IS434 - Social Analytics and Applications

Group Project Final Report

G1T7

Done By: Clarice Alyson Ho

Devyn Tan Wei Hung

Liew Su Yee

Lin Cheng Wei

Peh Yi Jia

Prepared for
Professor Kyong Jin Shim

12 November 2019

Table of Content

1. Executive Summary

This report describes the results of social analytics on the company Seamless Bespoke (SB).

There are two business problems:

a. To discover the overall users' sentiments towards SB on HardWareZone (HWZ) and identify who are the main drivers of opinions, whether good or bad, towards SB.
b. To identify the factors that contribute to the failure of their previous marketing campaign and how to increase the likelihood of a more successful marketing campaign in the future.

For problem a, the team will focus on HWZ forum and perform the following analysis:

Topic Modelling	To identify the different type of topics discussed in the forum's fashion/clothes section
Sentiment Analysis	To identify the sentiments among the topics
Social Network Analysis	To identify the impact of the top negative and positive users towards SB

For problem b, the team will focus on Instagram and perform the following analysis:

Demographic Analysis	To identify the types of SB followers
Social Network Analysis	To identify the potential influencers in the first and second layer of the social network

For problem b, the team will also focus on Facebook and perform the following analysis:

User engagement analysis	To identify the content type that will appeal to customers and the best post timing

Note:

Due to privacy concerns, the team decided to mask out the username and gave each user a dummy User ID

- For HardwareZone Forum, from user001 to user144
- For Instagram, the first layer, from user 1 to user 3163
- For Instagram, second layer, from user 0 to user 73185

2. About Seamless Bespoke

Seamless Bespoke (SB) is a bespoke tailoring and shoemaking boutique from Singapore that houses both a boutique and an in-house atelier.

SB is founded since 2012. Data collected from the customers show that most of them visit the shop through word of mouth. It cannot be proven as of yet that their social media platforms are helping to boost the visibility of the shop and increase customer drop-ins. Most of their customers are found to visit the HardwareZone Forum as well but SB is not sure exactly whether the customers are speaking well of them or not.

In their prime, Seamless Bespoke was the 'talk of the town'. It drew huge crowds from the upper-middle class and upper class. However, in recent times, their customer base is on the decline and we would like to help them re-establish the status they once honed in the tailoring sector.

Recently, there has been an increase in negative comments on HardwareZone. Rumours have started and have bred negativity among the users of HardwareZone towards SB as threads extend from negative or sarcastic remarks about SB on HardwareZone. SB hopes to identify the key opinion leaders to better manage and minimise the spreading of negative sentiments among the HardwareZone users.

In addition to that, SB's target audience is ideally the upper-middle class to upper class as their items are made with premium quality and workmanship but coupled with a higher price tag. When SB outsourced their digital marketing to reach out to more potential customers, it was a flop. One possible reason could mainly due to the advertising agency reaching out to the wrong target audience who could not understand or justify the price tagged to the products of SB and as a result, could not drive sales.

3. Business Problem

3.1 The bad/good influence on HardwareZone and how much impact they bring

Currently, there are active discussions on the HardwareZone forum that compare the different tailor shops in Singapore. From the survey results collected by SB, 15% of the customers discovered SB through HardwareZone forum discussions (Refer to Figure 1). SB is concerned about what people are discussing about them on the forum because of the significant influence HardwareZone has on the inflow of new customers, and the impression that existing customers would have of SB.

Our group will help SB have a sensing of the overall users' sentiments towards SB on HardWareZone and identify who the main drivers of opinions, whether good or bad, towards SB are.

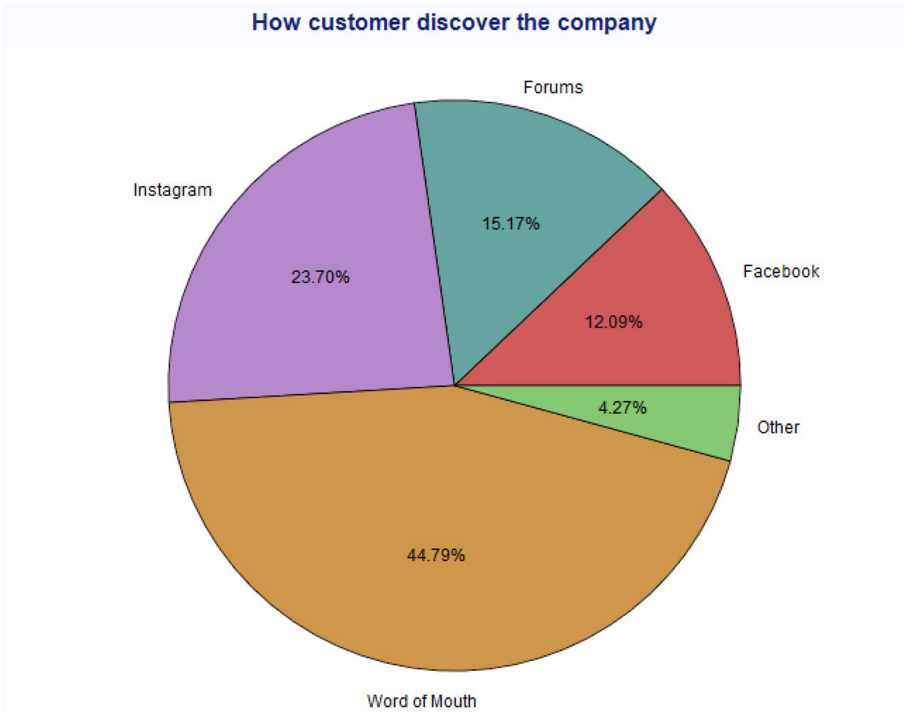

Figure 1: How customers got to know of them

3.2 How to improve from the previous failed marketing campaign

For Great Singapore Sale this year, SB outsourced its Facebook and Instagram marketing campaign to a digital marketing agency. However, the campaign was considered a failure to them as it did not drive more sales.

Our group is tasked to help SB to identify the factors that contribute to the failure of their marketing campaign and how to increase the likelihood of a more successful marketing campaign in the future.

4. HardwareZone (HWZ) Forum

4.1 Problem Statement

To discover overall users' sentiment towards SB on HWZ forum and identify the main drivers of opinions.

4.2 Approach

The team will be doing sentiment analysis and topic modelling on the HWZ posts and form a network graph between the communications.

4.3 Data Information

The time period of data that was decided upon to be scrapped is from 1st September 2018 to 30th September 2019. Any further post data dated further back will not be relevant to determine the impact of HardwareZone on them in recent times. The team started off using an online scrapper (https://github.com/freedaemons/hwz-scrape). However, the team soon came to realise that this scrapper is not able to provide much insight, especially in the area of forming a network graph. This is because the scrapper grouped the replies and original comments altogether.

After further research, the team decided to re-scrape the data using ScrapeStorm. The re-scrapped data showed higher levels of accuracy as replies are now correctly classified to the user who wrote it. The slight limitation with this tool is that each user is limited to scrape 100 posts a day. However, this was easily overcome as there are 5 members in the group, allowing us to scrape up to 500 posts a day.

The data extracted is in the form of a CSV(comma-separated value) file with nine columns. They are the user, date of the post, the message posted, post number in that thread, number of lifetime posts they made, user they are replying to, message they are replying to, user and message replied to and the reply. These data are then cleaned and formatted to prepare for analysis.

4.4 Data Preprocessing

For the post message, we preprocessed the text by removing stop words, irrelevant words (manually identified by us), any weird symbols and digits, as well as performing text contractions and lemmatization.

The section will first cover the topic modelling and sentiment analysis followed by the network graphs.

4.5 Topic Modelling

Latent Dirichlet Allocation (LDA) was used to do the topic modelling. As the number of clusters had to be specified, a for-loop of 3 to 10 clusters was executed. It was then determined that when the number of clusters was set at 5, the clusters had the most distinction and meaning.

The 5 topics identified by LDA:

```
Topic #0:
0.031*"shirt" + 0.024*"fabric" + 0.015*"suit" + 0.013*"jacket" + 0.010*"pant" + 0.007*"price" + 0.007*"budget" + 0.006*"trouser" + 0.005*"lai_en" + 0.005*"people"

Topic #1:
0.041*"suit" + 0.020*"pant" + 0.014*"price" + 0.009*"fabric" + 0.008*"shirt" + 0.008*"sze_sze" + 0.008*"recommend" + 0.007*"budget" + 0.007*"guy" + 0.006*"back"

Topic #2:
0.027*"shirt" + 0.014*"suit" + 0.008*"pant" + 0.007*"recommend" + 0.006*"yeossal" + 0.006*"first" + 0.006*"budget" + 0.006*"fabric" + 0.006*"new" + 0.006*"experience"

Topic #3:
0.022*"suit" + 0.013*"jacket" + 0.012*"shirt" + 0.009*"price" + 0.008*"agree" + 0.007*"fabric" + 0.006*"pant" + 0.006*"feel" + 0.006*"probably" + 0.006*"nice"

Topic #4:
0.016*"shirt" + 0.016*"trouser" + 0.015*"sb" + 0.008*"budget" + 0.008*"pant" + 0.007*"suit" + 0.007*"fabric" + 0.007*"price" + 0.006*"cost" + 0.006*"friend"
```

Figure 2: Screenshot of LDA output

From the above results, these are five main topics discussed by the users on HWZ forum.

	Topic
1	Lai En*
2	Sze Sze*
3	Yeossal*
4	General discussion on tailoring
5	Seamless Bespoke

**Lai En, Sze Sze and Yeossal are the competitors of Seamless Bespoke (SB)*

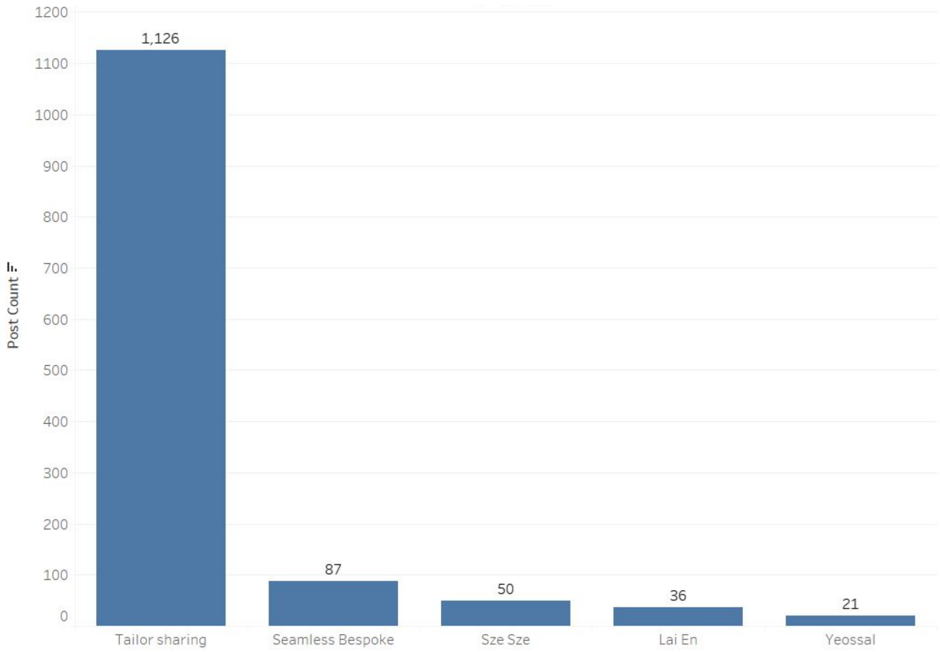

Figure 3: Post count of each topic

Focusing on the cluster of posts that mainly discusses SB, a word cloud was used to find out what did the HWZ users talked about.

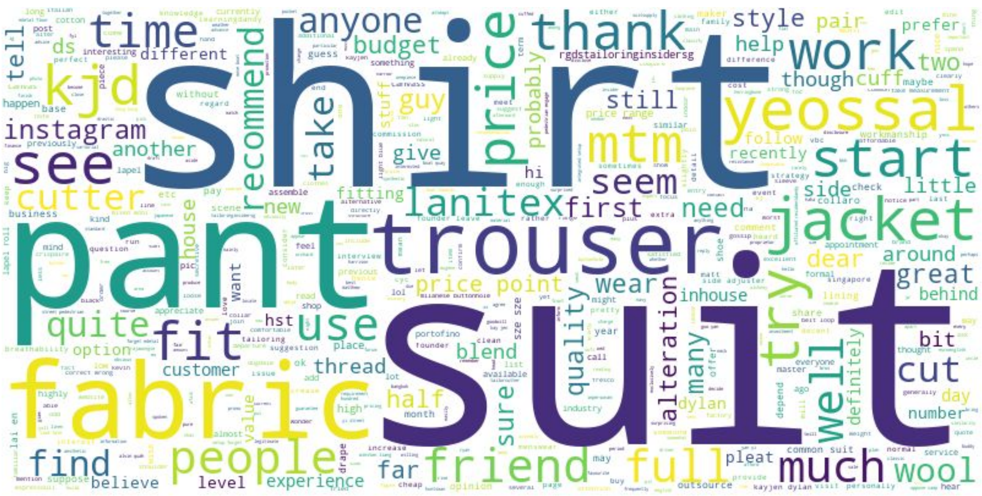

Figure 4: Word Cloud on the topic SB

Based on the above word cloud, the discussion is mainly about “shirt”, “pant”, “suit”, “jacket” and “trouser”. This is expected but is not insightful. Hence, another word cloud was run without those words.

Figure 5: Word Cloud on SB Without Common Words

The second word cloud generated (figure 5) gave more meaningful insights. It indicated that most of the HWZ users compared SB with its competitors. Users compared SB with "Yeossal" and "KJD". It also showed that users are interested in "MTM" suits, as well as fabric "Lanitex" and "Wool" by SB. There were mixed feelings about the "price" point of SB. Those that disagreed merely compared the prices with other tailors as an absolute while others argued for its reasonability on the basis of "quality".

4.6 Sentiment Analysis

Google Natural Language API and TextBlob were used to analyse the post and replies written on HWZ. Vader was also used initially but was subsequently dropped due to the inaccuracy in sentiment scores. The scores generated by the Google Natural Language API and TextBlob are very similar. There are some occasions where both the scores generated do not tally. This tends to happen to post with sarcasm. To account for these cases, the average of both scores were taken. The sentiment score ranges from -1 to 1. The range for the sentiment scores "positive" (green), "neutral" (yellow) and "negative" (red) is shown below.

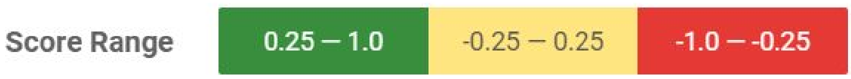

Figure 6: Range of Sentiment Scores

From the aforementioned topics identified by LDA, it was decided that the focus will be on the topics regarding SB and its competitors.

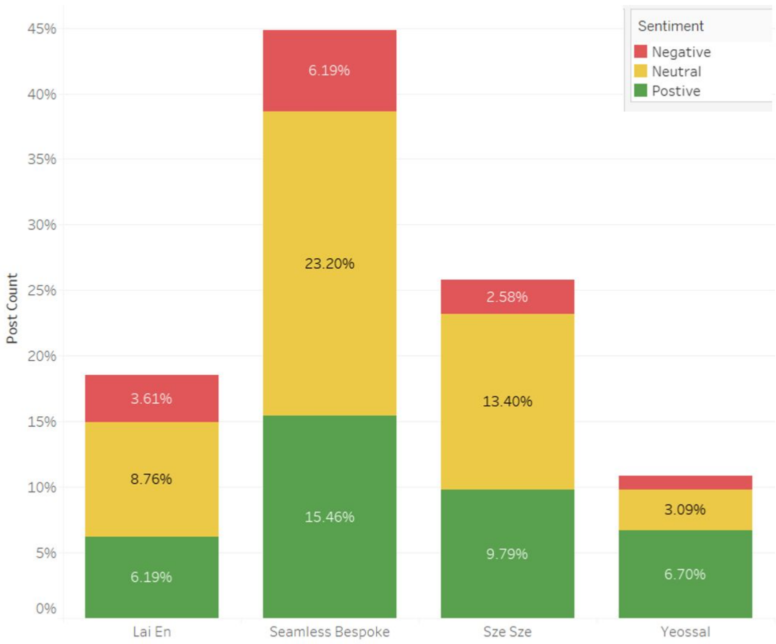

Figure 7: Distribution of Sentiment for Each Topic

Referring to the graph above, SB has both the highest negative and positive posts on HWZ forum. Also, SB has the second-highest negative to positive post ratio.

The top 5 HWZ users that post **positively** about SB is shown below.

	User	No. of Posts
1	user011	3
2	user082	2
3	user135	2
4	user033	1
5	user083	1

The top 5 HWZ users that post **negatively** about SB is shown below.

	User	No. of Posts
1	user135	3
2	user061	2
3	user033	1
4	user082	1
5	user083	1

After identifying the top 5 positive and negative users towards SB, the team wanted to find out how much influence and impact can these users bring. Hence, the next step was to do a social network graph analysis for HWZ.

4.7 HWZ Social Network Graph Information

After filtering out all the isolated nodes in the HWZ social network graph, the final graph is a directed graph with a network size of 144 nodes and 454 of edges, where a node represents a unique user and an edge represents communication between two users.

4.7.1 Degree Centrality

The team first looked into the distribution of degree in the HWZ population.

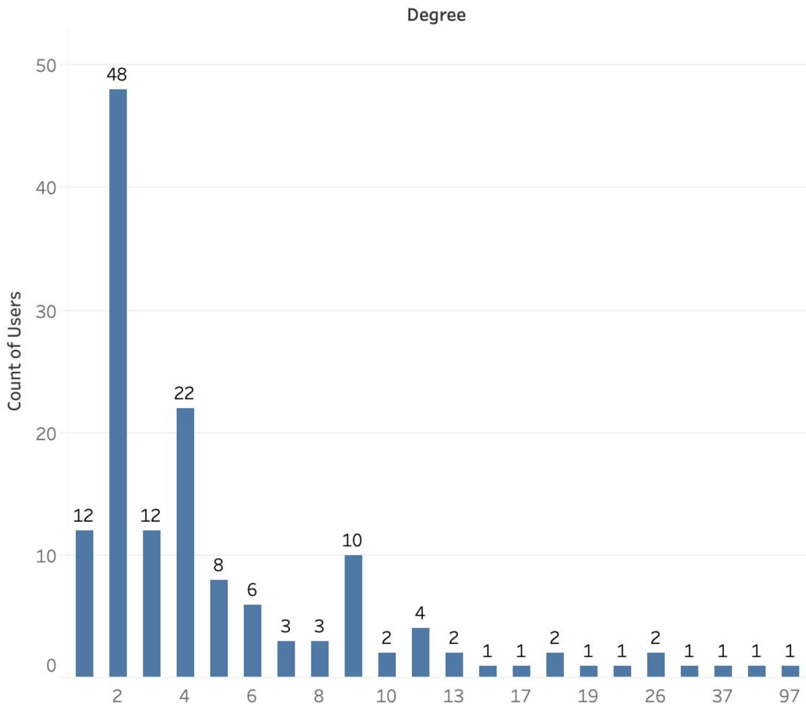

Figure 8: Count of Users by Degree

Based on Figure 8, it was seen that only 1 user has the highest degree of 97, which meant that the user managed to communicate with more than half of the population in the HWZ social network. It was also seen that half the population (72 users) only communicated with a maximum of 3 users.

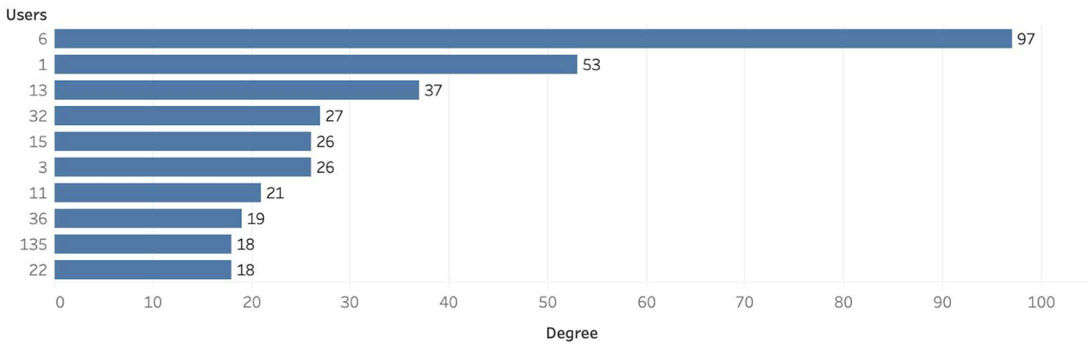

Figure 9: Top 10 User on Total Degree

Based on Figure 9, the team identified **user 6, user 1, user 13, user 32 and user 15** (listed in order) as the top 5 most connected users based on total degree.

Referring to the network graph below, with degree centrality as the node size and replies as the edge weight, there are four big communities in the HWZ social network, which are denoted by green, pink, blue, and purple nodes. In the green community, user 6 has the highest degree centrality. User 1 (second highest degree centrality) is the ego in blue community. User 13 and user 32, the third and fourth highest in terms of degree centrality, are the biggest nodes in the purple community. User 15 (fifth-highest degree centrality) is the ego in the pink community.

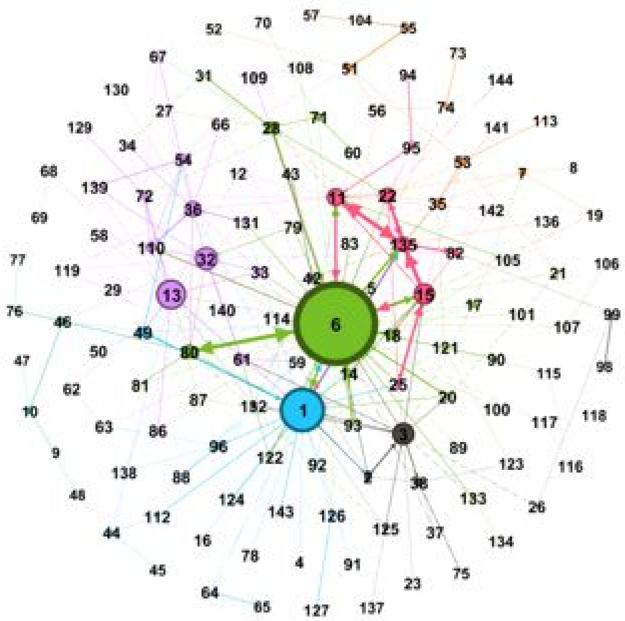

Figure 10: Social Network Graph of HWZ (Node size: Degree; Edge weight: Replies)

The team decided to look into in-degree and out-degree specifically as this is a directed graph. In-degree refers to the reply a user receives from another user whereas out-degree refers to the user replying to another user.

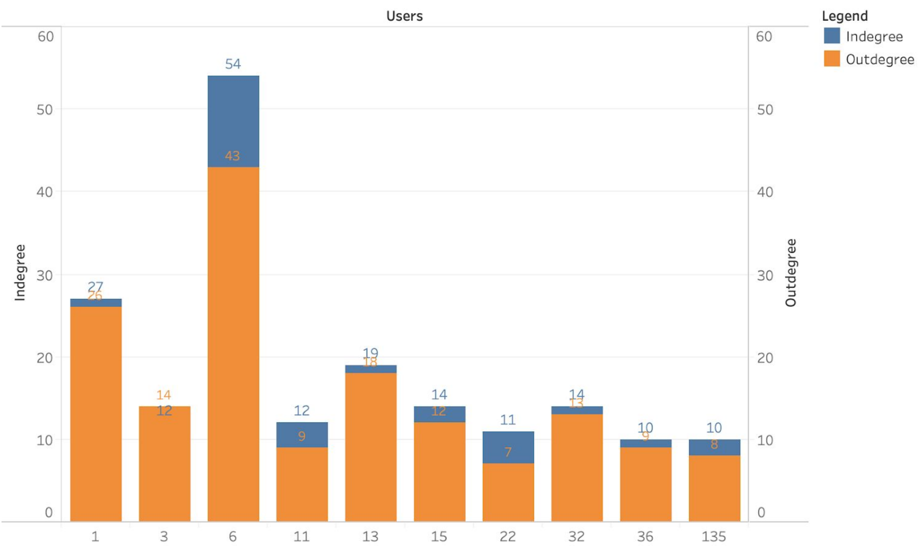

Figure 11: Top 10 Total Degree (In and Out Degree) User

Based on Figure 11, the top 10 identified users seem to be very active. They have been communicating with many other users and have loyal followers. This is seen from their posts/replies often receiving responses from other users.

However, just based on the total degree centralities alone, it is not possible to determine how fast and far good/bad news can spread among these 144 users.

4.7.2 Closeness Centrality

The top 10 users in terms of closeness centrality are shown below.

User ID	Closeness Value
6	0.512397
1	0.460967
13	0.423408
80	0.420339
49	0.409241
110	0.400000
15	0.400000
3	0.396166
18	0.396166
33	0.389937

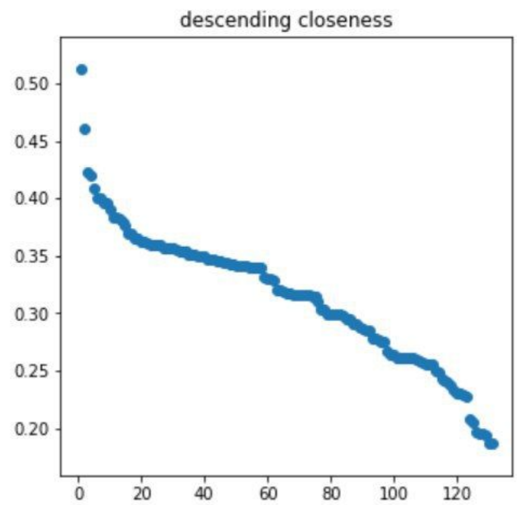

Figure 12: Graph of Closeness Centrality against UserID

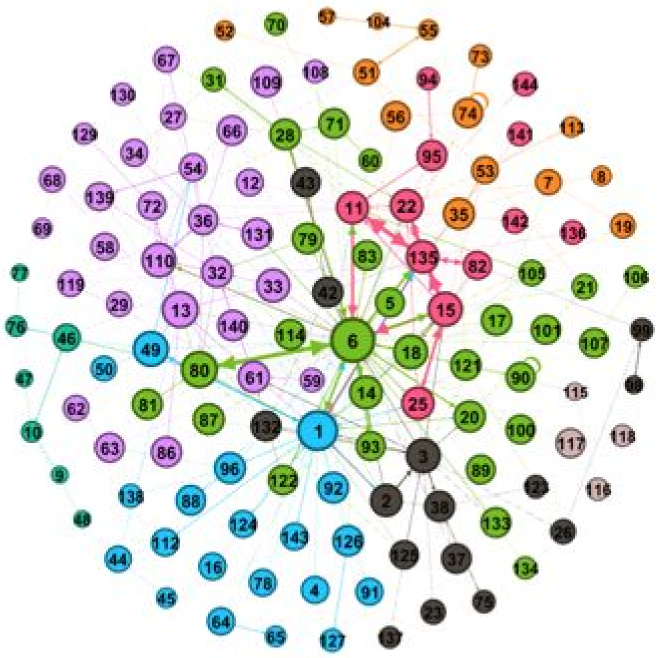

Figure 13: Social Network Graph of HWZ (Node size: Closeness; Edge weight: Replies)

As observed in Figures 12 and 13, the closeness centrality scores are very close to each other, making most of the node size are similar. Regardless of the topic discussed, users still reply to one another at a reasonably constant rate. This meant that information travels very quickly in this network. When B replies to post by A and C replies to B's reply, C is in some way replying to A's post as well.

4.7.3 Betweenness Centrality

The table below shows the top 10 users in terms of betweenness centrality.

User ID	Betweenness Value
6	8877.15737
1	3479.46567
13	2630.03318
46	1482.42395
80	1275.7485
15	1149.60029
32	1145.50057
51	1136.97692
49	966.350719
3	821.942319

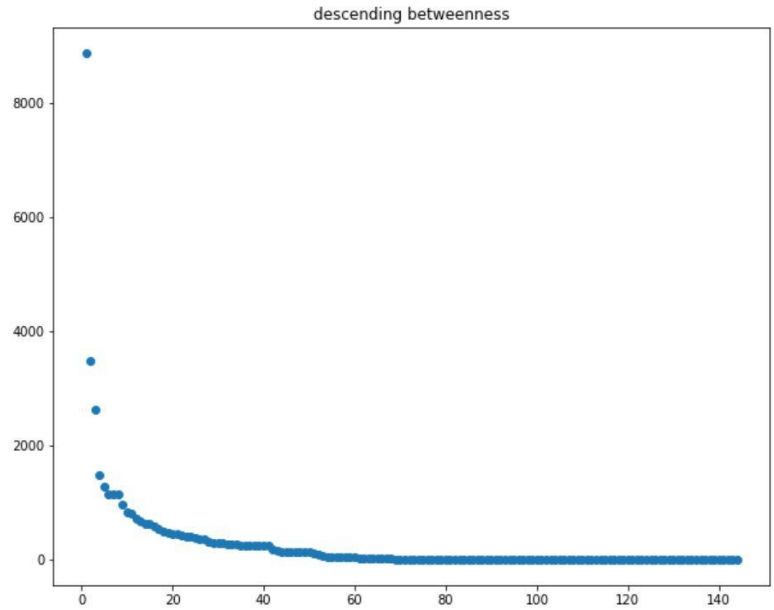

Figure 14: Graph of Betweenness Centrality against UserID

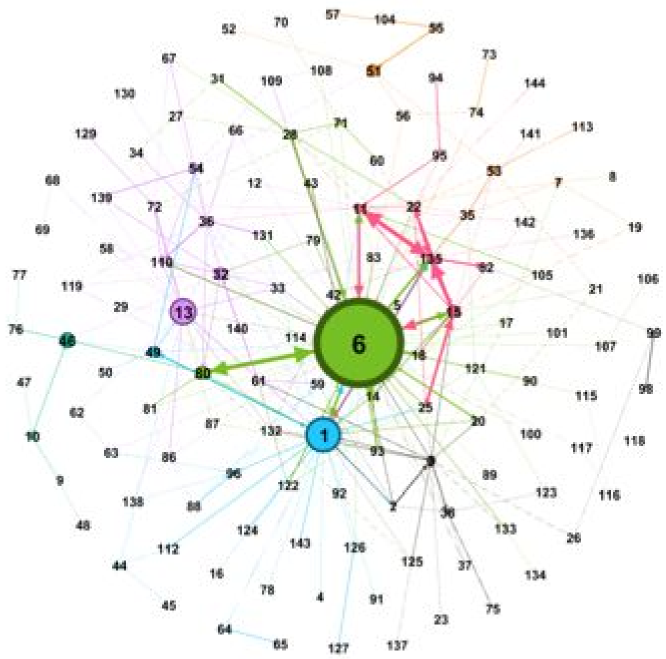

Figure 15: Social Network Graph of HWZ (Node size: Betweenness; Edge weight: Replies)

The graph of betweenness centrality is similar to the graph of degree centrality. This means that the three biggest users (user6, user1 and user13) are the ones who always appear in between the thread of multiple conversations. It can be seen as A posting, B replying and C replying to B rather than to A directly. B has the power to control information being disseminated, meaning that if B does not reply, C might not reply either. As a result, A would not have obtained the information from C.

4.8 Insights

By utilising the HWZ social network graph, we found out the degree, closeness and betweenness centrality values of each top 5 positive and negative users towards SB.

Top 5 HWZ users that contribute **positive** posts about SB:

	User	No. of Posts	Degree	Closeness	Betweenness
1	user011	3	21	0.375758	806.618
2	user082	2	9	0.331551	246.6185
3	user135	2	18	0.381538	577.3358
4	user033	1	9	0.389937	107.0792
5	user083	1	3	0.347339	0

User033 has the fastest spread of positive posts as he is one of the top 10 users based on closeness centrality. User135 has similar closeness centrality as user033, and he has a higher degree centrality and betweenness centrality, which means his spread of positive posts is further. User011 has the highest degree centrality and betweenness centrality among these 5 users, which means his positive posts can reach most people in the HWZ social network. Overall, the spread of positive posts about SB is considered fast and the reach is far.

Top 5 HWZ users that contribute **negative** posts about SB:

	User	No. of Posts	Degree	Closeness	Betweenness
1	user135	3	18	0.381538	577.3358
2	user061	2	11	0.369048	268.5437
3	user033	1	9	0.389937	107.0792
4	user082	1	9	0.331551	246.6185
5	user083	1	3	0.347339	0

The speed of spread of negative posts is about the same as positive posts, but the reach of negative posts is smaller than that of positive posts. The highest betweenness centrality among these 5 users is 577.3358, which is about 28% smaller than the furthest reach for positive posts.

There were overlaps among positive and negative users. User135, user033, user082 and user083 have both posted negative and positive replies/posts regarding SB. Since they had mixed sentiments to SB, this shows that they are unbiased towards SB and the opinions they posted were genuine. Their unbiased attitude will make them appear to be more credible. Thus, gaining the trust of more users on HWZ.

Hence we have identified user135, user033, user082 and user083 as the main drivers of the posts regarding SB.

5. Instagram

5.1 Problem Statement

To increase the likelihood of a more successful marketing campaign in the future.

5.2 Approach

To find out SB's type of followers and how in the future they can seek active followers/influencers to help them post or market their product. The team analysed the behaviour of the SB followers and did a social network graph analysis between SB and its followers (1st layer), as well as its followers' followers (2nd layer).

5.3 Data Information

The data scraped off Instagram are as follows:

- Seamless Bespoke's Followers (1st layer)
  - Username, the total number of likes and comments contribute to SB's posts, first like date, last like date, Is private or Is business account, business category, number of followers, number of following, number of media posts.
  - This data has 14 dimensions.

- Seamless Bespoke's Followers' followers (2nd layer)
  - Username
  - This data has 2 dimensions (ie source and target).

- Likes and comments of the posts by Seamless Bespoke from September 2018 to September 2019.
  - Comments, the username of each comment
  - This data has 2 dimensions (ie user and comment).

The tools used are Helper Tools for Instagram, Instaloader for scraping, LDA for topic modelling and Vader emoji sentiment analysis.

5.4 Data Preprocessing

Since the length of comments is generally short and the number of comments is relatively small, text preprocessing was not performed. The comments were separated into emojis only and text only. This is because the tools used to generate sentiment for text is not suitable for emojis. Since the number of text comments is very small, sentiment analysis was only performed on the emoji comments.

5.5 Topic Modelling on Instagram Comments

Latent Dirichlet Allocation (LDA) was used to perform the topic modelling on 601 comments scraped from Instagram. No concrete topic could be identified after multiple iterations of LDA. The topics identified were too broad (refer to Figure 16) due to large amounts of noise in the data. Hence, it was better to go through the comments manually.

```
Topics found via LDA:

Topic #1:
nice awesome make check look shoe love post want life

Topic #2:
wow love lovely like cool perfect look classy jacket right

Topic #3:
beautiful like look send suit amaze page dm collaborate congrats

Topic #4:
price great good dm picture thanks like follow love available
```

Figure 16: Topics identified by LDA

After manual inspection, it was concluded that the comments can be categorised into "bots", "collaboration" and "user tagging (mention)". Shown below are examples of each comment type.

Figure 17: User Tagging (Mention) Comment

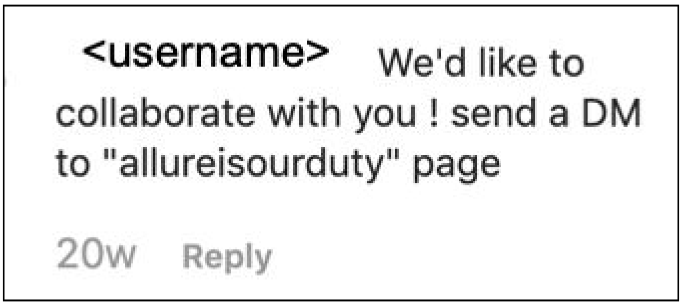

Figure 18: Collaboration Comment

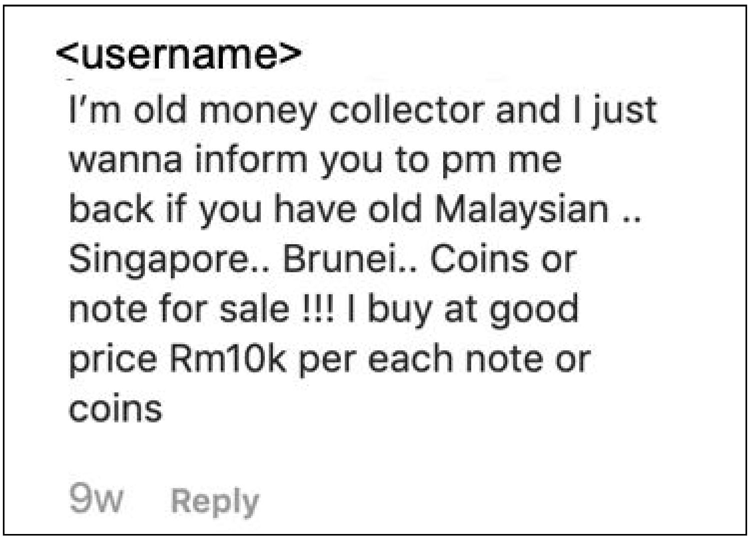

Figure 19: Bot Comment

After excluding comments with the above-mentioned categories, the word "Congratulations" kept appearing at a high frequency. There was uncertainty as to the contexts in which that word would be used. Hence, an investigation into the comments which contained the word "Congratulations" was launched. The conclusion was later drawn that those were posts regarding SB's new store location and the wedding of one of the owners' friends, of which they tailored the outfit.

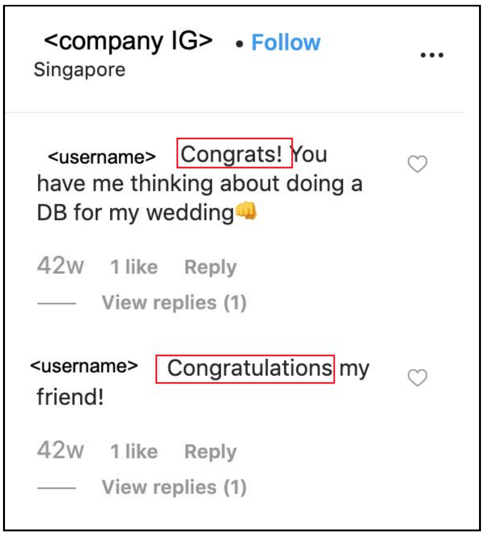

Figure 20: The word "Congratulations" in a Wedding Post

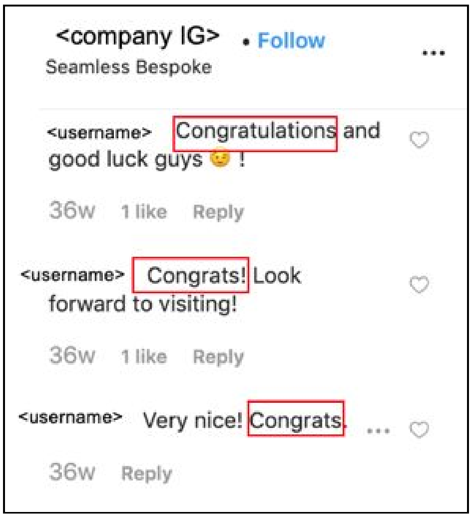

Figure 21: The word "Congratulations" found in a post showing the new store

It was finally decided that the comments with "congratulations" under the topic of "greetings".

The next issue was that many comments were just pure emojis and LDA could not detect them. Hence, Vader was used to perform a sentiment analysis on the comments with emojis. All of emojis are detected as positive hence another topic of "positive emojis" was created.

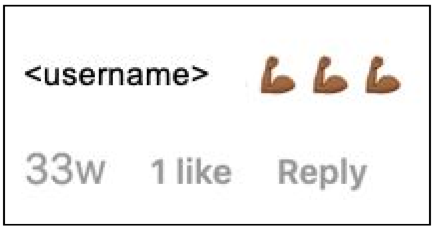

Figure 22: Pure emoji comments

After labelling some of the comments as "bots", "collaboration", "user tagging", "greetings" and "positive emojis", LDA was ran again on the remaining unlabelled comments. In the end, LDA identified 2 topics as shown below.

Topic #0	Great, beautiful, lovely, wow, perfect, cool, awesome, order, suit, detail, linen, purchase, stylish, dm_price, go, still, yanko, nice_work, classy, pm_price
Topic #1	Nice, love, like, good, amazing, photo, shoes, fabric, could, beautiful, great, picture, right, best, top, amaze, collar, wow, gorgeous

Topic #0 is geared towards "query price/product" because the words related to querying appears frequently ("order", "purchase", "dm_price", "pm_price", "online", "cost" and "available_online").

Example of a comment that is labelled as Topic #0: The material in photo 3 looks **beautiful**, may I know what material that is? How much would a 2 pc **suit** in that material **cost**?

Another example of a comment that is label as Topic #0: Amazing, **DM** the **price** please* (dm_price = DM the price as the team removed stop words and bigram it)

On the other hand, all the words in Topic #1 are generally positive. Hence, this topic was labelled "praise".

Example of a comment that is label as Topic #1: **Wow love** these

Example of a comment that is label as Topic #1: **Nice shoes**

Upon secondary inspection, a few comments still did not fall under any category of the identified topics. It was decided that they will be labelled under "others" because those comments have keywords that appeared less frequently. LDA was probably not able to detect them due to their low frequency of appearance.

Weight was given to the commenting. The logic behind the assignment of weights is because comments have a higher impact as compared to "likes".
A user liking a post could simply due to:

a. user habits of liking every post while scrolling
b. liking it without any specific reasons

When a user comments on SB's posts, it shows that they are interested in SB. However, not all comments are meaningful. Therefore, the comments were further weighted based on category.

The weight assigned to each topic is shown below.

Topics	**Weights**
User tagging (Mention)	5
Query price/products	4
Praise	3
Greetings	2
Others	2
Positive Emoji	1
Collaboration	0
Bots	0

The highest weight (5) was assigned to “user tagging (mention)” because if a user tagged another user under SB’s post, the user acts as a connector that helps to spread the company product to others. This increases brand awareness.

The second highest weight (4) was assigned to “query price/product” because when a user asks for the price it means that the user is interested in the products.

A weight of 3 was assigned to “praise” because it takes effort for the user to express his/her positive emotions on SB’s products/posts.

Following which, “greetings” and “others” was given a weight of 2 and “positive emojis” a weight of 1.

Finally, 0 is given to “bots” and “collaboration” as those comments do not contribute to engagement or awareness. These comments are not helpful to the company’s social image as the appearance of such comments make it seems like the company’s Instagram has fake followers.

5.6 Insights

5.6.1 Instagram comments

Out of 601 comments, 256 comments came from SB’s followers, while the rest came from non-followers. Based on figure 23 shown below, most of SB’s followers posted comments about “praises” to the product/content and “query” about the products/cost.

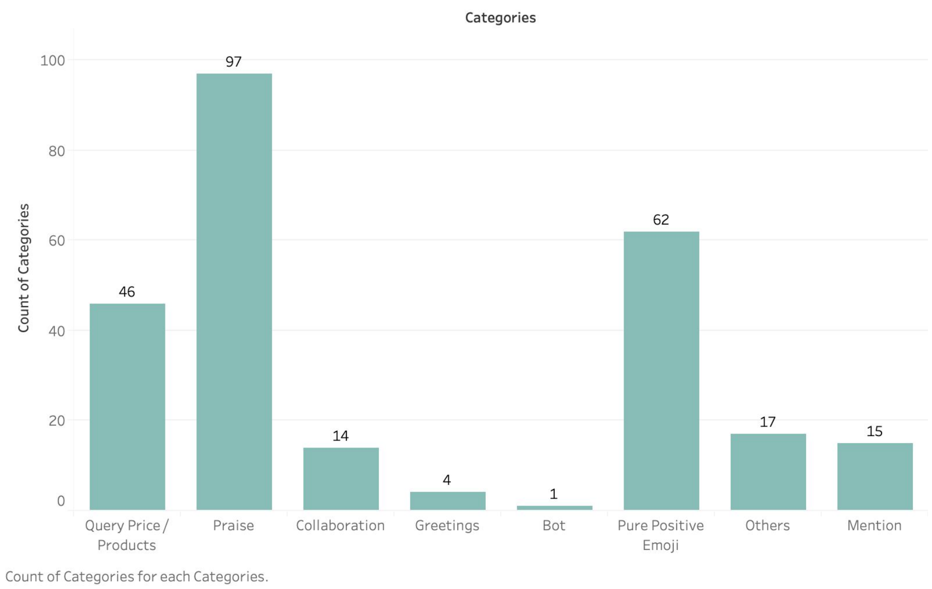

Figure 23: Follower Comment Count by Category

Referring to figure 24, most of the non-followers posted comments about “praises” to the product/content and “positive emojis”.

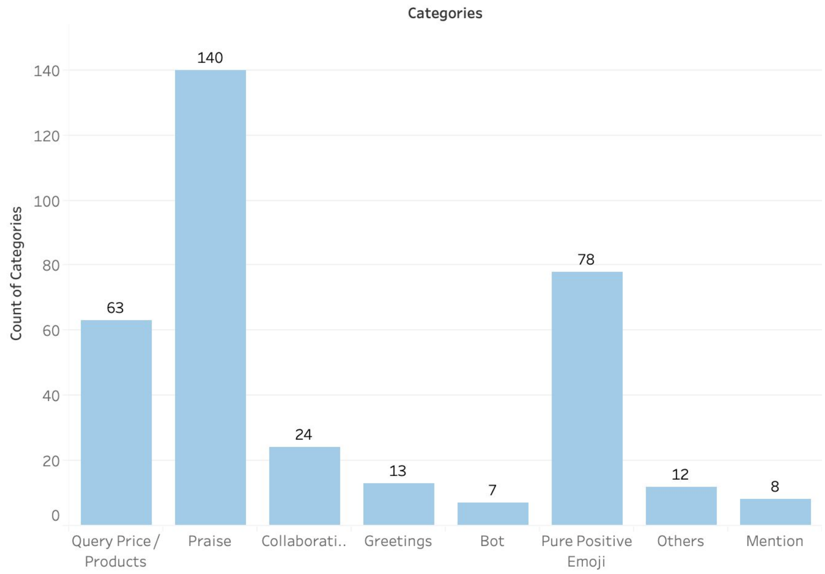

Figure 24: Non-follower Comment Count by Category

Overall, the comments are generally positive. Improvements can be made by SB in the “query” on price and products category. More will be elaborated under the recommendations section where comparisons will be made between HardwareZone Forum comments and Instagram comments.

5.6.2 Types of Followers

A total number of 13,313 users liked or/and commented on SB’s Instagram posts. However, only 3162 of them were from SB’s 9862 followers.

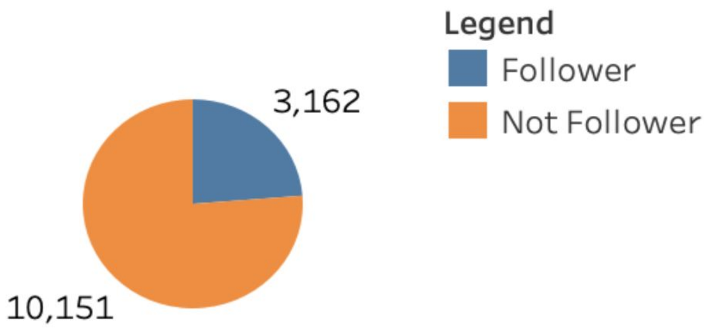

Figure 25: Number of Followers vs Non-Follower Liked / Commented

Focusing on the 3162 followers, a metric was created to determine whether they are active user based on:

1. Follower Engagement = number of like + (sum weight of each comment * number of comments)

2. today()* - last like date
   * the date is 3 October 2019 as it was the day of analysis

Metrics	Range	Reason
3	Less than or equal 30 days	Very active
2	31 to 90 days	Active
1	91 to 180 days	Irregularly active
0	181 and above	Inactive

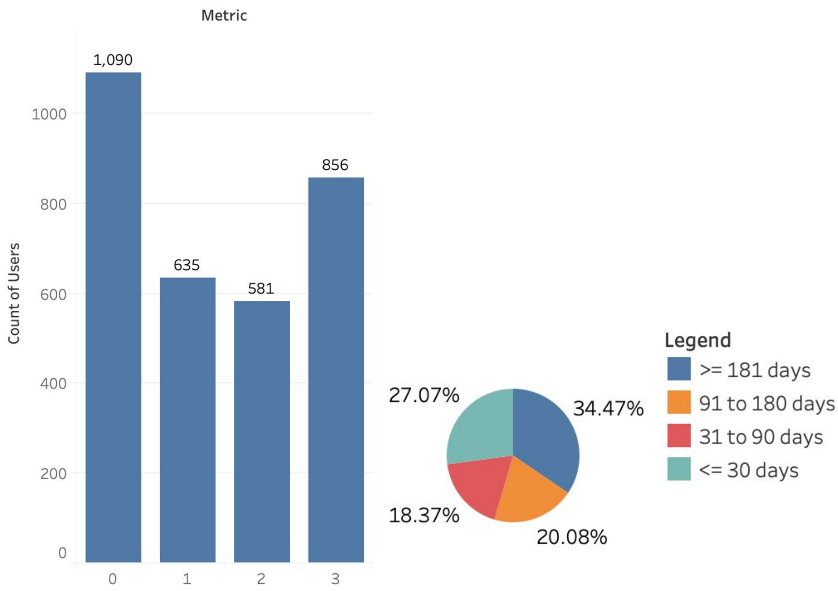

Figure 26: Count of Users in Each Metrics

Based on figure 26, the team found that most of the followers had stopped liking the company post for more than 180 days. The team decided to dive into it and found out that 53% of the 1090 followers (metrics 0) are no longer active on Instagram as their last media post date on their own account could be as old as 7th July 2014.

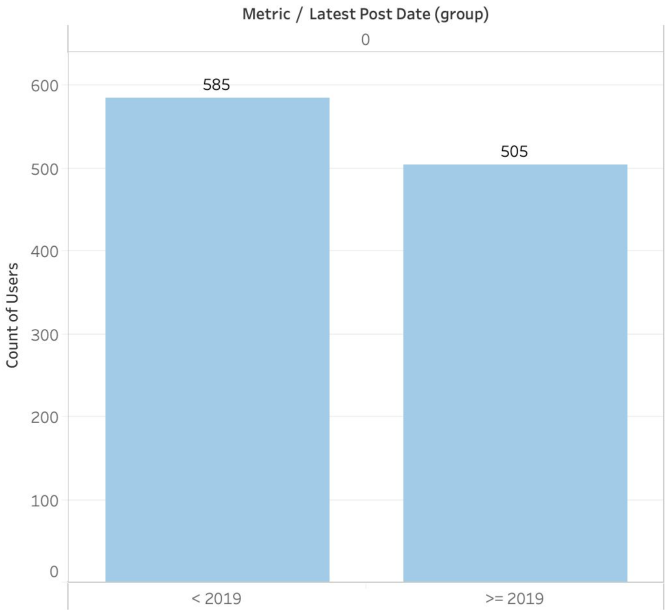

Figure 27: Count of users (metrics 0) with the latest post date before and during 2019

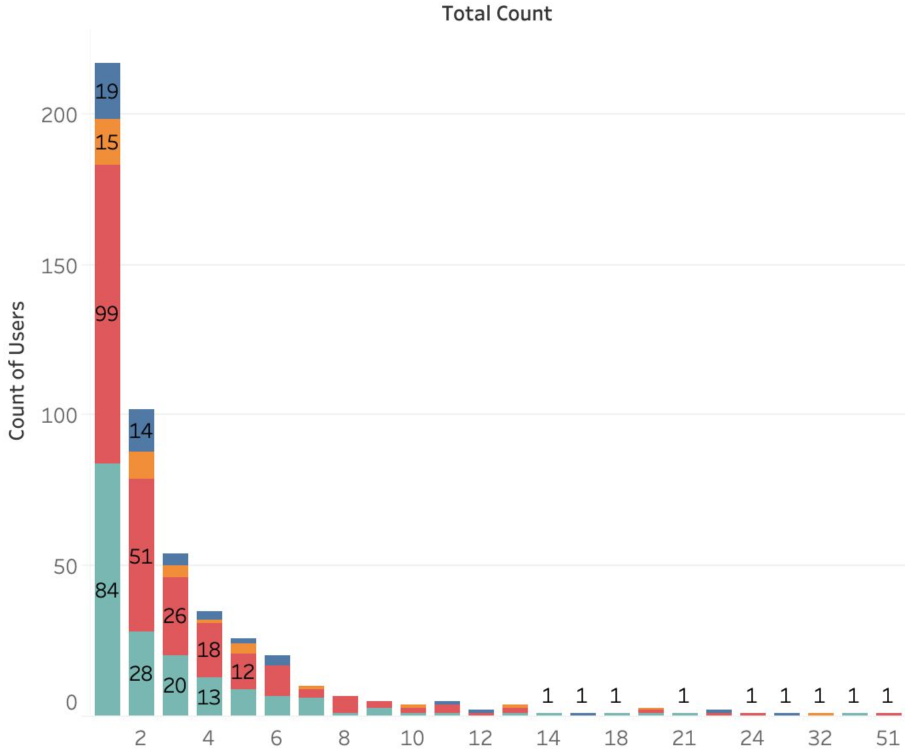

Figure 28: Count of users (metrics 0) by number of SB's posts liked in 2019

The colours denote the media count of the user: Orange - 4 digits, Green - 3 digits, blue - 2 digits, red - 1 digit.

Based on figure 27, 47% of the users with metrics 0 are still active in 2019, but most users only liked SB's posts once and only 1 user liked 51 SB's posts (refer to figure 28). This could be due to a lack of stickiness of SB's posts which in turn made it difficult for them to retain their followers. To support this point, the followers' profiles were looked into. Most of them have 100-1000 media posts. This means that these users (metrics 0) are active on their own social media circle, but they are not actively responding to the posts by SB. The lack of active followers could be the main reason why SB’s previous social media marketing campaign failed.

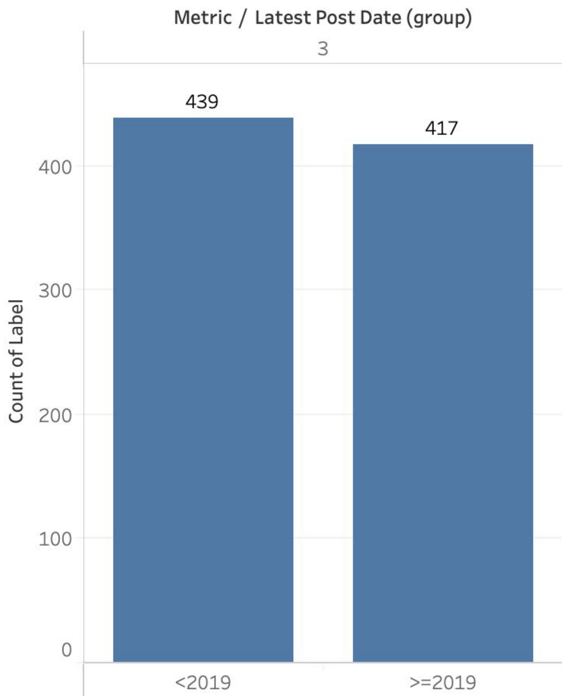

Figure 29: Count of users (metrics 3) with latest post date before and during 2019

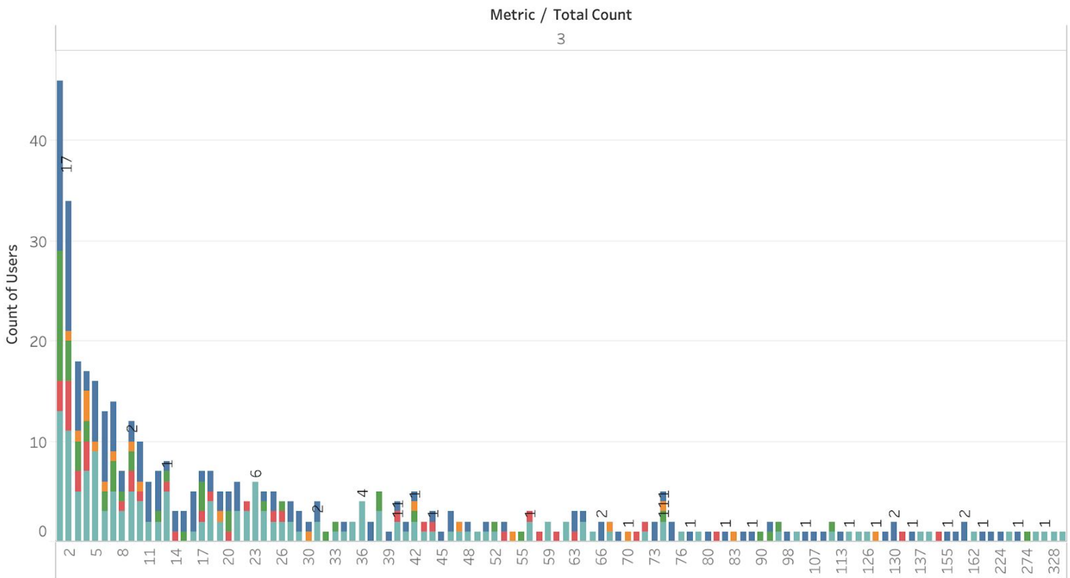

Figure 30: Count of users (metrics 3) by number of SB’s posts liked before 2019

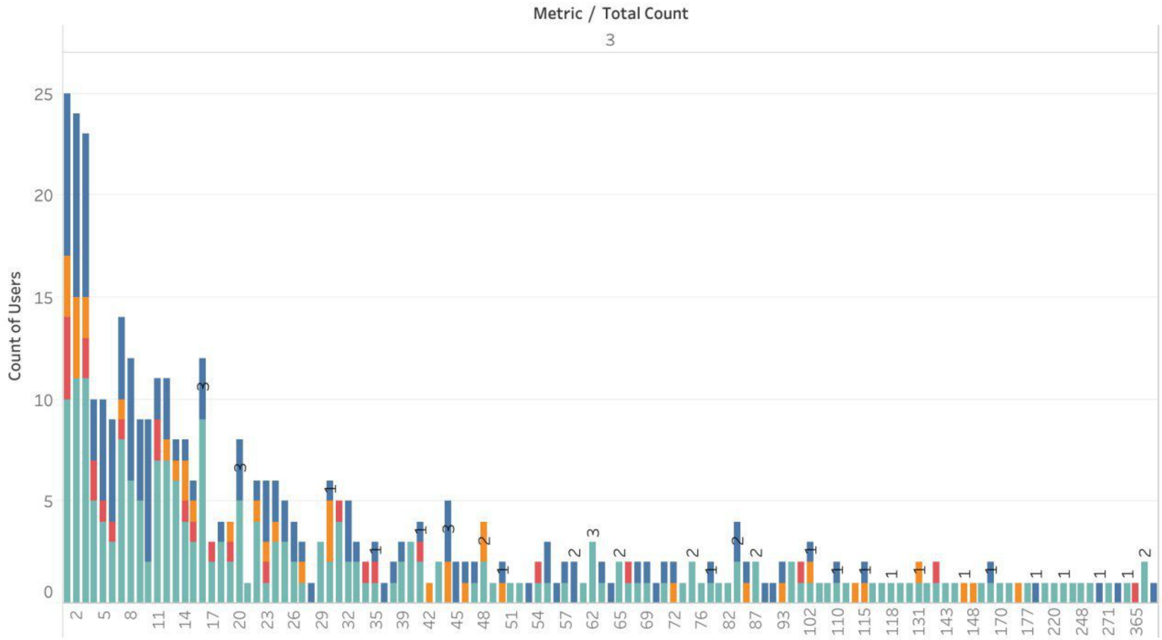

Figure 31: Count of users (metrics 3) by number of SB's posts liked in 2019

The colours denote the media count of the user: Orange - 4 digits, Green - 3 digits, blue - 2 digits, red - 1 digit

Based on graph 29, 30 and 31, users with metrics 3 are relatively more active towards SB posts as compared to users with metrics 0.

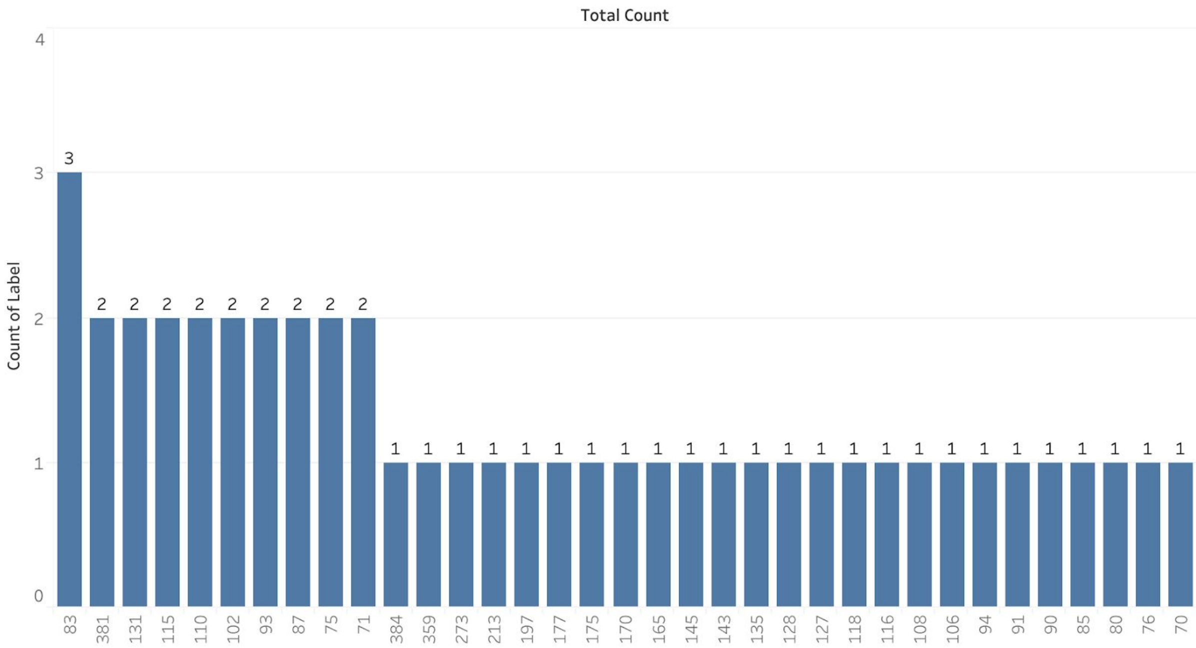

Figure 32: Count of users (Metric 3) who liked and/or comment above 70 of SB's posts

Users who liked/commented SB posts for more than 70 times were also identified (refer to figure 32). These users will be taken into consideration when selecting users to perform social network analysis, which will be explained in the next part.

5.5.3 Instagram Social Network Analysis

The total number of users with metrics 3 is 46 but 2 users were dropped because their followers-following ratio is too big (Instagram would potentially be able to detect and ban the team for scraping data off their platform).

The group identified the top 44 users with metrics = 3, latest post date = 2019, number of SB posts liked/commented >= 70, Private = False, Business Account = False and media post count at 2 digits and above.

The users with the above-mentioned characteristics would mean that they had been following SB's posts closely for more than 2 months and they are actively engaged with the posts by SB. They are also considered to be active on social media since they have a media post count of double digits and above. Business accounts are excluded due to a conflict of business interests. Private accounts are also excluded because their betweenness and closeness are limited only to people who followed them.

5.5.3.1 First Layer Analysis (between SB and its followers)

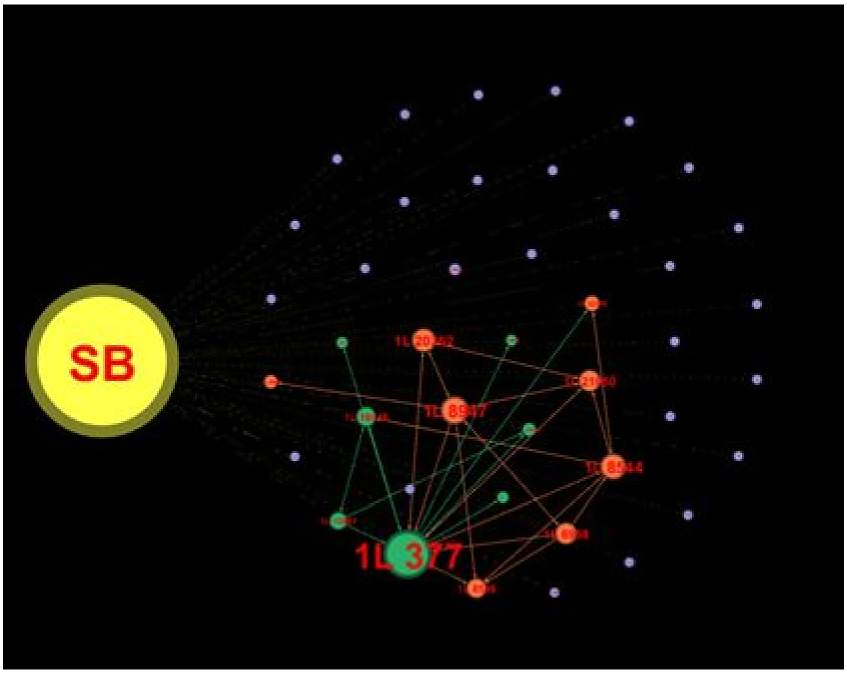

Figure 33: Degree Centrality

Degree Centrality

Node 377 (user 377) has the highest degree centrality. Only roughly ⅔ of these 43 nodes connected to SB, the rest of the nodes are mostly connected to node 377. This indicates that for people who follow SB, they also follow node 377. This means that node 377 has some form of influence on SB's followers.

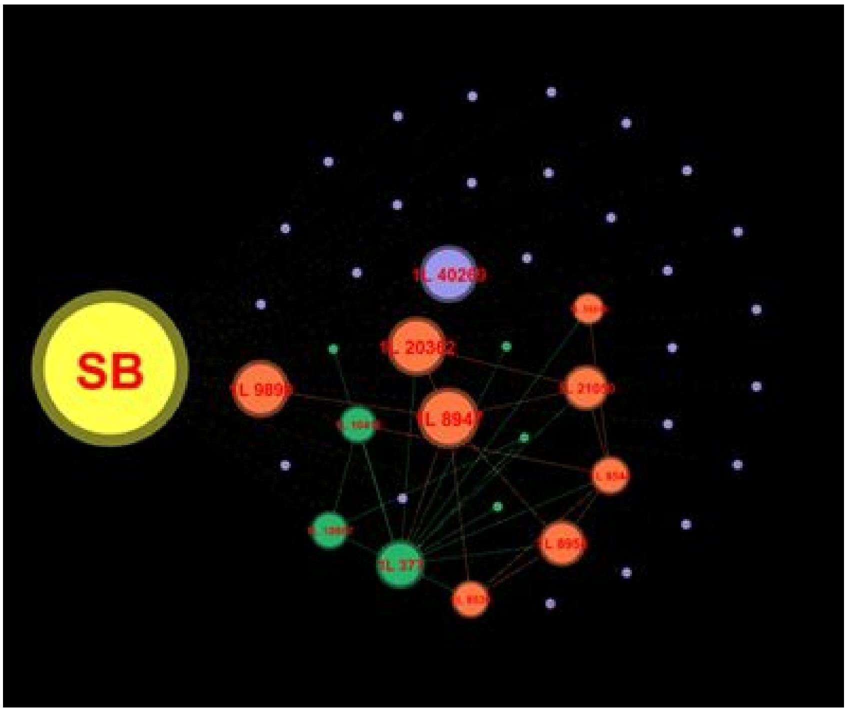

Figure 34: Closeness Centrality

Closeness Centrality

Nodes 9899, 8947, 20362 and 40269 have high closeness centrality. These are the users who can spread information about SB very fast.

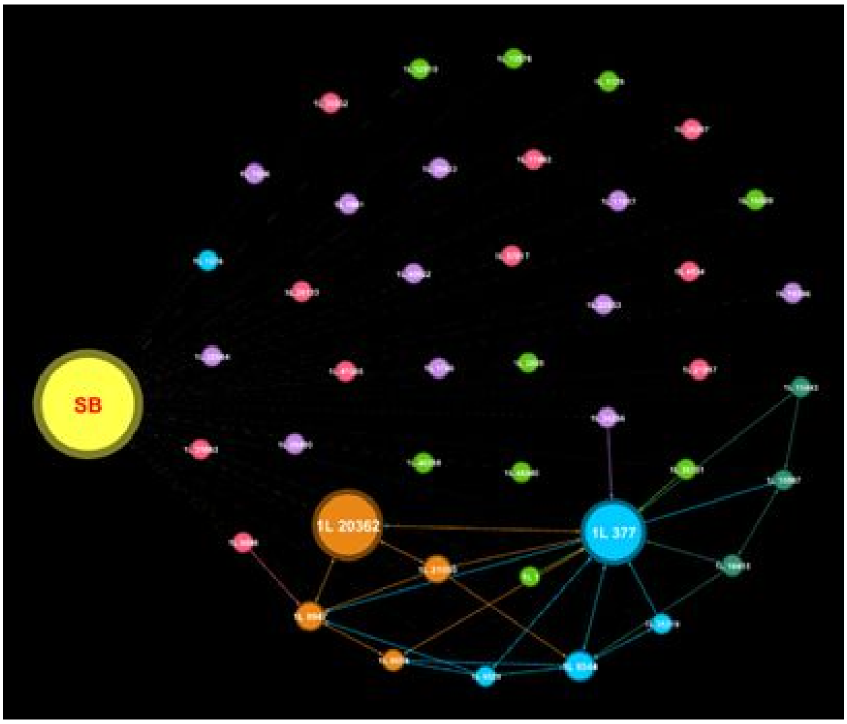

Figure 35: Betweenness Centrality

Betweenness Centrality

When it comes to betweenness centrality, an interesting fact was discovered. Both nodes 377 and 20362 become the largest node (excluding SB) in the network. It is worth to point out that node 377 has the highest degree centrality while node 20362 is one of the nodes with high closeness centrality.

Based on the 1st layer analysis between SB and its followers, it was found that node 377 and node 20362 could be the potential influencers that are helpful to SB.

Looking into user 377's Instagram profile (username: dk.gtm), it was found that he is a menswear enthusiast and shoe reviewer based in Hong Kong. He has 6235 followers and his average posts like is 266.

Looking into user 20362's Instagram profile (username: m_lyc), it was found that he is also a shoe lover. He posts his purchases on Instagram such as SB's shoes and chino pants. Although he only has 331 followers, he still has considerable influence within the community because of his high closeness centrality and betweenness centrality.

Therefore, user 377 and 20362 can be approached as potential influencers of SB for future marketing.

5.5.3.2 Second Layer Analysis (between SB and its followers)

	Id	Label	Degree	indegree	outdegree	closnesscentrality	betweenesscentrality
64	seamlessbespoke	SB	46	4	42	0.507332	26501.213108
305	yeossal	2020	33	6	27	0.490547	18085.940049

Figure 36: Second layer Yeossal

Among the second layer, Yeossal has the highest closeness, betweenness and degree centrality. This shows that users who follow SB also follows Yeossal. From here, tight and close competition can be seen among these 2 companies. Users see posts from both SB and Yeossal. This will naturally lead to comparisons between them. Moreover, both of their closeness centrality are very similar. It would mean that both of them can spread information just as fast. SB's strongest competition would be Yeossal. This knowledge is crucial when making business decisions.

6. Recommendations based on HWZ and Instagram Insights

While doing the analysis, the team noticed that users post their suit/clothes on HWZ and curious enthusiasts will reply to them. Based on this model, it is may be a good opportunity for them to encourage such behaviour on Instagram. SB can encourage consumers to post their SB bought outfits on Instagram and either hashtag or tag them. This could bring about greater brand awareness as well as potentially bring their support base to Instagram. This is supported by the fact that many potential customers tend to look for reviews and comparisons between them and their competitors.

Also, the two influencers identified could act as brand ambassadors. Both of them have high betweenness which allows them to spread the word of SB far and wide. User 377 has a high degree centrality and User 20362 has a high closeness centrality on top of their betweenness. Multiple groups of like-minded people about tailor will be better able to know of their products and campaigns. It is believed that this is a better marketing strategy than paid advertisements as a portion their direct followers are in the first place inactive accounts. Also, it is already seen that this social network responds better to fellow user reviews.

Lastly, for their when posting, it would be nice if SB could include the price of items. That seems to be the main concern of potential customers. Also, it would be good if SB does not hashtag themselves too often when posting. The reason for this is that when potential customers clicks on their hashtag, they will be able to see posts by tailoring enthusiasts rather than all of it just by SB themselves. The network tends to believe other users rather than the tailor themselves in this community.

7. Facebook

7.1 Problem Statement

To increase the likelihood of a more successful marketing campaign in the future.

7.2 Approach

The goal is to find out how well is their current social reach and engagement. Based on the past user engagement level, we will identify the best timing for posting and the content type that will appeal to more users. This information will help to guide SB for future content creation for the next marketing campaign.

7.3 Data Information

The team obtained the data of Seamless Bespoke's Facebook posts directly from them. It is a large CSV file that contains all the post captions, likes, comments, shares, hides, time of post, the reach of post etc. Basically, it was a file with 49 dimensions with all the posts Seamless Bespoke made within the past year. With this information in hand, there was no need to do any scraping on Facebook. Moreover, scraping on facebook will result in a ban almost instantaneously with their high levels of security.

7.4 Data Preprocessing

Much cleaning had to be done so as to understand all the column and remove those that will not provide much analysis. It was understood that the client did not make use of this data extracted as there were too many dimensions and they did not know which ones to zoom into. The number of dimensions was reduced to 34 after introducing our engineered features of gender which the post is targeted towards, the colour of the item(s) that was posted, type of post and what the post captured.

7.5 Analysis and Insights

7.5.1 User Engagement (Positive)

The likes, comments and shares have been aggregated to engagement with weights 1, 3 and 5 respectively. This is because there are not many comments and shares of their post and doing it separately with sparse data would not be accurate.

7.5.1.1 User Engagement by day of the week

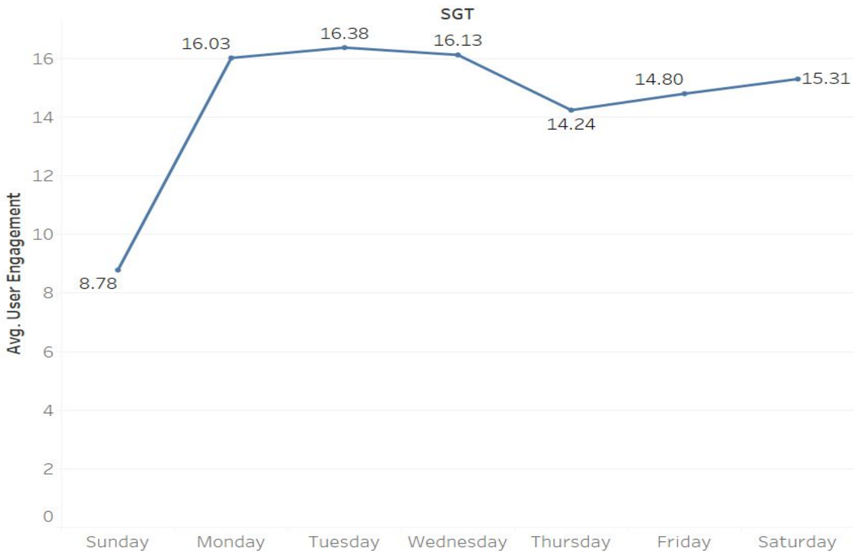

Figure 37: Average User Engagement on different days of the week

It can be observed that Sundays yield the lowest levels of user engagement as compared to Monday to Saturday. It would be recommended to post more on Mondays-Wednesdays to yield even higher user engagement while users are already active on Facebook during those periods.

7.5.1.2 Overall User Engagement by Time of the Day

12am - 6am : Sleeping Hours
7am - 12pm : Morning
1pm - 6pm : Afternoon
7pm - 12am : Evening

a. Time of the Day overview

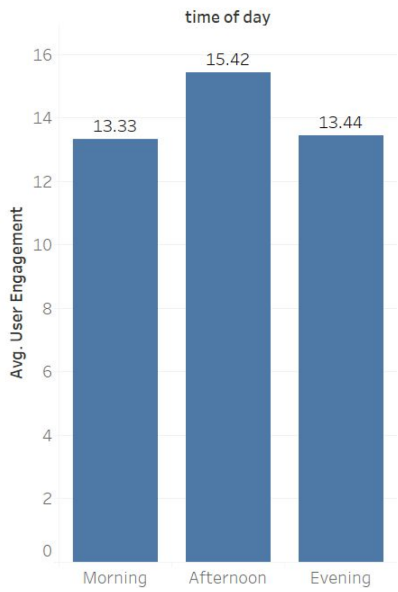

Figure 38: Average User Engagement in Different Time of the Day

Based on a general overview of all types of posts, posts posted in the Afternoon yield the highest engagement while posts posted either in the Morning or Evening perform about the same.

b.User Engagement of Female and Male products

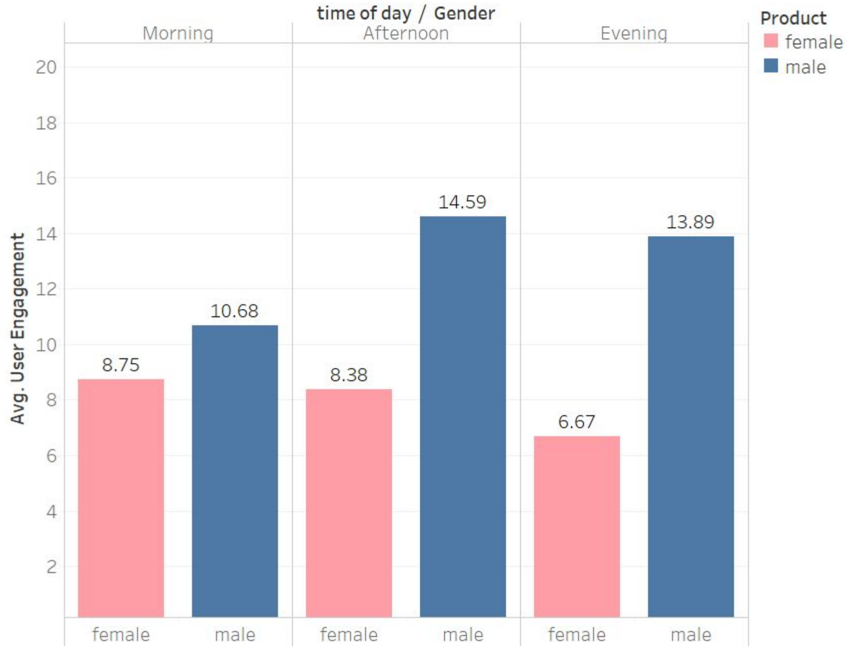

Figure 39: Average User Engagement of Female and Male products on different time of the day

When the post types are further broken down to products by gender, posting Male products in the afternoon yield the highest engagement, followed by evening then morning. Female products are most well engaged during the morning, and almost the same in the afternoon, and worst in the evening.

It is recommended that SB posts more male products during the afternoon and evening and post more female products in the morning and afternoon so that it would produce greater engagement in total.

c. User Engagement of Product Colours

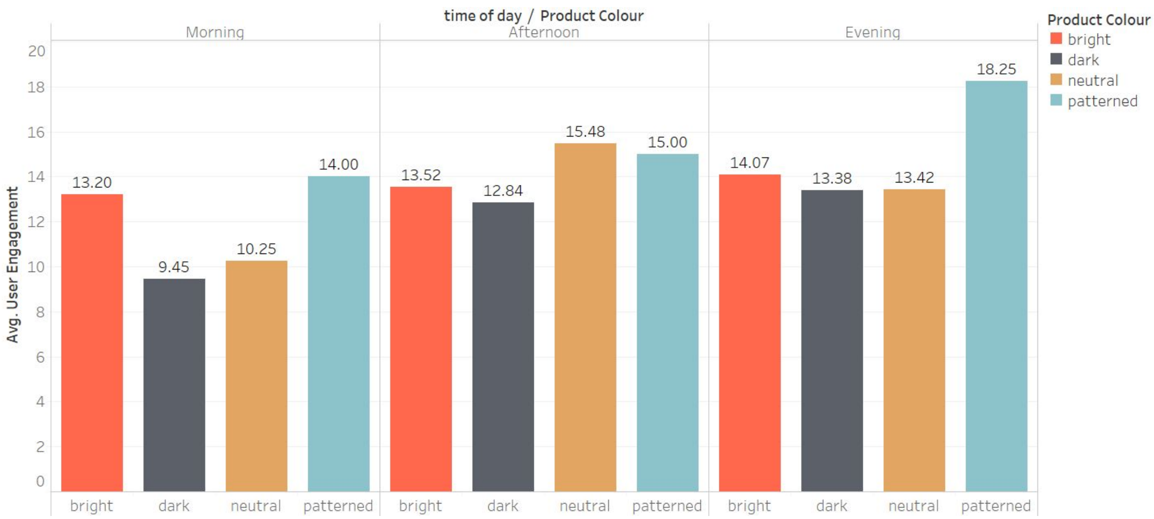

Figure 40: Average User Engagement by product colours on different time of the day

When further broken down to the product colours and grouped by the time of the day, there is a slight dip in the user engagement during the morning, with significantly lower engagement of dark coloured and neutral coloured products in the morning. On the flip side, there is significantly higher engagement with patterned products in the evening.

SB could post more bright coloured or patterned products in the morning as it seems that customers prefer that, and post more patterned products in the evening as well.

d. User Engagement of Miscellaneous posts by time of the day

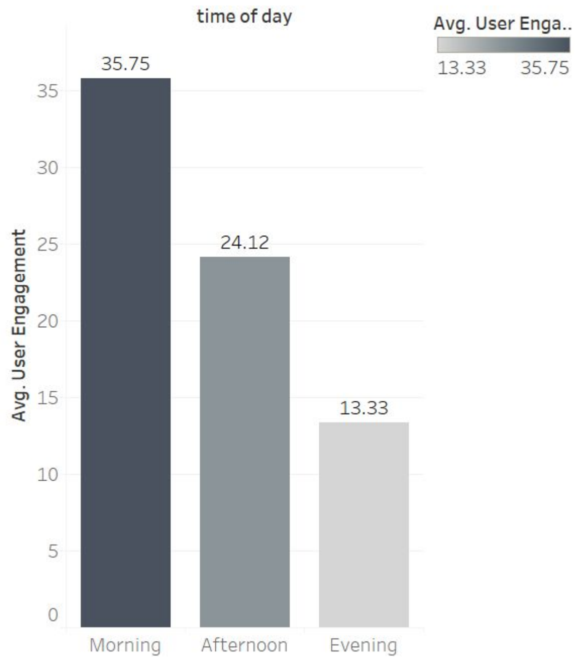

Figure 41: Average User Engagement of miscellaneous on different time of the day

Miscellaneous posts include: Events, Promotions, Announcements

When focusing only on Events, Promotions, Announcements, SB should post more of these non-product related posts in the morning and minimise these posts in the evening as the user engagement decreases significantly throughout the day.

7.5.1.3 User Engagement grouped by Female, Male, and Gender-Neutral products

Gender Neutral Products include Fabric samples and Miscellaneous which is are events, promotions and announcements

a. Overall User Engagement of Female, Male and Gender-Neutral products

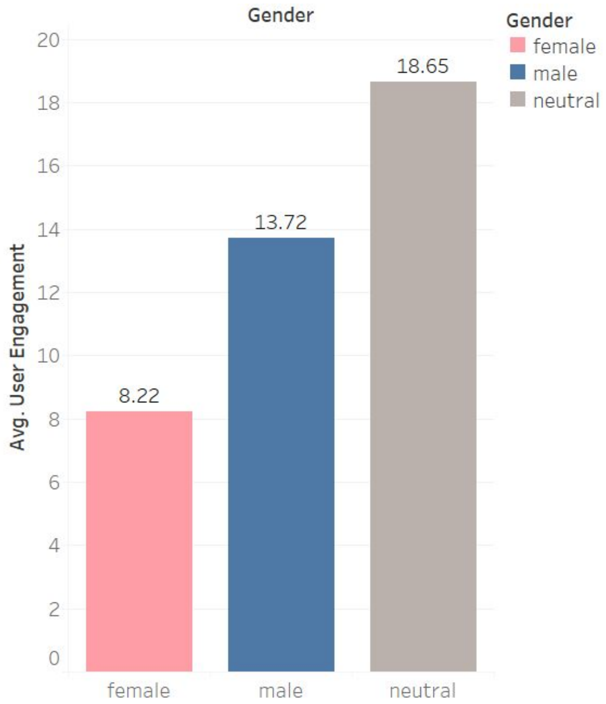

Figure 42: Average User Engagement of female, male and gender-neutral products

When segmented into the gender of the products featured on the posts, it is evident that female products featured gather the least engagement, followed by male products, and gender-neutral products gather the most engagement.

b. Product Colours

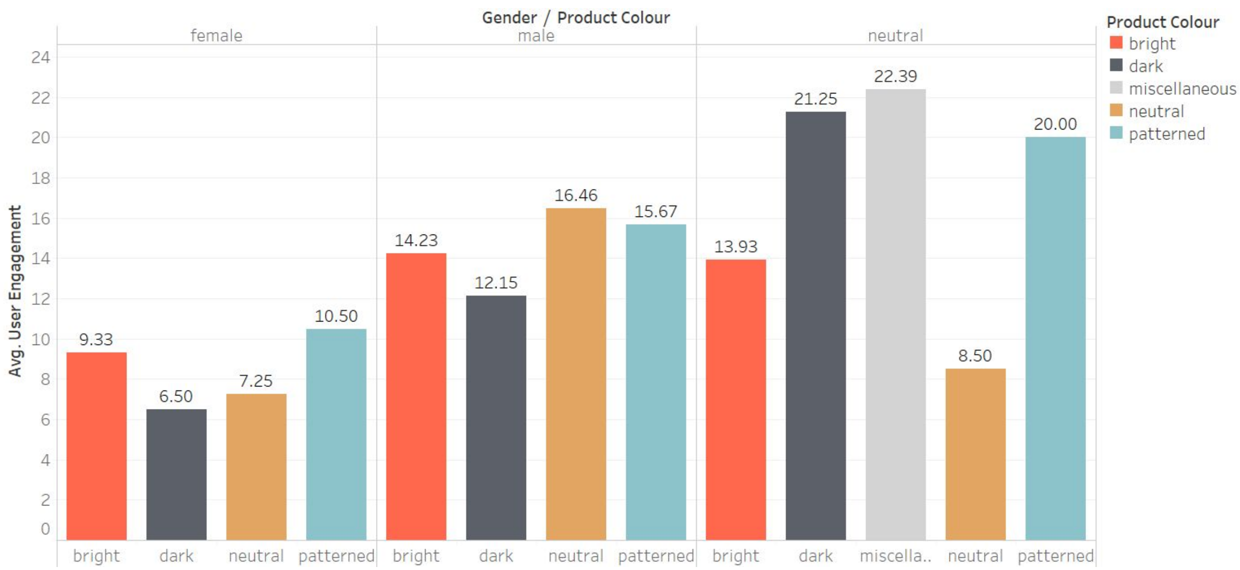

Figure 43: Average User Engagement of different product colours by female, male and gender-neutral products

Under the gender-neutral column, miscellaneous refers to all events, promotions and announcements while all the other colour categories are for the fabric samples.

When further broken down into the colour of the products grouped by gender, for the Female products, there is most user engagement among the bright coloured and patterned products. Among the Male products, the neutral coloured products yield the highest engagement, and among the gender-neutral products, aside from the promotions and events, the dark coloured and patterned fabric samples are the most popular.

Based on this analysis, we can get a rough sense of the overall preference of the different gender groups.

c. Product Type

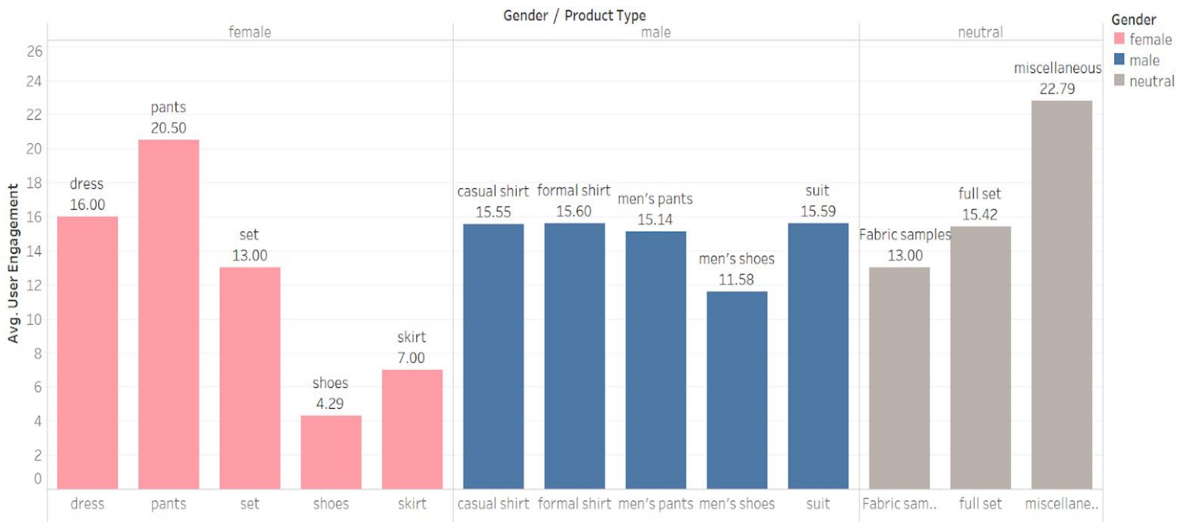

Figure 44: Average User Engagement of different product type by female, male and gender-neutral products

From this graph, we can see that among the Female products, there is a strong preference for pants and little interest in shoes. When we look at the Male products, most of the responses towards the different products are homogenous, but their shoes have significantly less user engagement.

7.5.2 Photo Views (Positive):

The patterns of user interactions with the photos of their post will be analysed below (ie. the number of people who click the photos on the posts to view). This is important as it shows the number of users that are interested in the items that they post.

7.5.2.1 Photo views by day of the week

Figure 45: Average Photo views by day of the week

It is clearly shown here that posts on Saturdays garner the amount of interest in terms of viewing the products Seamless Bespoke has to offer. This is closely followed by Monday. This leads to the insight that while posts on Saturdays may not have gained the most number of interactions. It certainly has one of the greatest reach. Posting photos of products on these 2 days may be a good marketing strategy to reach interested consumers.

7.5.2.2 Photo views by the time of the day

12am - 6am : Sleeping Hours
7am -12pm : Morning
1pm - 6pm : Afternoon
7pm -12am : Evening

a. Time of the Day overview

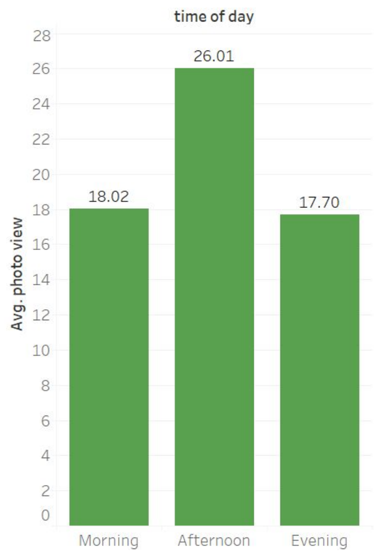

Figure 46: Average Photo views by different time of the day

Posting in the afternoons garnered the most interest in their photos. On average, it can be seen that afternoon posts generates about 10 more views as compared to mornings and evening posts.

b. Photo views of Female and Male products

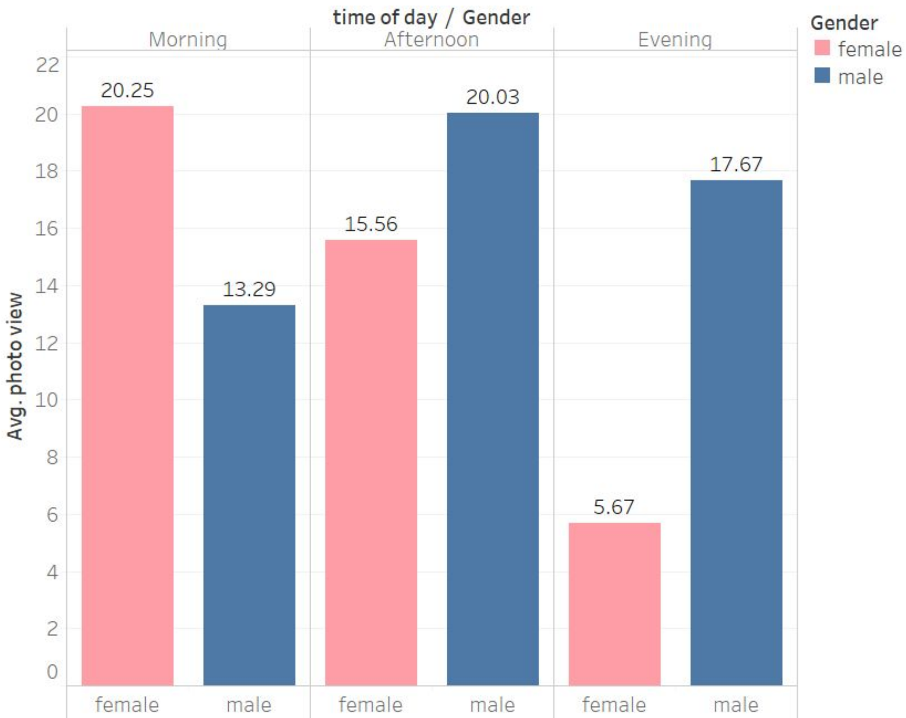

Figure 47: Average Photo views of female and male products by different time of the day

However, after diving into further analysis, we can see that females actually prefer to view their products in the morning compared to the afternoon and the evening. This led us to the insight that males prefer to do their shopping in the afternoons and evenings. The reason behind the previous graph reflecting such high views in the afternoon is because most of the posts are male apparels.

c. Photo views of Product Colours

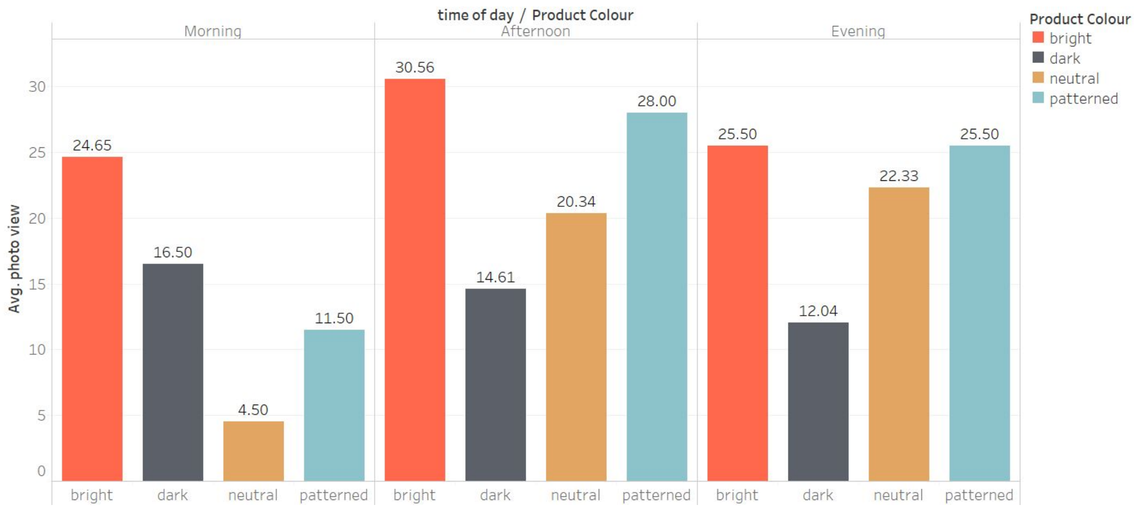

Figure 48: Average Photo views of different product colours by different time of the day

It is quite clear that apparels of pattern and of neural colour tones are the least of interest early in the morning. Users are generally more drawn towards bright posts. Morning seems to also be the best time to post darker colours as compared to the rest of the day.

d. Photo Views of Miscellaneous posts by time of the day

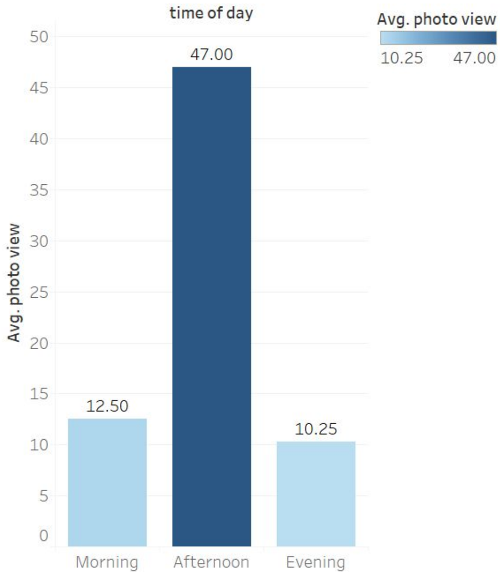

Figure 49: Average Photo views of miscellaneous posts by different time of the day

Miscellaneous posts include: Events, Promotions, Announcements
The most interest generated for announcements and events is very clear in the afternoon.

7.5.2.3 Photo views grouped by Female, Male, and Gender-Neutral products

Gender Neutral Products include Fabric samples and Miscellaneous which is are events, promotions and announcements

a. Photo views of Female, Male, and Gender-Neutral products

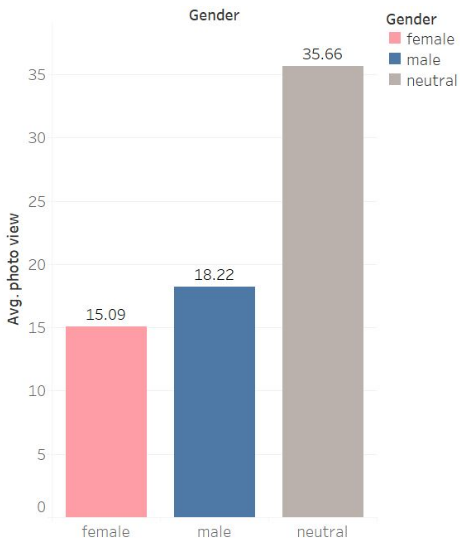

Figure 50: Average Photo views of Female, Male, and Gender-Neutral products

The above graph reinforces what is already known. There are more male consumers in this market and events as well as sales tend to draw the most interest.

b. Photo views of Product Colours

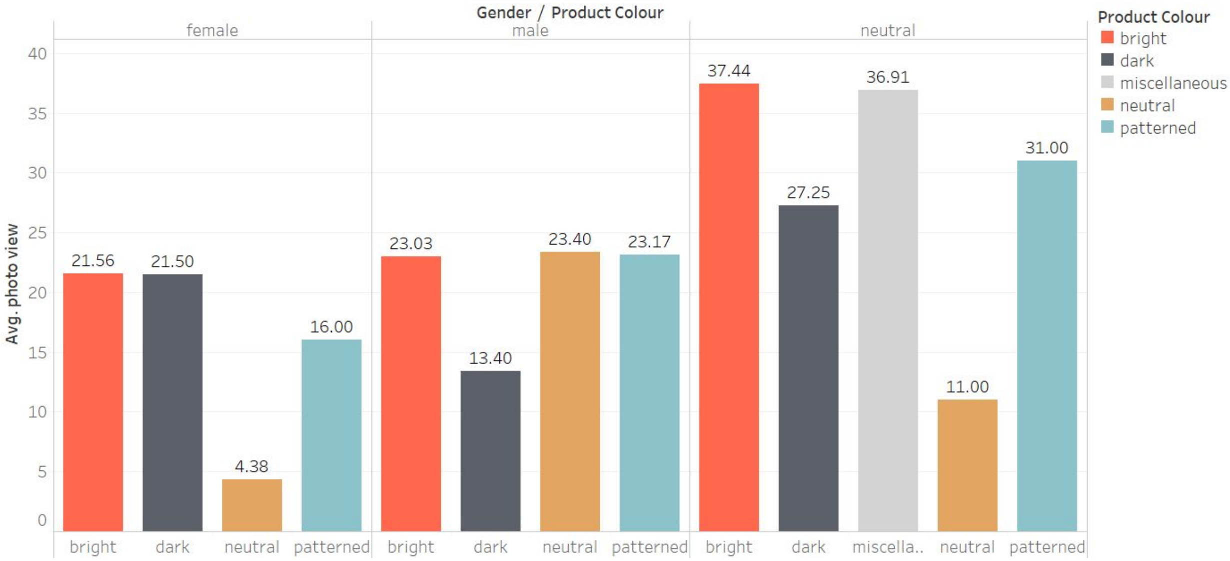

Figure 51: Average Photo views of Product Colours

Under the gender-neutral column, miscellaneous refers to all events, promotions and announcements while all the other colour categories are for the fabric samples.

When broken down for further analysis, it was quite surprising that bright fabrics actually interest the consumers. Gender-neutral items tend to attract the most number of consumers.

c. Photo views of Product Type

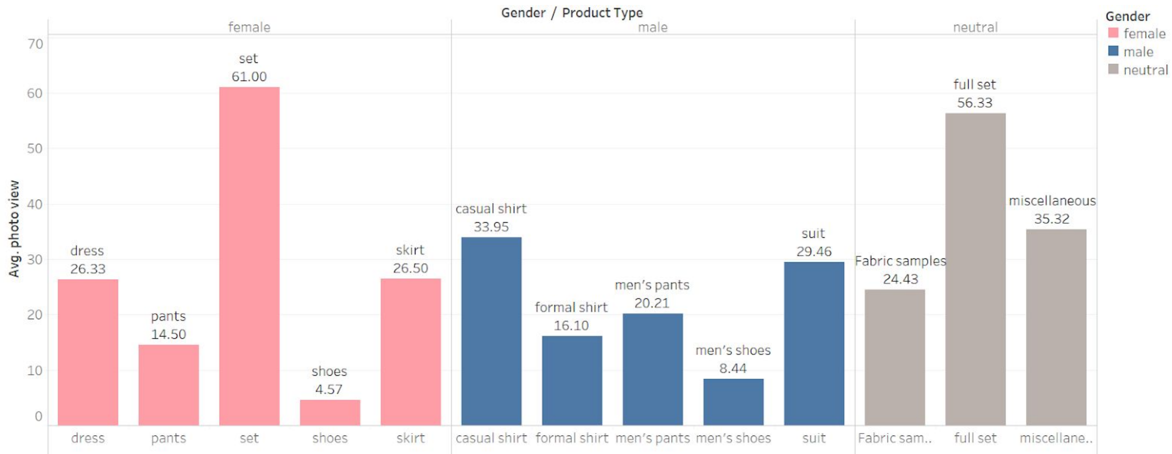

Figure 52: Average Photo views of Product Type

There is a clear insight that both males and females will be more interested to see their items all in full sets. There have not been significant full sets post of males but from the graphs seen in the gender-neutral column, where male and female full sets are posted together, the most amount of interest is generated.

7.5.3 All Hides (Negative)

This is an analysis of the number of people who hide the specific post or all posts from SB when they see a specific post to trigger the action.

7.5.3.1 Total Hides by the time of the day

12am - 6am : Sleeping Hours
7am - 12pm : Morning
1pm - 6pm : Afternoon
7pm - 12am : Evening

a. Total Hides of Product Colours

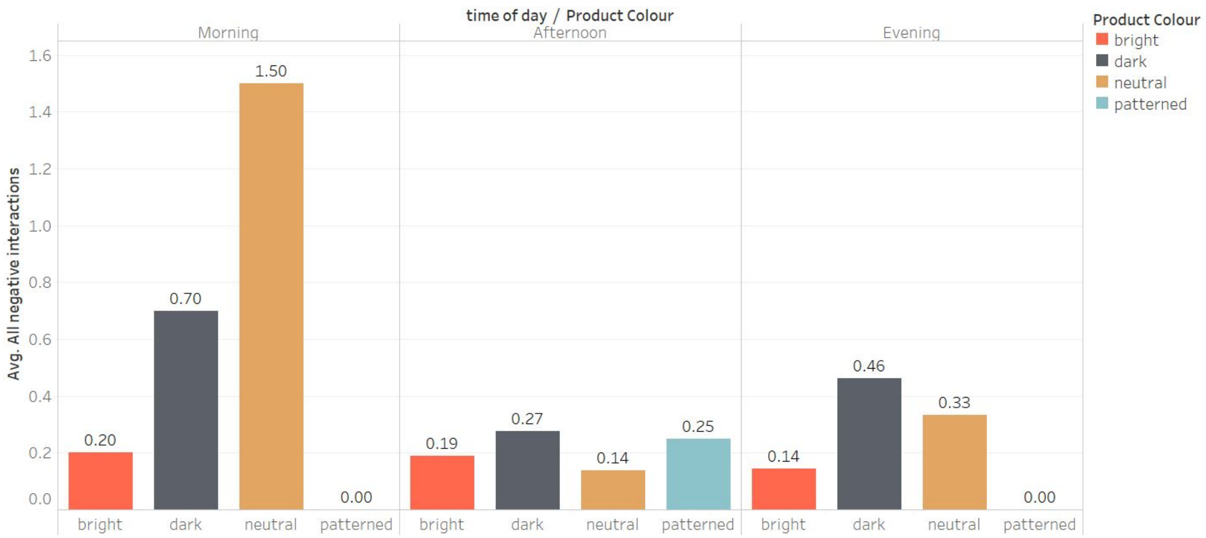

Figure 53: Total Hides of Product Colours

From the graph above, there is a significantly higher chance of users hiding neutral coloured products in the morning as compared to the rest of the colours. Seamless Bespoke could post more patterned and bright coloured products in the morning instead as the users respond better to it.

b. Total Hides of Miscellaneous posts by time of the day

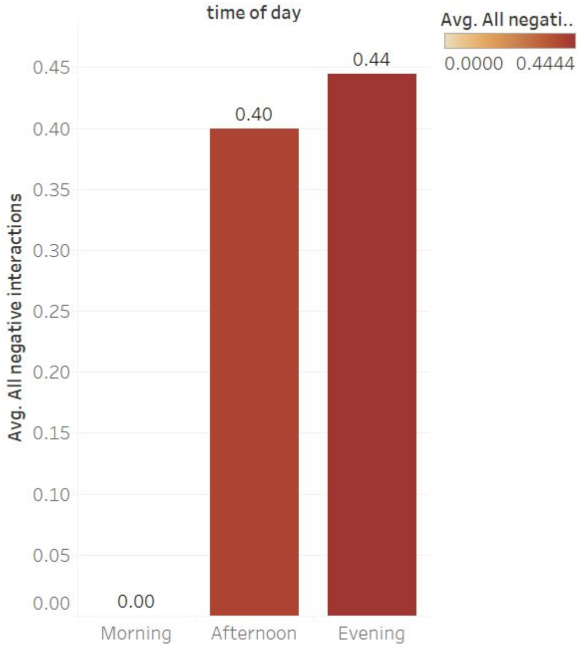

Figure 54: Total Hides of Miscellaneous posts by time of the day

Miscellaneous posts include: Events, Promotions, Announcements

Based on this graph, it is best to post Events, Promotions, Announcements in the morning as the users are least likely to hide it and will be able to view it again.

7.5.3.2 Total Hides grouped by Female, Male, and Gender-Neutral products

Gender Neutral Products include Fabric samples and Miscellaneous which is are events, promotions and announcements

a. Total Hides of Product Colours

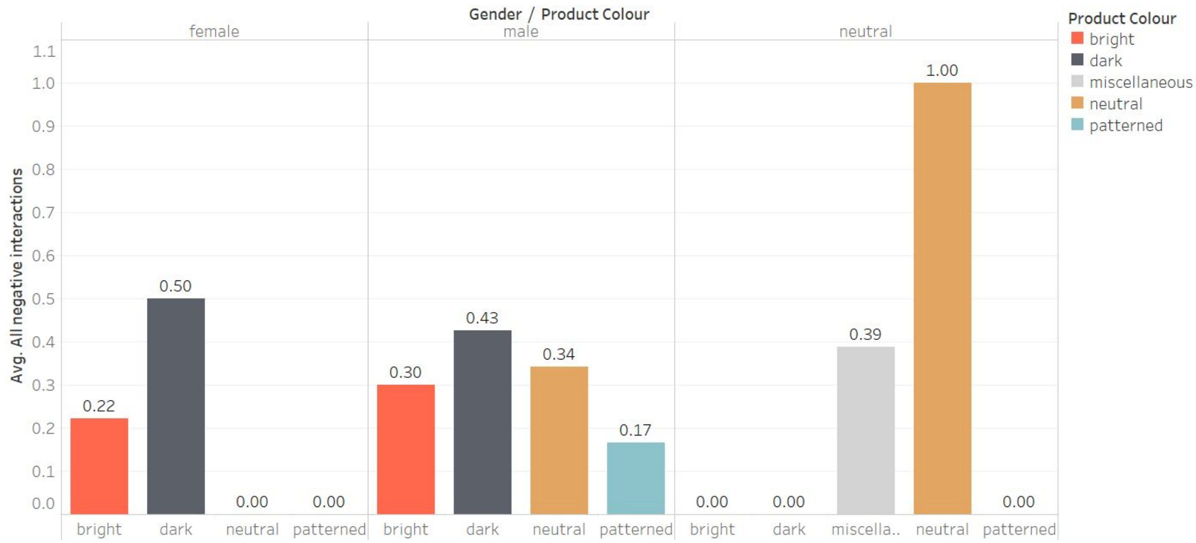

Figure 55: Total Hides of Product Colours

Under the gender-neutral column, miscellaneous refers to all events, promotions and announcements while all the other colour categories are for the fabric samples and full set outfit of BOTH male and female.

When splitting the products up by colour and gender, within the female products, the least negative interactions (least hides) come from the neutral and patterned products which shows favour towards these products.

Within the Male products category, they seem to favour the patterned products as well.

Within the gender-neutral category ignoring the announcements and events, neutral coloured fabric samples are hidden the most. This may mean that users like more bold coloured fabric samples or full set outfits.

b. Total Hides of Product Type

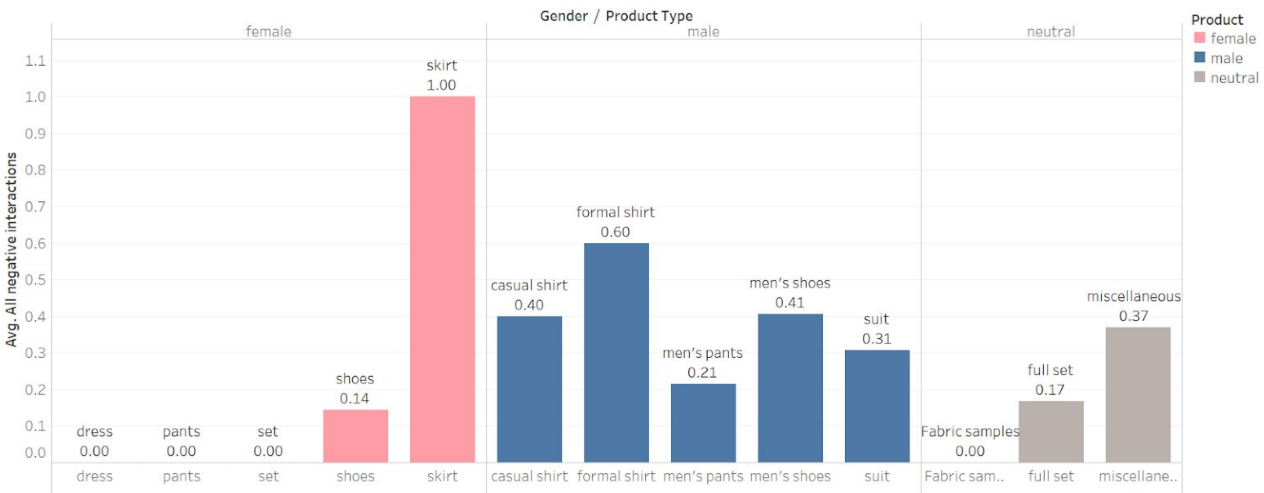

Figure 56: Total Hides of Product Type

The product that is worth noting in this category is female skirts. Perhaps the photographs could be presented in another way.

8. Recommendations based on Facebook insights

Bright coloured products do the best in capturing users' attention, while people are generally not as interested when they view a product of neutral colour. Users generally dislike seeing neutral coloured products, especially in the morning.

It is recommended that SB post more bright coloured products or posts to capture users' interest, and if they were to advertise for neutral coloured products, do not post that in the morning.

For events, promotions and announcements, it is best to post them in the afternoon as people are most receptive to it during that time, and they are interacting with these posts the most based on the graph. In the evening, interaction is the lowest and more people are likely to hide these posts.

In terms of posting gender-specific apparel, it is better to post female products in the morning while it is better to post male products in the afternoon. Overall engagement level is lowest in the evening. In terms of product types, full outfits posted are better for marketing.

Posting on Mondays and Saturdays is recommended as they garner the most interest and engagement. Avoid posting on Sundays as the reach is not that wide and level of engagement is poor.

9. Team Contribution

The team felt the everyone contributed equally towards the success of this project. As such, the team feels that every member should be given 5/5 for contribution. Yijia got in the client, led the HWZ scraping and presentation. Suyee headed the topic models and sentiment analysis. Chengwei led the Instagram scraping and plotting of the social network graphs. Devyn managed the analysis and insights of the platforms and the report. Clarice headed the facebook data analysis and the poster. Though the above merely stated the areas each team member took lead in, every member contributed a fair amount toward every section and step of this project. Scraping together using every single one of our machines, discussing recommendations to bring SB to greater heights and fighting for what each member believes is the most meaningful insight is what the team will remember for a long time to come. This is a unique experience, unlike any other projects. The team would like to take this opportunity to thank Seamless Bespoke for entrusting the team with their data as well as Professor Kyong for the advice and guidance since the very beginning of this project.

10. Conclusion

In conclusion, the team managed to identify what are the top 5+1 topics in HWZ forum related to Seamless Bespoke. On top of that, the team also managed to find the top 5 positive and negative users towards SB in the forum.

For Instagram, the team managed to find out that there is a lack of stickiness between the users and SB's posts. The team found 2 potential influencers and realised that SB's competitors are closely matched (in terms of degree, closeness and betweenness centrality) to them in the 2nd layer of the social network.

The team had also found out the best timing to post different content type on Facebook.

The team had also come out with 4 recommendations for SB:

- Avoid liking their own posts and tagging themselves on social media
- Launch "post-review" campaign to encourage users to post reviews about their products
- Collaborate with the two influencers (user 377 and 20362) for future social media marketing
- Fine-tune post content based on the content type and best post timing we identified