Age Range)
Date	Lifetime Total Likes	F.13-17
Month	Daily New Likes	F.18-24
	Daily Unlikes	F.25-34
	Daily Negative feedback	F.35-44
	Net Likes	F.45-54
	Daily Page Engaged Users	F.55-64
	Daily Total Reach	F.65+
	Engaged users	M.13-17
	Daily Total Impressions	M.18-24
		M.25-34
		M.35-44
		M.45-54
		M.55-64
		M.65+

Other attributes were removed from analysis after discussion with our project sponsor. The main reasons for doing so would be: 1) our focus on net rather than organic metrics, 2) redundancy of paid posts as no posts were paid for, 3) redundancy of check-in metrics since this was not a hotel or destination services business, and 4) the omission of video formats for current analysis which is limited to pictorial posts.

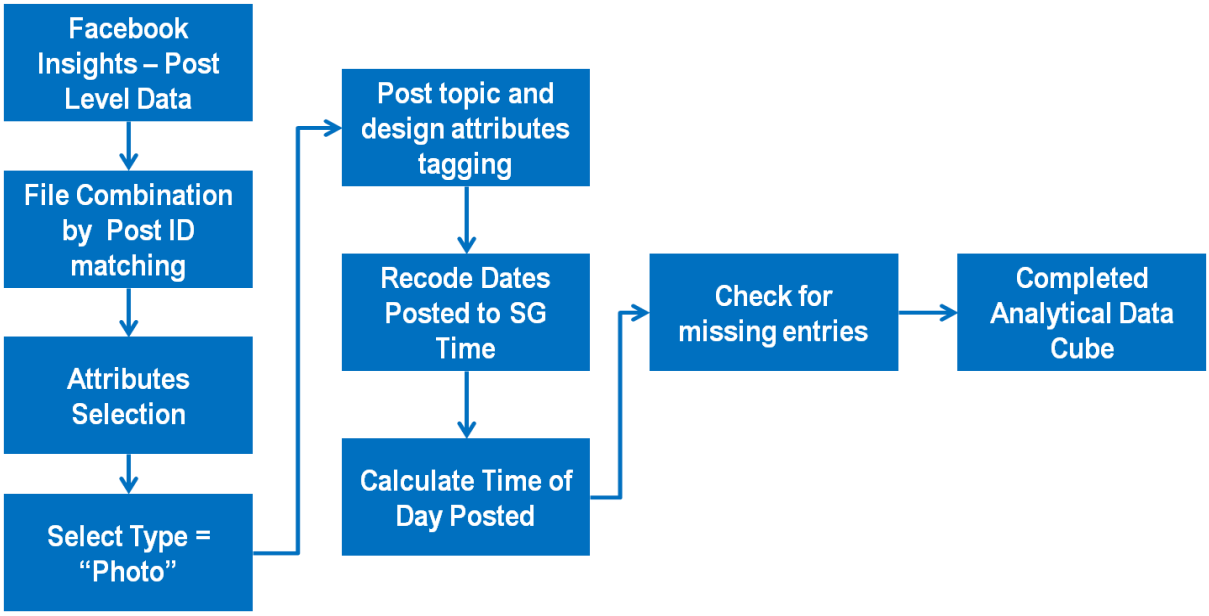
Secondly, we checked the file for missing data entries. Since data was recorded on a daily basis, we found that all days were accounted for our period of review, 1 Jan 2015 - 31 Dec 2015. Thus, there was no need to treat any missing data points.

Lastly, with the above two processes in place, we now have our completed Page-Level analytical data cube ready to be used for data analysis.

4.1.2 Data Preparation - Post Level

Data preparation for Post Level data was more complicated due to its granularity and larger number of observations recorded.

FIG 4. DATA PREPARATION - POST LEVEL



Firstly, post level data extraction resulted in a combination of data sheets each recording different aspects of post performance. Some examples are: Key performance metrics.tab, Lifetime talking about this.tab, Lifetime Negative Feedback.tab, etc. These different metrics are identified for individual posts via unique Post IDs. Thus, the function =index(match) on Excel was used to recombine all these metrics into a single data sheet, matched by Post IDs.

Secondly, attributes were selected for analysis, which were

Generic Post Information	Performance Metrics
Post ID	Lifetime Post Total Reach
Permalink	Lifetime Post Total Impressions
Post Message	Lifetime Engaged Users
Type	Lifetime Negative feedback
Posted	Lifetime Negative Feedback from Users
	Lifetime Post Impressions by people who have liked your Page
	Lifetime Post reach by people who like your Page
	Lifetime People who have liked your Page and engaged with your post
	comment
	like
	share
	hide_all_clicks
	hide_clicks
	report_spam_clicks
	unlike_page_clicks

Similar to page level attributes selection, other attributes were removed from analysis after discussion with our project sponsor. The main reasons for doing so would also be: 1) focus on net rather than organic metrics, 2) redundancy of paid posts as no posts were paid for, 3) the omission of video formats for current analysis which is limited to pictorial posts, and 4) only direct engagement response metrics "like", "share", "comment" and direct negative feedback metrics "hide_all_clicks", "hide_clicks", "report_spam_clicks" and "unlike_page_clicks", were included in addition to general performance metrics since the other optional additional attributes were already well represented by the general performance metrics.

Thirdly, of the three types of post formats recorded ("Types"), we selected only "photo" and removed "links" and "video" formats. "Photo" indicated the bulk of SGAG's content, which are memes. "Link" and "Video" indicated secondary content types of listicles and youtube videos, which will not be the focus of our study.

Fourthly, each selected post was tagged according to their topics and design attributes. As discussed above, we used a textual tagging system to indicate post topics and themes. For design attributes, once again in consultation with SGAG, we identified three main areas of design, namely 1) character use, 2) number of frames, and 3) number of description lines within the picture. Character use had 6 main character types, and dummy coding was used. Number of frames was indicated with either "1", "2" "3" or ">3", since most of the posts were designed to attempt to fall within three or less frames. Similarly, number of description lines is also indicated with either "1", "2", "3" or ">3".

Fifthly, the team noticed that the attributed "Posted" which recorded the date and time of post release contained a large number of posts which were released between 000h-0600h daily. This is strange as these are the sleeping hours of Singaporeans and would thus make no sense for anyone to be releasing the posts so late at night. Through online research, we found that Facebook Insights recorded "Posted" according to Pacific Time rather than local time. As such, there was a need to recode "Posted" forward by 16 hours to match with local SG time. This was done by adding 16 hours (=+0.667) to the previous recorded time. With a check on the newly calculated local time, majority of posts were released within the expected timings of 0900-2100h in local time.

Sixthly, the team examined the data for missing values and found some observations with missing values for performance attributes. The cause for these missing values is not known, though we suspect the cause to be another limitation in Facebook Insights attribute retrieval for specific types of posts, such as linked posts. However, the number of such missing values is very small, comprising around 2% of our dataset. As such, we have decided to omit these missing values during our analysis.

Lastly, with the above processes in place, we now have our completed Post-Level analytical data cube ready to be used for data analysis.

4.2 Revised Analytics Data - Analytical Data Cube

The team's analytical data cube comprises of two parts: Page-Level and Post-Level.

Page-Level data cube comprises performance metrics for the SGAG Facebook page as a whole, recording positive performance metrics such as number of likes, and negative performance metrics such as number of unlikes. There are daily observations for each day of the period 1 Jan 2015 - 31 Dec 2015.

The main purpose for this data cube is in exploratory data analysis, and trends in page performance will be cross compared with trends in post performance and design, which is obtained from the post-level data cube.

Post-Level data cube comprises performance metrics for each individual posted pictorial content on the SGAG Facebook page. Similarly, it records both positive and negative performance metrics, but these are recorded on an individual post basis. Since SGAG posts at least 5 times daily, there are more observations in this data cube, and all pictorial posts from the period 1 Jan 2015 - 31 Dec 2015 are recorded in this data cube.

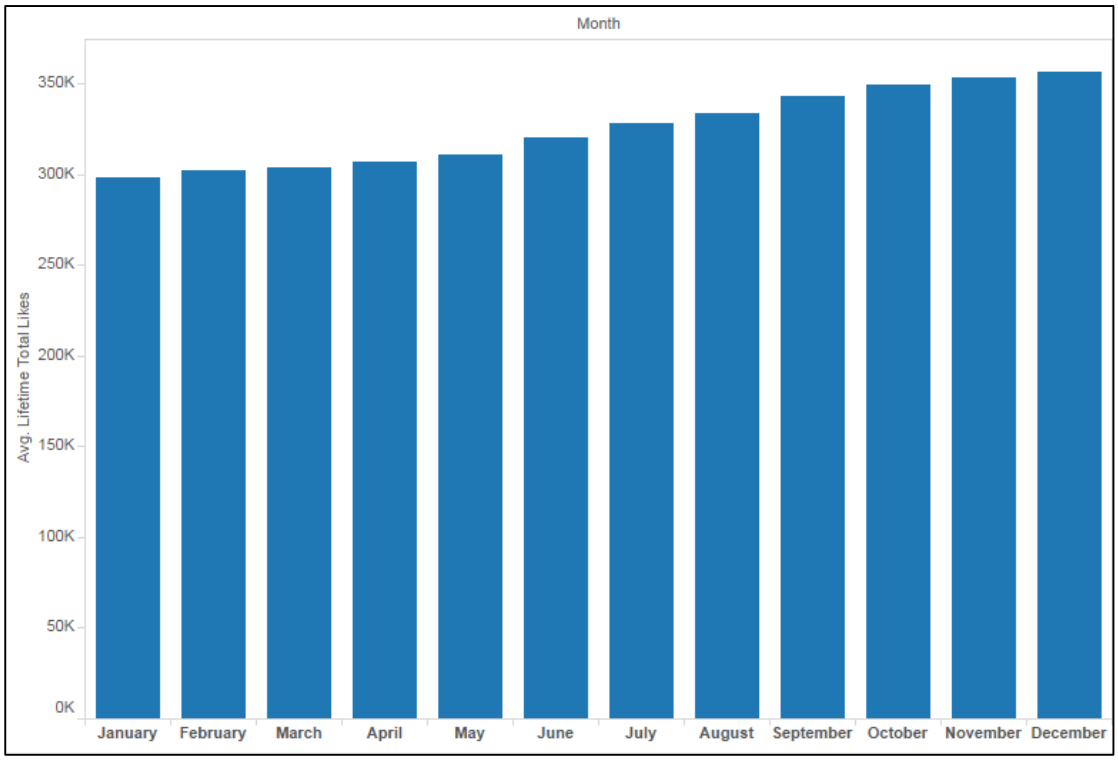
The main purpose for this data cube is in all stages of our analysis, including exploratory data analysis, cluster analysis, topic modelling, content analysis, and multi-linear regression analysis.

4.3 Exploratory Data Analysis - In Progress

The team has currently embarked on exploratory data analysis for both page-level and post-level data cubes. Although the process is still in-progress, we record our progress in this area thus far:

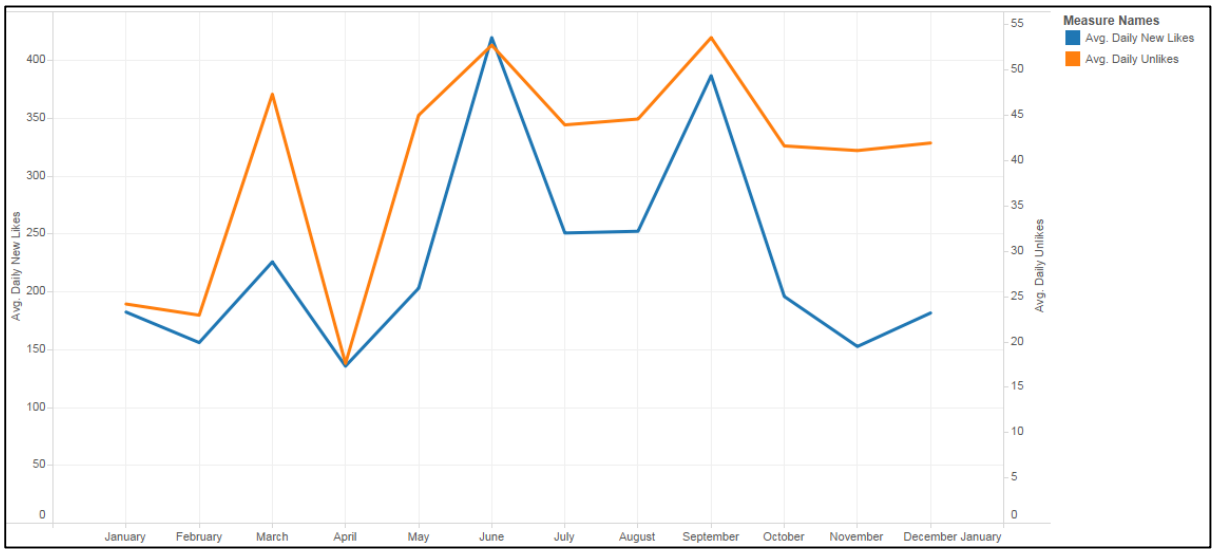
4.3.1 EDA - Page Level

The team looked at changes in **lifetime total likes** of the SGAG Facebook page, which is visualised in the following histogram:



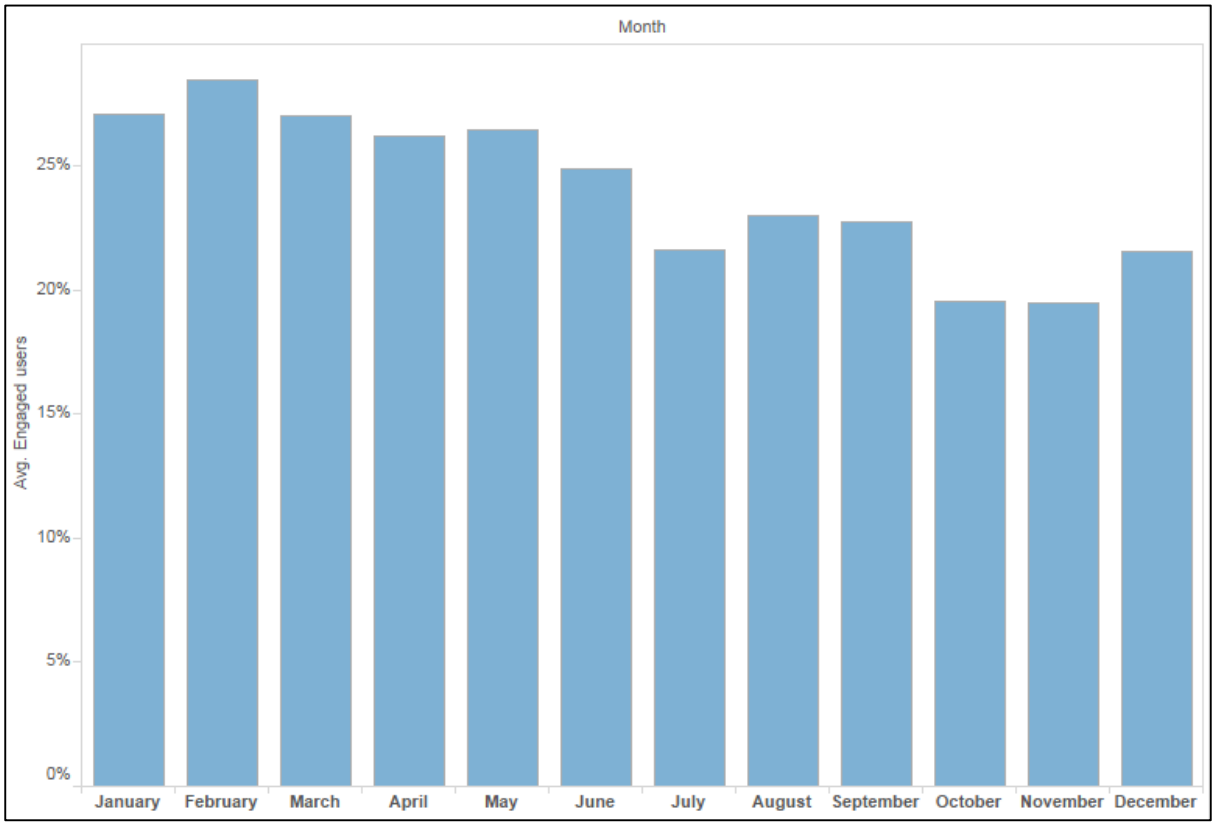
From the months January to December 2015, there is an increasing trend in number of lifetime total page likes. The sharpest increase are in June with 9658 likes, and September with 8923 likes. We hypothesize that this can attributed to content created for two very popular events during the same time period: the Southeast Asian Games (June), and Singapore's General Election (September).

The comparison between **average daily new likes and average daily unlikes** also obtained the following results:



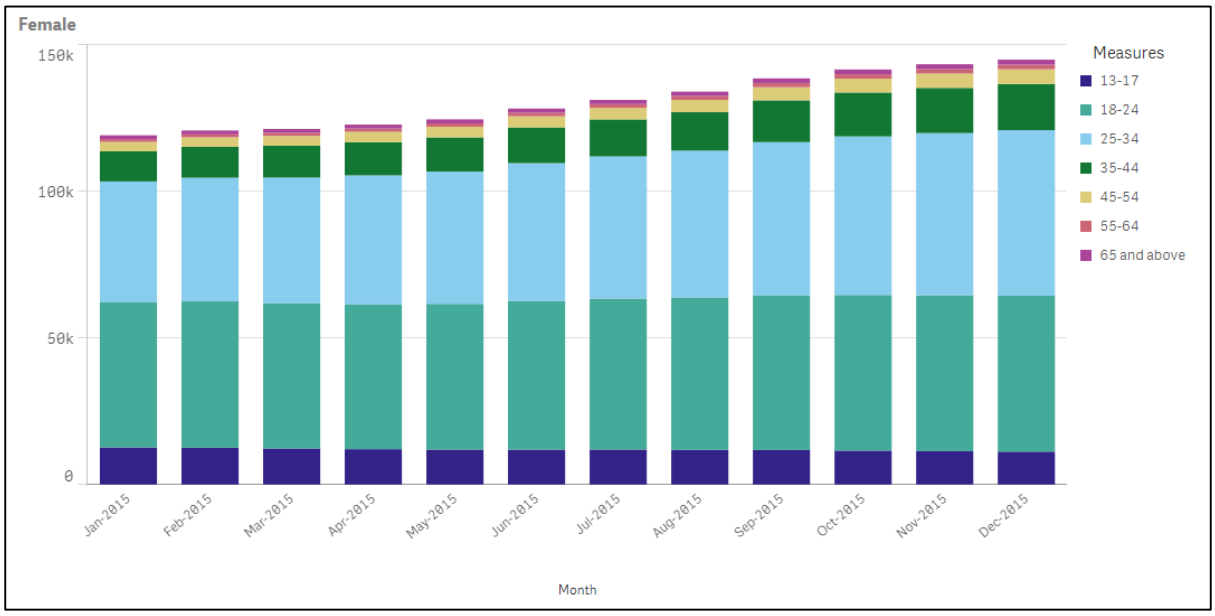
Both daily new likes and unlikes appear to move in the same direction, sharing similar peaks and troughs, with an overall slightly increasing trend. In particular, we note that the average number of daily unlikes increased sharply in March, and dropped to a minimum of 18 in April. We attribute this to the content created in March, the most significant category of which includes the passing of Mr. Lee Kuan Yew, which may have incited some dislikes from "haters" of the late prime minister. In comparison, SGAG typically created feel-good content for the month of April, including funny conversations and everyday aspects of Singaporean life. While content for March may be viewed by some as having political intonations, April's content was comparatively neutral. On the other hand, the average number of daily likes increased most for the month of June. We attribute this to content created for SEA games, which was widely popular. In October, number of likes dropped sharply, which might be the after effect of the popular general elections, and the lack of any similarly notable event in October.

Analysis on the **rate of engaged users** yielded the following:



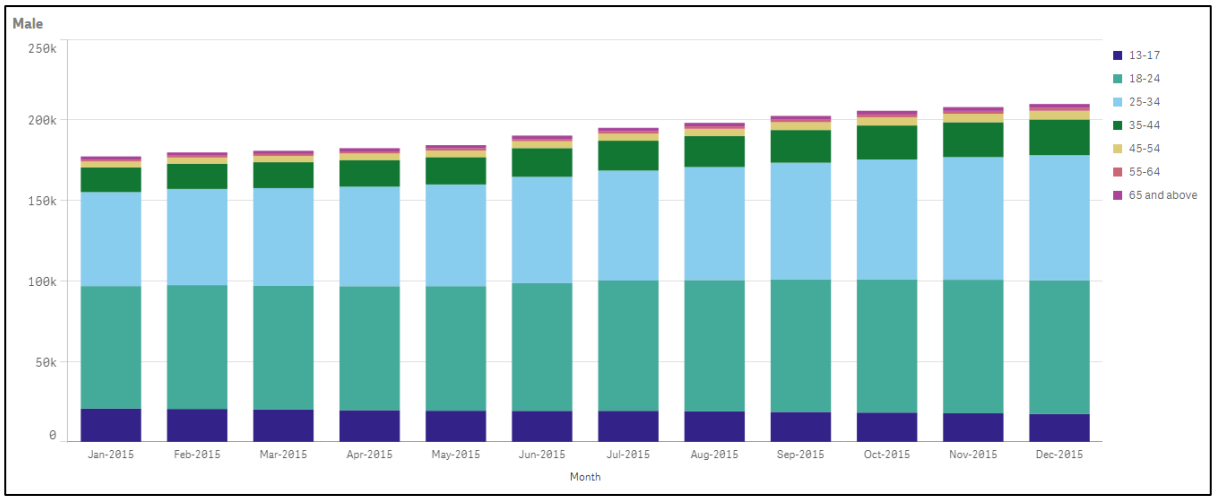
We see that the rate of engaged users generally shows a decreasing trend. Rate of engaged users is the proportion of users who engaged with the SGAG Facebook Page (=daily page engaged users / daily total reach). It is thus interesting to note that although SGAG has earned an increasing number of page likes, this has not necessarily translated into greater audience response and engagement. Thus, SGAG's content generation still needs to enable audience to become more engaged

Female Age Group Demographics



We see that SGAG has indeed successfully reached out to their target group, which audience members of ages 18-34, forming the bulk of their female audience. In terms of growth rate, the segment with ages 25-34 has shown the greatest growth. This is probably because SGAG's content resonated most with this group of audience members.

Male Age Group Demographics



Similarly, we see an increasing trend in the number of page like among SGAG's male audience. Also, SGAG has also had most success in reaching to their target group of 18-34 year old males, with the age range of 25-34 years old showing greatest growth for the year 2015.

In conclusion, we note that SGAG has been successful in communicating to the 25 to 34 years old as seen from the consistent increasing trend in both male and female segments. On the other hand, there is a lack of significant growth from those who

are 18 to 24 years old. In order for SGAG to continue generating greater growth and reaching out to their target audience, contents posted are recommended to better cater to those within the age group of 18 to 24 years old as well.

4.3.2 EDA - Post Level (Proposed Work Plan)

The focus of post level exploratory data analysis would be to observe the distribution of posts across our key performance metrics, "likes", "shares" and "comments". During our data preparation, we noted that the three attributes shared different scales, with likes ranging in the thousands, and shares and comments ranging in the hundreds. Through exploratory analysis, it will be important for the team to better grasp the distribution of the performance metrics as well, to understand whether these are normally distributed, or if the distribution is skewed as we suspect it to be.

The team expects the main results of exploratory analysis to be critical in further preparing our data for cluster analysis in later stages of our study.

5. REVISED WORK SCOPE AND PLAN

In the following 5 weeks, the team expects to complete the following analysis in our workscope: 1) Topic Modeling, 2) Cluster Analysis, 3) Content Analysis, and 4) Regression Modeling.

5.1 Topic Modeling

The focus of Topic Modeling is to sieve out prominent themes from our topic tags. The team aims to take reference to the work of Lai & To : Social Media Content Analysis, A Grounded Approach (Lai & To, 2015) to discover techniques on how to refine our topics for content analysis in SGAG's perspective.

With some of the suggested methodology in mind, we will meet with Prof. Kam shortly to discuss the methods used, and seek further advice on how they may be suitably applied in our project. A key technique likely to be used is text mining, with SAS Enterprise Miner being the main software tool.

5.2 Cluster Analysis

The team expects to use cluster analysis to segment and profile content posts according to their performance indicators. Despite preliminary attempts to use the k-means clustering method to segment our data, the team found that clustering results were less than ideal, since it frequently resulted in one large supercluster, and multiple small clusters. As such, we have decided to revert back to deeper exploratory analysis to better understand the dynamics of performance indicators in our dataset, to mitigate the distribution of these indicators before k-means clustering is attempted again. At the same time, we may explore other clustering methods, such as nearest neighbour or Wald's to identify outliers or anomalies in the dataset. Similarly, the team aims to meet Prof. Kam shortly to seek clarification on these

methods so as to achieve better execution. The main analysis tool will also be SAS Enterprise Guide or JMP.

5.3 Content Analysis and Regression Modeling

With the above two analysis steps completed, the team will use findings derived from the above analysis as inputs for content analysis and regression modeling. Further discussion is required to clarify how content analysis and regression modeling is to be achieved and what further data preparation is required to do so. However, the team currently expects that cluster profiles will reflect certain prominent topics that contribute to the performance of such posts.

6. GANTT CHART OF WORK PLAN

TASK	WK 1	WK 2	WK 3	WK 4	WK 5	WK 6	WK 7	WK 8	WK 9	WK 10	WK 11	WK 12	WK 13	WK 14
Gather requirements from client	N, S													
Brainstorm ideas for analysis	N, S													
Research on social media content analysis	N, S													
Explore analytical tools	N, S													
Develop proposal		N, S												
Submit Proposal		N, S												
Scope refining with sponsor and sup		N, S												
MILESTONE SUBMISSION: PROPOSAL														
Data collection			N, S											
Data cleaning			N, S											
Exploratory Data Analysis: Page-Level						N, S								
Midterm report write up						N, S								
Update wiki								N						
MILESTONE SUBMISSION: MIDTERM														
Exploratory Data Analysis: Post-Level									N, S					
Scope refining with sup										N,S				
Topic Modelling											N, S			
Cluster Analysis											N, S			
Content Analysis											N, S			
Regression Model												N, S		
Meeting with sponsor and sup												N,S		
Preparation for final presentation													N,S	
Update and refine wiki for finals													N	
MILESTONE SUBMISSION: FINAL														
Set up structure for research paper													S	
Research paper													N, S	
MILESTONE SUBMISSION: RESEARCH PAPER SUBMISSION														
Poster													N	
MILESTONE SUBMISSION: POSTER PRESENTATION														

N - Amirah, S - Sze Huey (Refer to appendix for clearer image)

With the above work scope in mind, the team agrees that both members, Amirah and Sze Huey would be equally heavily involved in the various upcoming analytical steps. The team plans to meet weekly to discuss data preparation and specific analysis steps for each type of analysis, while spending the rest of the week executing the discussed methods, before comparing results at over the weekends through online discussions. This method of communication has worked well for the team so far, and we expect to continue to do so.

At the same time, the team is due to give our project sponsor an interim update on our project progress as well, and intends to do so in Week 10, either via a visit to the client's office, or through an email discussion. Furthermore, the team expects more frequent meetings with Prof. Kam to seek advice and clarify the analysis techniques

in the upcoming weeks, and will proceed to book consultation timeslots whenever possible in the upcoming weeks.

Lastly, the team expects to complete all main analysis and record findings by the end of week 13, after which we would prepare the final report and presentation for an in-depth discussion of our study and findings, with our project supervisor and sponsor.

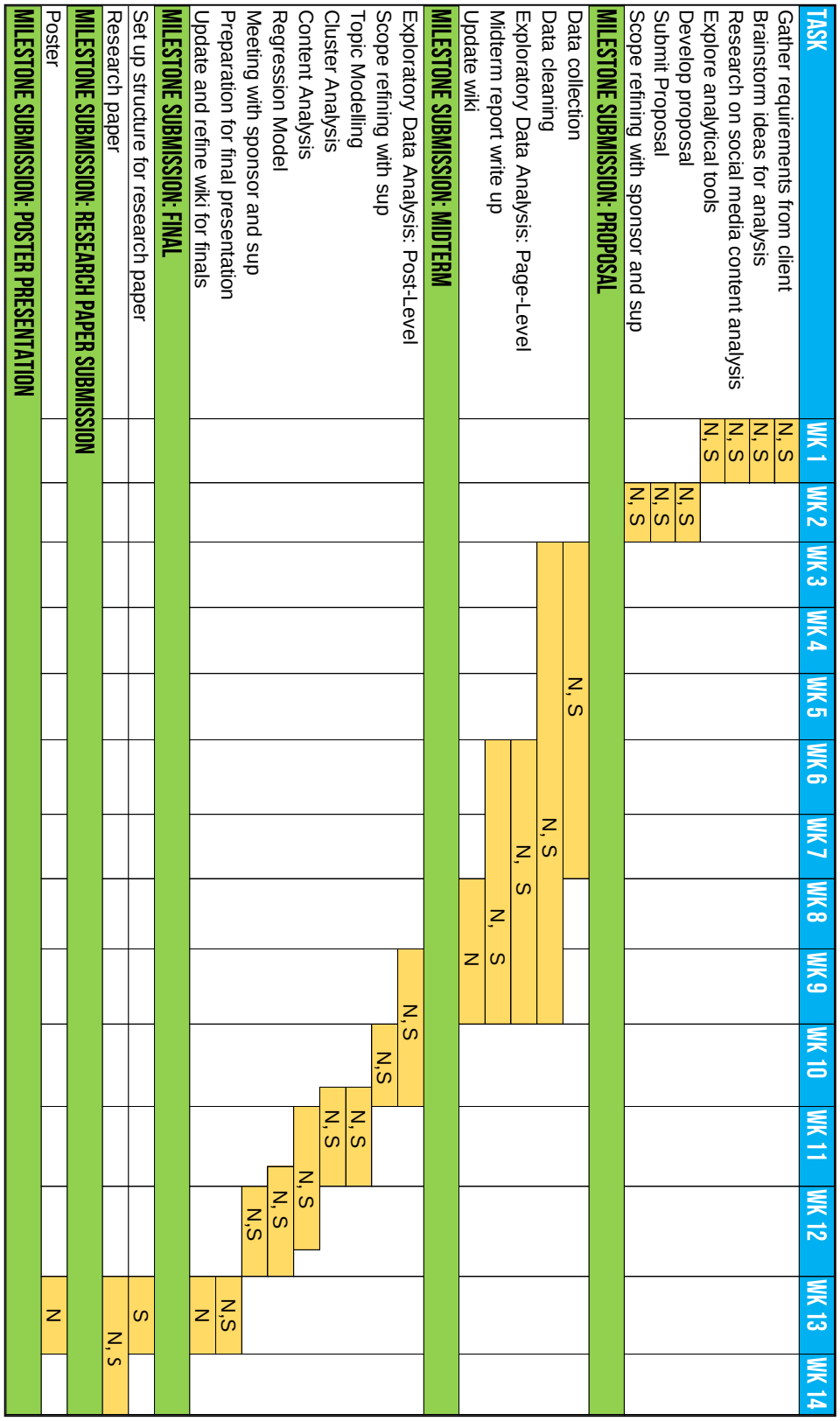
7. REFERENCES

Lai, L. L., & To, W. (2015). Content analysis of social media : A grounded approach. *Journal of Electronic Commerce Research, Vol 15, No. 2*, 138-152.

SGAG. (3 January, 2015). *Humble policeman says I might be handsome but...* Retrieved from SGAG Facebook Page: https://www.facebook.com/sgag.sg/photos/a.378177495530578.106131.378167172198277/1087332334615087/

Yi, H. J. (12 September, 2015). *Channel News Asia, Singapore News*. Retrieved from Channel News Asia: http://www.channelnewsasia.com/news/singapore/opposition-parties/2121768.html

APPENDIX (I)


TASK
WK 1
WK 2
WK 3
WK 4
WK 5
WK 6
WK 7
WK 8
WK 9
WK 10
WK 11
WK 12
WK 13
WK 14
Gather requirements from client
Brainstorm ideas for analysis
Research on social media content analysis
Explore analytical tools
Develop proposal
Submit Proposal
Scope refining with sponsor and sup
MILESTONE SUBMISSION: PROPOSAL
Data collection
Data cleaning
Exploratory Data Analysis: Page-Level
Midterm report write up
Update wiki
MILESTONE SUBMISSION: MIDTERM
Exploratory Data Analysis: Post-Level
Scope refining with sup
Topic Modelling
Cluster Analysis
Content Analysis
Regression Model
Meeting with sponsor and sup
Preparation for final presentation
Update and refine wiki for finals
MILESTONE SUBMISSION: FINAL
Set up structure for research paper
Research paper
MILESTONE SUBMISSION: RESEARCH PAPER SUBMISSION
Poster
MILESTONE SUBMISSION: POSTER PRESENTATION
N, S
N
N,S
S