RMIT

RMIT International University Vietnam

Assignment Cover Page

Subject Code:	**COSC2789**
Subject Name:	**Practical Data Science**
Location & Campus (SGS or HN) where you study:	**SGS Campus**
Title of Assignment:	**Assignment 2: Data Modelling and Presentation**
Student name:	**Tran Tien An, Do Minh Huy**
Student Number:	**s3699000, s3751305**
Teachers Name:	**Yongli Ren**
Group Number:	
Assignment due date:	**13 Sep 2019**
Date of Submission:	**13 Sep 2019**
Number of pages including this one:	**11**
Word Count:	**3003**

RMIT

TABLE OF CONTENT

RMIT

Data set: Facebook live sellers in Thailand dataset

Data set information: The variability of consumer engagement is analysed through a Principal Component Analysis, highlighting the changes induced by the use of Facebook Live. The seasonal component is analysed through a study of the averages of the different engagement metrics for different time-frames (hourly, daily and monthly). Finally, we identify statistical outlier posts, that are qualitatively analysed further, in terms of their selling approach and activities.

Data Set URL: *http://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand#*

Investigating Thai Seller's Facebook Status

I. Abstract / Executive summary

The main purpose of this project is investigating the Online selling market and customers' behaviors of Facebook in Thailand, then provide some recommendations for the sellers so that they could approach their marketing and selling goal. The data set was conducted by collecting every single data of a Facebook's post such as number of likes, number of reactions, … from 10 Thai cosmetics and fashion retail seller's post on Facebook. We could separate the likes and reactions into smaller categories in order to know how do the customers feel about the posts. This study aims to determine how Facebook social environment can help resellers approach their selling plan, based on the research which was collected from Thai reseller's Facebook pages. The results indicate that the photos have been used the most and it also have the most reactions. Moreover, the users should not post links or simple status quotes because the consumers give no attention to those status types. The research concludes that most of the post are in photo or video type will gain a lot of attention. It is recommended that the Thai resellers should have their post as a video with great contents in order to reach the customer's interaction.

II. Introduction

Social media has become a popular phenomenon in this 4.0 digital era. One of the most familiar social networks is well known as Facebook. It has become an essentials element of every citizen's daily life. Users are familiar with the definition of online selling and trading market. Facebook Live is a platform which allow retail sellers to sell their products thought their post (Videos, Photos, Status). The number of users in this platform is increasing day by day for marketing and selling purpose. The main goal of this project is to determining the intrinsic group of unlabeled posts. From those investigation, some recommendations might be drawn for the sellers in order to help them reach their online selling goals. Live selling on social media, Facebook in particular, is becoming increasingly popular in Asian countries. Small suppliers are now able to reach a wider field of audiences and connect with thousands of customers. In order to archive this project, there are 4 main steps for in order to deal with this problem: cleaning the data, exploring the data, modelling the data by clustering task, providing and concluding the data by plausible hypothesizes and recommendations.

III. Methodology

The main achievement of this project is investigating the seller's post so that we could determine which type of post is good and suitable for this online selling and marketing field… Applying clustering task will help we achieve this assignment. the goal here is to determine the intrinsic grouping in this unlabelled data set. The clustering task is suitable for this achievement because it can help us find the useful and suitable groups for the sellers. In addition, the clustering task works excellently in the marketing field. It could group seller's posts which have similarity in the same groups or the posts which have the same characters into the same respective categories.

The research data set was conducted by collecting data from 10 Thai fashion and cosmetic retail sellers' Facebook activities. The data set is separated into 12 different attributes. The first three columns are status id (the id of the status), status types (the types of the status: Photo, Post, Video, …), status published (date of publishment). The 3 following column is Number of reactions, comments and share. The last 7 columns are the amount of each type of reaction (Like, Love, Wow, Ha Ha, Sad, Angry). The data set also has 4 extra blank columns so we will remove them in the cleaning data step.

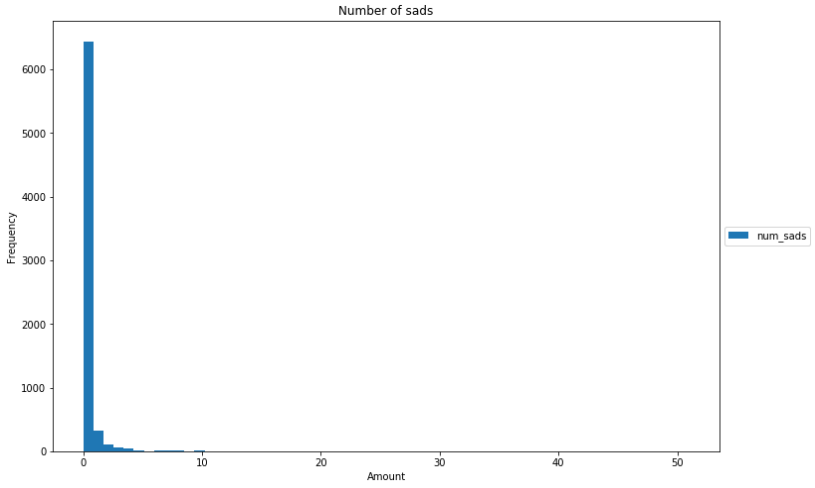
The first issue we have to deal is cleaning the data which is an important process. The data might have some mistakes or glitches so we have to handle those issue in order to have a high accuracy analysis. A small sanity check has been created in order to check if the data make any senses. After every mistake has been replaced or removed, the following step is data exploring. At this step, many graphs have been visualized so that we can investigate the relationship between columns. Many plausible hypothesises might be drawn in this process. The third step is grouping the data set by applying the clustering task to the data set. Clustering is a Machine Learning technique which will involves the grouping of data points. Clustering algorithm could be used to classify each data point into a specific group. In this situation, we could divide the data set into two main groups: status types and reactions. Finally, there will be a small discussion and some recommendation for the retail sellers in order to help them reach their selling and marketing goal.

In the training and validating model process, a good strategy is to split the data set into 2 parts, the first part is training (80% of the data set) and the second part is testing (20% of the data set). The idea here is to get the high accuracy of the model which the machine could learn. The first part is used to train our model on this data set, while the second part is used to make actual prediction using the trained model.

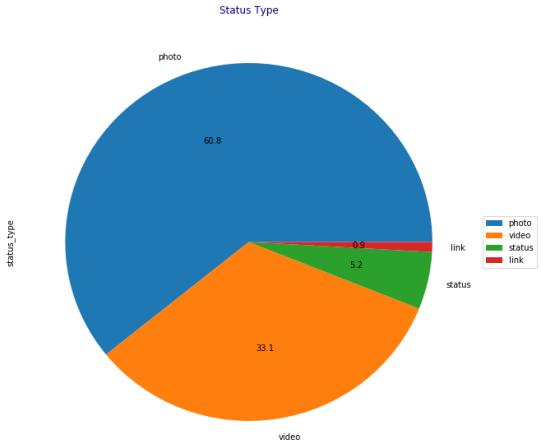
In the data modelling process, clustering task will be applied by using K-means and DB Scan technique. The fact that choosing k value for k-means is quite challenging so we decided to use the DB Scan in order to determine the number of cluster and then apply those number of cluster to the k-means technique. We will try to separate the main data set into different categories such as interesting posts, boring posts, interested and non-interested customers, … and then some recommendation could be drawn from the investigations for the resellers.

IV. Result

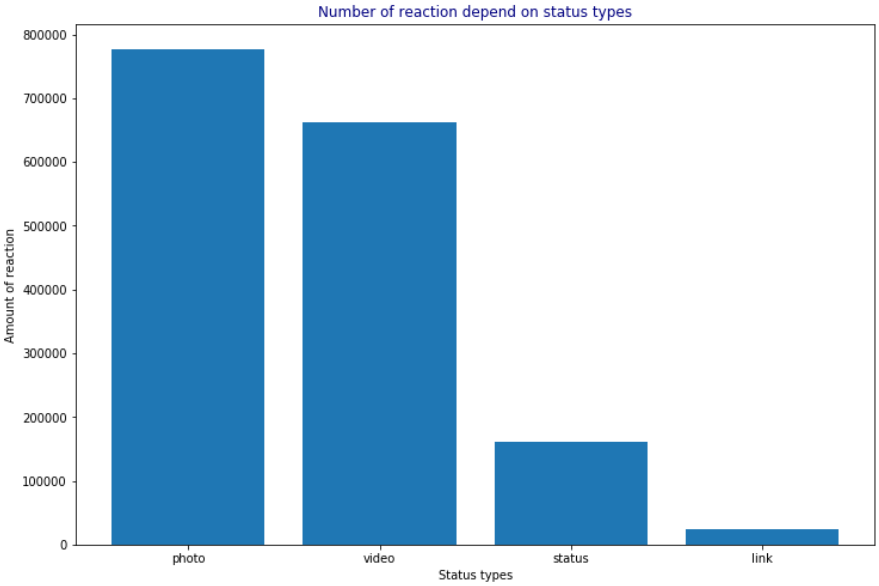
The numerical columns such as number of likes, loves, wows, Ha Ha, we will apply the histogram in order to show the density of the reactions. For example, from the histogram, we concluded that most of the status had the number of sad is 0 (more than 6000 rows). First of all, Facebook just have updated this functions since 2016. Secondly, there is no reason to be sad for a selling or trading post. A small number of consumer might be regret that they did not purchase the limited products. Not only the number of sad is extremely low, but also the number of angry and other negative types of reaction (except likes) are also at a low level. The like reaction has been used since the very first day of Facebook so every user is familiar with that reaction.



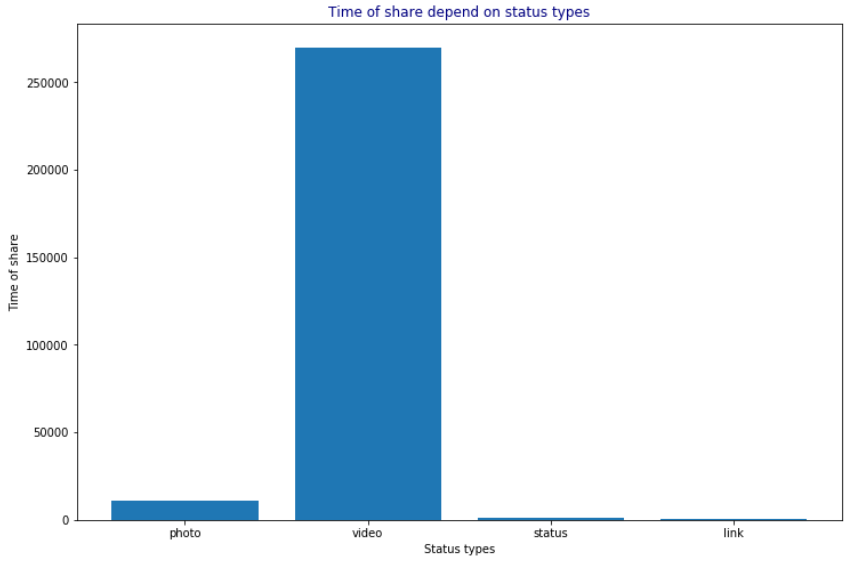
For the status type, the pie chart is applied because there are only 4 types of status: photo, video, status and link. As we can see above, the photo type occupies more than half of the chart with more than 60 percent. The second largest percent belongs to the video type. Normal status post and the link type is the smallest with approximately 7.0 percent in total. An assumption for this graph is that sellers tend to have their selling post as photos or videos. On the other hand, posting a link or a status is not a good idea.



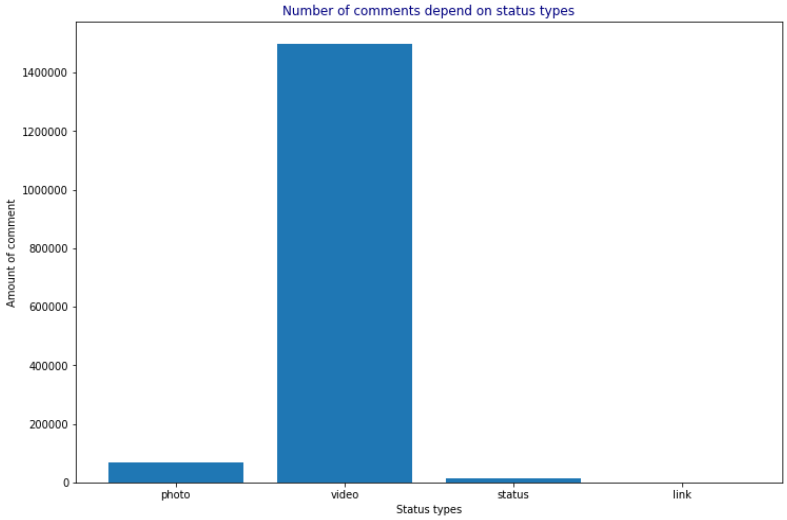
As we can see below, the photo type has the most reaction. Video type is in the second place. status type is in the third place and like type is in the last. The reason why we have this graph is consumers rather see and watch the post than read and click in the link. We could conclude that sellers should have their posts in photos or videos. They could have their customer's attention and interaction. The photo type has the most reaction because they could be funny memes or funny advertisements.



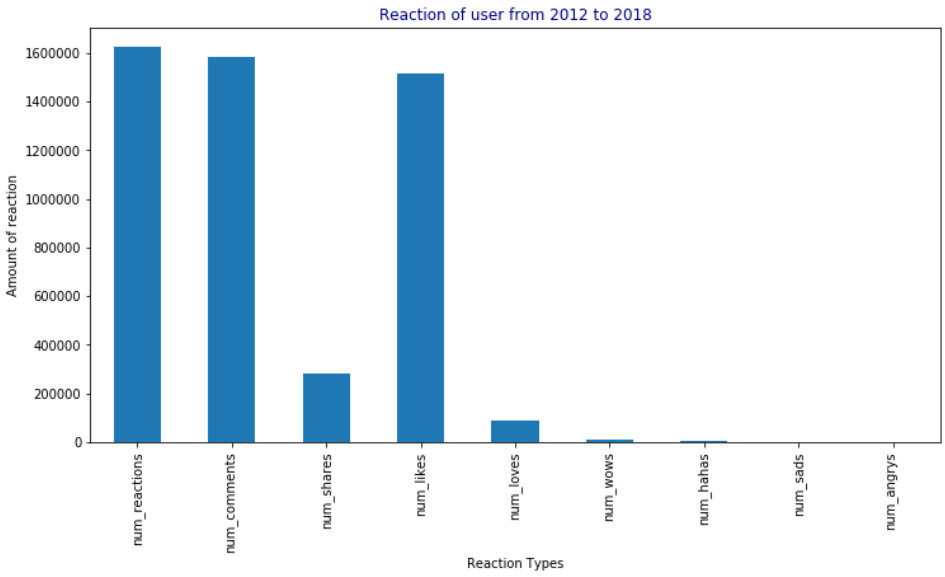
The graph below show the number of share depend on the status type. As we can see, the video type is dominating the others type in this chart with the highest number of share is more than 25 thousand. This phenomenon was happened because users tend to share videos funny clips with their friends. We can conclude that if the retail sellers want their advertisements shared widely, they should make videos clip or short movies instead of posting pictures or normal links and quotes.



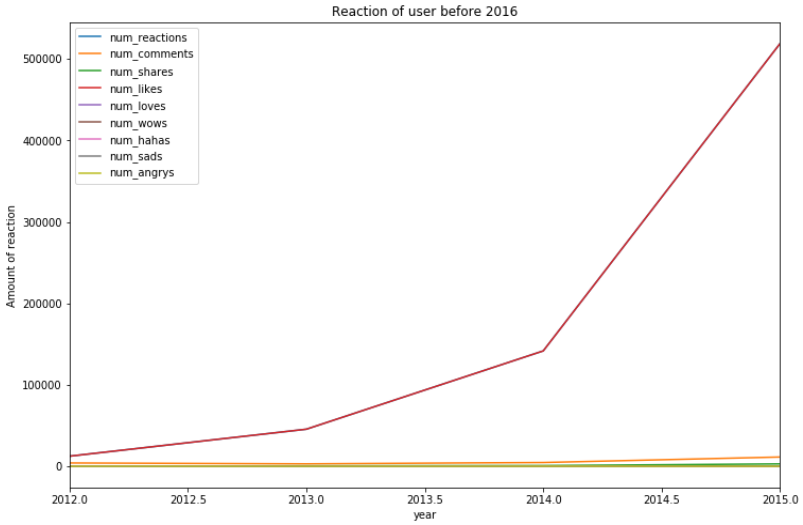
This bar chart shows the number of comments depends on the status types. The video type has the most comments and on the other hand, the link type has almost no comment. A conceivable assumption for this chart is that in the video type, the users tends to comment more because they want to interact with other people and know how other thinks when they finish the clip. Moreover, Facebook also have the live stream function so that users will have to comment so that they can speak and chat with the streamers. On the other hand, it is hard for them to read the link and give any comment or feedbacks because it takes them several step to read and comment on the link post.



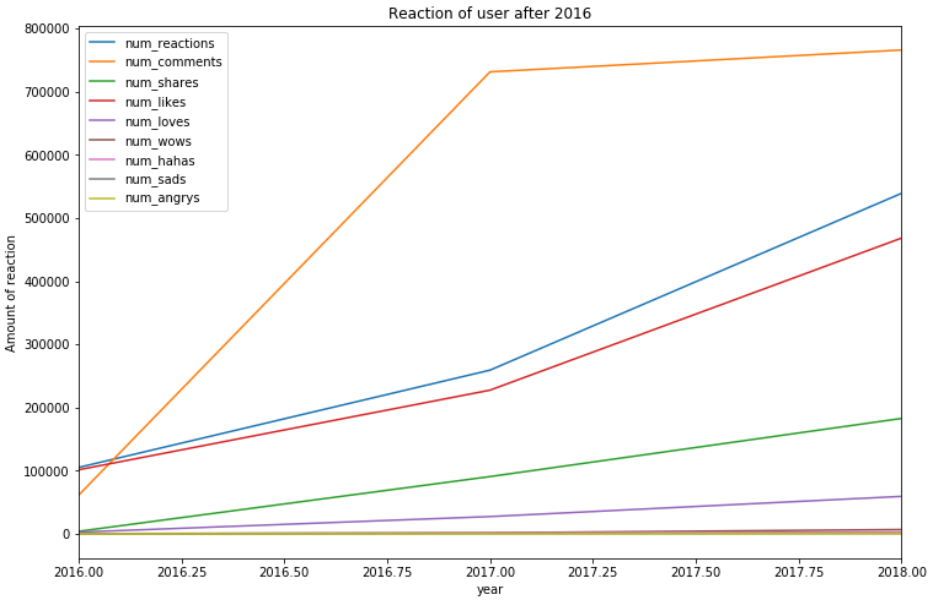
This bar chart shows the reaction of users thought year (from 2012 to 2018). This chart visualizes the amount of reactions, comments and shares of the users from 2012 to 2018. As can be seen in the chart, most of the customer will react and comment rather than share the post to their friends. A possible assumption could be think about in this chart is that these are the selling posts so costumers tend to react and comment about the products and the quality of the product.



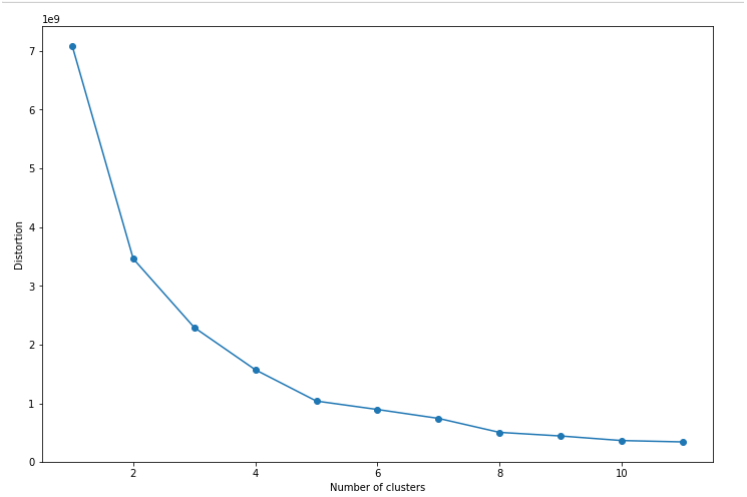
The 2 graphs below show the reaction of user before 2016 and after 2016. As we can see, before 2016 there is only one rising red line which is the number of likes. In this time line, there were no others reaction because the reaction function has just been updated since 2016. Moreover, we can conclude that Facebook Live selling market was not developed and the users were not familiar with this definition of online shopping.



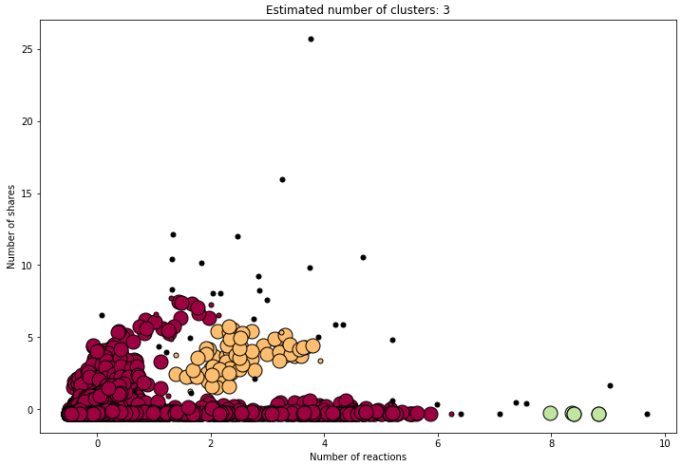
After 2016, Facebook has come up with the new reaction system. At that moment, users not only have like button but also they can react the post by choosing love, ha ha, wow, sad or angry. This innovation is really important because it could help the dealers know their customers well. They could understand the customer's behaviors and thoughts. It could create a social network trend. Users and explain their feeling or how do they feel after watch or read the seller's posts in an easier way. For instance, if they watch a funny clip and they do feel entertained, they could choose the Ha ha reaction and share it to their friends.



The chart below is the elbow method in order to find the k in the k-means technique. It can be seen that both 2 and 3 are possible values so we have to build 2 clustering model with k = 2 and k = 3. The value k in the chart is hard to choose because it can be both 2 or 3. In this case, we have decided to use the DB Scan technique in order to find the correct number of cluster, after that we can apply it to the k-means technique.

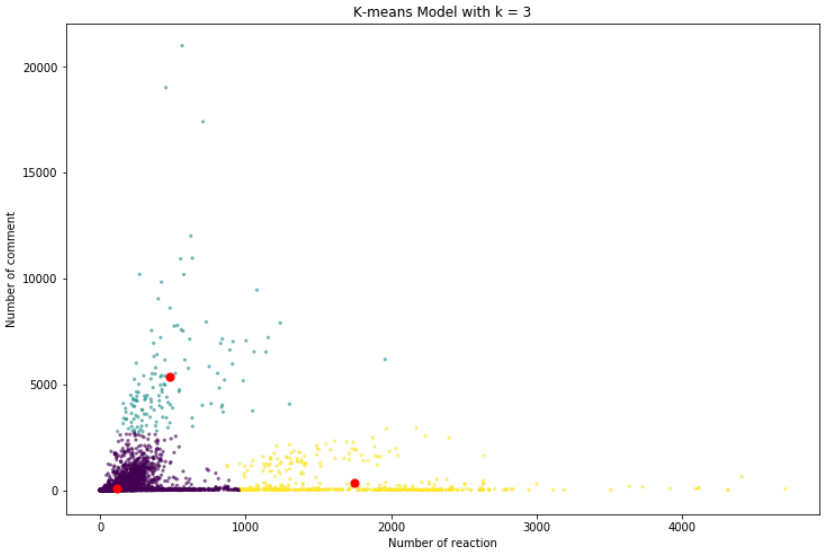


The model below is made by applying the DB Scan technique. The corresponding distance value on the first value or the knee point in the graph indicated the *Eps*. If the *Eps* is too small, many data will not be clustered, and on the other hand, if it is too huge, the clusters will be merged and the majority of the object will be in the same cluster. We initially consider the minimum points is equal to 4 (MinPts = 4) and the *Eps* = 0.5. The cluster model below has 4 clusters. The first cluster is the post with small number of reactions and shares. This is the majority of the data set with the red dot in the clustering table. The second cluster is the posts which has many reactions but less shares (white dots). The third cluster is the post with medium number of reaction and share. The final cluster is that the post with high amount of share but having a few reactions. the noises point (black dots) these are neither a core point nor density –reachable.




```
('number of noise(s): ', 33)
```

However, the main technique for this grouping task will be the k-means. After finishing the DB Scan technique, we have decided to use the strongest value k is k = 3. If the value k is equal to 2, there will be only 2 clusters in the model. As it can be seen from the below chart that with the value k = 3, the data set will have 3 clusters (purple, yellow, blue) with their centroids. The purple cluster is the posts which have no attention from the users. In this cluster, the number of comment and reaction are extremely low or even 0 for both axes. The second cluster which is yellow, this cluster is the posts with many reaction and without or less comments. The number of comment in this cluster can be zero. The last and the most important cluster in this modelling is the blue one. The posts in this cluster have many reactions and comment at the same time. This cluster is important for our project goal because we want the seller's post to be in this group.



V. Discussion & Recommendation

After the result part, many recommendations can be made for us in order to archive the project goal. As we can see above, the exploring task has many plausible hypothesises which we can think about. First of all, the number of reactions is not distributed equally to the reaction type because most of the customers find it easier for them to press the like butter than other reactions or they are just simple do not like the post. Secondly, most of the post is in the photo type which has the most reactions, however, the customers might react the photo because they might be memes or jokes and they do not really care about the product in the post. The video type is the second largest status type and they have the highest number of share.

From the modelling task, it can be seen that there are 3 main post group. The purple cluster is the group of posts without or less reactions and comments. These posts do not have any customer's interaction due to the fact that these posts might be boring or its contents are not suitable for the users. The posts which are in the yellow cluster are having a high number of reaction, however they have a small number of comment or just a small amount of comment. The possible theory for this cluster is the users liked the post without any attention. They just like the status because they are bored and they do not

really attend about the content of the post. The final cluster with blue colour is the most important ones in this project. The posts of this cluster are having many both reactions and comment. It means that the customers are having high attention for the content of these statuses. The main project goal is to lead the seller's port in to this cluster so they could reach their marketing and selling goal.

The best recommendation for the retail sellers is that they have their post in videos in order to approach the customer's attention. They should not post a simple quote or the product link because it will have a bad influence on the posts' interaction. Moreover, sellers also should care about the content of their posts due to the fact that there are 3 type of clusters and they will want their post to be in the blue cluster.

VI. Conclusion

In this paper, we have investigated the relationship between customers and retail sellers' Facebook posts, especially the characteristic of the posts. The main achievement of this project is to grouping the posts and customers which have the similar identity so that we could help the resellers approach their marketing and selling goal. Based on many charts and visualization, we have concluded that the sellers should have their post in video or photo type in order to reach the customers attention. Moreover, we have done the clustering task, exanimated those clusters, compared those clusters and determined which cluster will be used in order to archive the main project goal. In the near future, we plan to improve this model in order to help the retail sellers

VII. References list

Nassim Dehouche and Apiradee Wongkitrungrueng. Facebook Live as a Direct Selling Channel, 2018, Proceedings of ANZMAC 2018: The 20th Conference of the Australian and New Zealand Marketing Academy. Adelaide (Australia), 3-5 December 2018.