A drawing of a face

Description automatically generated

**IS434: Social Analytics and Applications**

**Group Project Report**

**Submission Date:** 20 July 2020

**Submitted To:** Professor Kyong Jin Shim

**Group 6 - FoodieBuddie**

Christal Poon

Gracia Yuwono Kwantalalu

Kok Jim Meng

Neo Yu Yao Terence

**Table of Content**

[**1.0 Introduction** 2](#_Toc45994077)

[**1.1 Background and Motivation** 2](#_Toc45994078)

[**1.2 Our Target Client** 2](#_Toc45994079)

[**1.3 Business Problem** 2](#_Toc45994080)

[**1.4 Deliverables** 2](#_Toc45994081)

[**2.0 Data Collection** 3](#_Toc45994082)

[**2.1 Instagram Scraping** 3](#_Toc45994083)

[**3.0 Data Engineering** 3](#_Toc45994084)

[**3.1 Corpus Creation for Data Labelling** 3](#_Toc45994085)

[**3.1.1 Corpus from Wikipedia Scraping** 3](#_Toc45994086)

[**3.1.2 Corpus from Web Scraping** 5](#_Toc45994087)

[**3.1.3 Identifying under-represented Influencers** 7](#_Toc45994088)

[**3.1.4 Corpus from Pareto Analysis** 9](#_Toc45994089)

[**3.2 Sentiment Analysis** 11](#_Toc45994090)

[**4.0 Data Analysis & Findings** 13](#_Toc45994091)

[**4.1 Dashboard for New and Existing Business Owners** 13](#_Toc45994092)

[**4.1.1 New Business Owners** 13](#_Toc45994093)

[**4.1.2 Existing Business Owners** 15](#_Toc45994094)

[**4.2 Social Network Analysis** 18](#_Toc45994095)

[**5.0** **Tools & Packages Used** 20](#_Toc45994096)

[**6.0 Limitations & Future Work** 20](#_Toc45994097)

[**7.0 Team Contribution** 21](#_Toc45994098)

[**8.0 References** 21](#_Toc45994099)

[**9.0 Appendix** 22](#_Toc45994100)

# 

# **1.0 Introduction**

## **1.1 Background and Motivation**

Singapore is a food paradise and it is a billion-dollar industry where it consists of S$8.3billion annual receipts, S$4.3billion annual contribution to the country’s GDP, and 48,000 workers are employed. According to the Ministry of Trade & Industry’s (MTI) Food Industry Transformation Map, there is a 4.5% of productivity growth and 2,000 new jobs are created in the food industry. Our team is interested in assisting the new and existing food business owners to maximise their share in this industry by recommending potentially trending food items or cuisine which they can sell and identifying suitable potential food influencers to work with to maximise their reach and exposure.

## **1.2 Our Target Client**

1. New business owners: people who want to open a food business
2. Existing business owners: People who already own a food business

## **1.3 Business Problem**

Due to the current COVID-19 pandemic, a lot of usual businesses in many industries are suffering. As the food industry continues to be stable despite the situation, we feel that it is an industry which business owners should consider going into. We recognise that opening and sustaining businesses are not easy as they need to take into account about the market demand and how to promote their products.

Through this project, we want to help new and existing food business owners by suggesting potentially trending food which they can consider selling, andtherespective suitable influencersto engage.

## **1.4 Deliverables**

The first deliverable is an interactive dashboard for business owners to identify potentially trending food based on cuisines, as well as suitable influencers to promote their businesses.

The second deliverable is asocial network graphwhich is used to analyse the network of food communities in Singapore.

# **2.0 Data Collection**

## **2.1 Instagram Scraping**

Instagram is our primary data source. We acknowledge the influence Instagram has on the food industry as it is a platform where food businesses and influencers post about food to promote or recommend them. Thus, we could analyse the posts of food influencers to identify potentially trending food. In this project, we narrowed down our scope to the accounts of the top 12 food influencers in Singapore based on their number of followers as follows:

1. *danielfooddiary* (239k)
2. *misstamchiak* (154k)
3. *ieatishootipost* (136k)
4. *sethluicious* (114k)
5. *foodkingnoc* (109k)
6. *eatbooksg* (75.7k)
7. *stormscape* (73.6k)
8. *8days\_eat* (64.8k)
9. *hungrygowhere* (62.7k)
10. *burpple* (61.4k)
11. *sgfoodonfoot* (54.9k)
12. *sheeatsshecooks* (53.6k)

Instagram posts were scraped using *Python Selenium WebDriver*, *BeautifulSoup* and *Pandas*. The posts data include the food influencers’ username, caption, number of likes, number of comments and a list of comments. These data will be used for text analysis using data engineering techniques, which will be discussed further in the following sections.

We also scraped the followers list of the food influencers using *Phantombuster*. These data will be used to generate the social network graph.

# **3.0 Data Engineering**

## **3.1 Corpus Creation for Data Labelling**

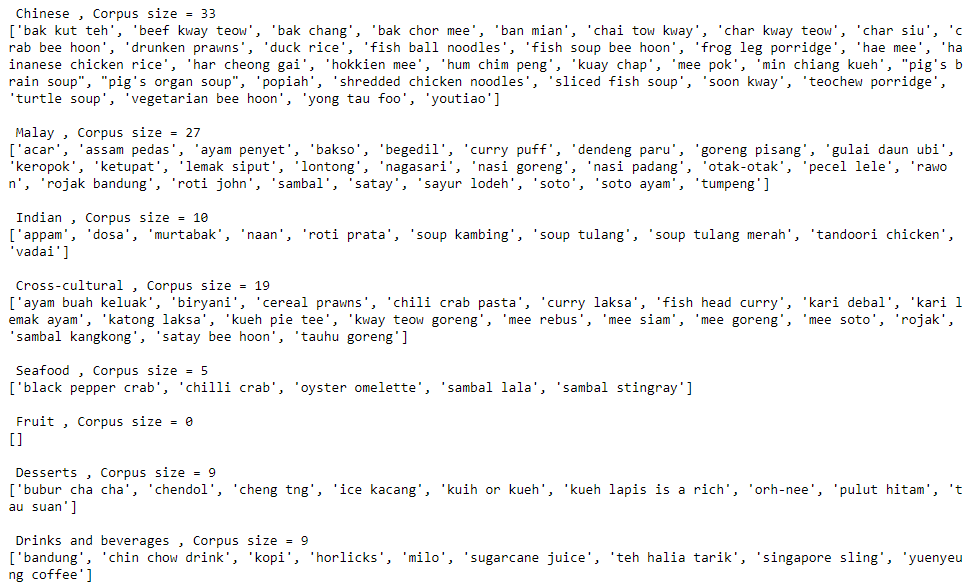
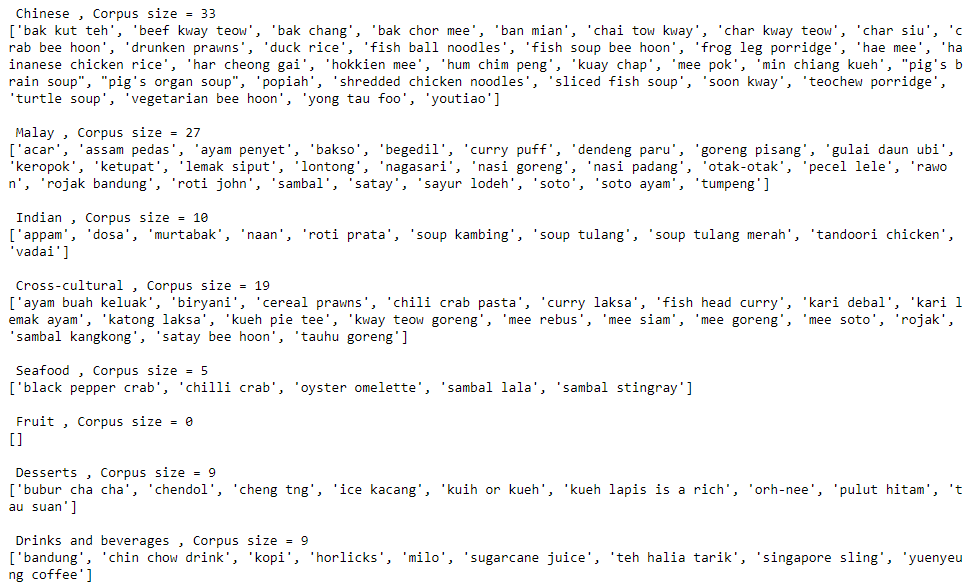
### **3.1.1 Corpus from Wikipedia Scraping**

In order to identify the food mentioned in the Instagram posts, we needed a corpus of food names and their corresponding cuisine names. With Wikipedia’s categories and multi-level subcategories features, a list of food is already associated to their respective cuisines, thus allowing us a quick way of generating a sizable corpus. In this project, we focused on the Singaporean cuisine sub-category as in Figure 1.



***Figure 1: Wikipedia Food and Drink Subcategories***

We scraped this data via the Wikipedia API using Python. Figure 2 shows the Singaporean cuisine corpus with 112 unique food names associated with 9 cuisines. Most food are categorised under Chinese and Malay cuisines.



***Figure 2: Initial Corpus from Wikipedia Scraping***

The data in the corpus is not clean. For instance, incorrect food terms such as ‘kuih or kueh’ and ‘kueh lapis is a rich’ are found. This can be due to the nature of Wikipedia being a voluntary-run project where information can be edited by anyone at any time. Thus, we proceeded to cleaning the data by amending the incorrect food names, as well as adding additional alternative food names as shown in Figure 3. This increased our search terms to 129 key words.



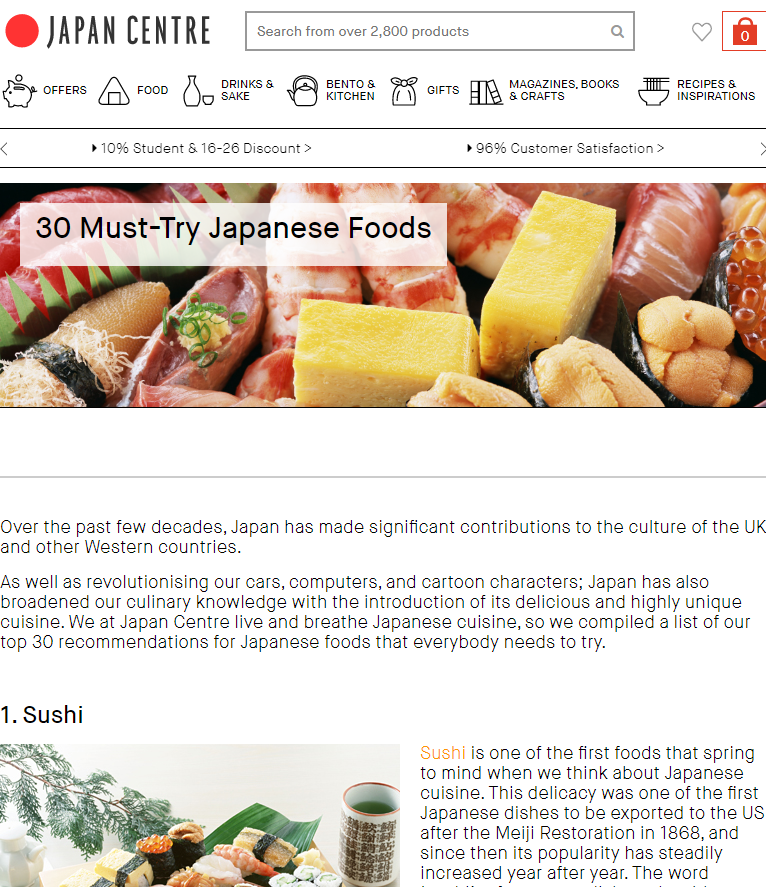
***Figure 3: Cleaned Corpus from Wikipedia Scraping***

After sorting the food into their respective cuisines, we realised that two major cuisines, Japanese and Korean, which are easily found in Singapore were missing as in figure 4. To add in these cuisines, we proceeded to web scraping.



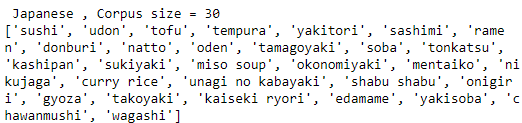
***Figure 4: Final Corpus from Wikipedia Scraping***

### **3.1.2 Corpus from Web Scraping**

***Figure 5: Food blogs for Japanese and Korean Food***

To make our food corpus as extensive as possible, we found 2 blogs as shown in Figure 5. We used *BeautifulSoup* to scrape from the Korean food blog. We also experimented with string parsing, which entails scraping code from scratch instead of using a pre-built package, to scrape from the Japanese food blog.

Our scraped results can be seen in Figures 6 and Figure 7. We successfully scrapped all the food names from the blogs, providing us with 30 Japanese and 29 Korean food names. To make our results more intricate, we have also included respective Korean characters for the Korean food in Figure 7.



***Figure 6: Japanese Cuisine Corpus from Web Scraping***

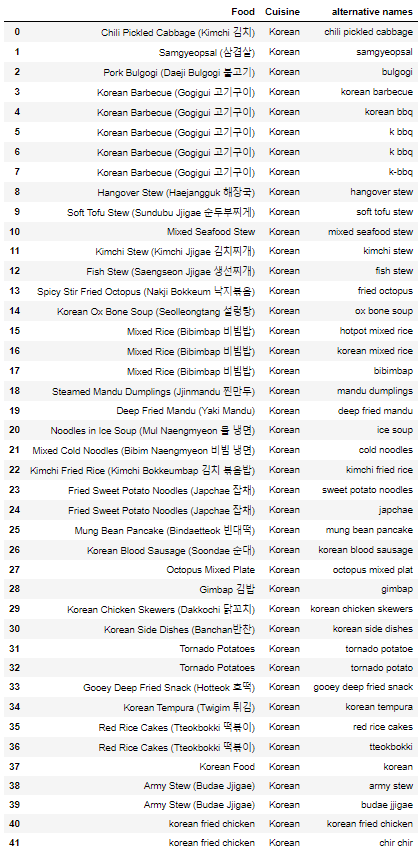
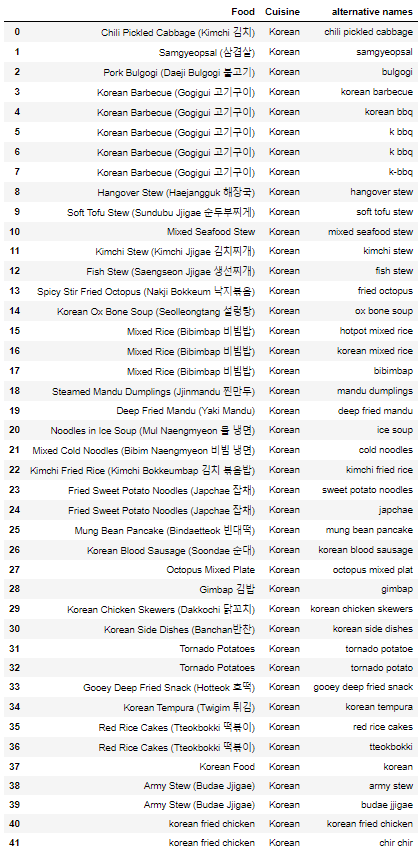


***Figure 7: Korean Cuisine Corpus from Web Scraping***

Further inspection of the corpus revealed two insights about Korean food:

1. Some commonly eaten food in Singapore were missing, such as Korean fried chicken, army stew, etc.
2. The scraped corpus uses food names which differ from names we use in Singapore (eg. KBBQ, ‘tteokbokki’ instead of ‘red rice cakes’)

Therefore, we expanded our corpus for Korean food, adding more common Korean food names in Singapore, as well as alternative food names, resulting in a corpus that extended to 41 search terms, shown in Figure 8.

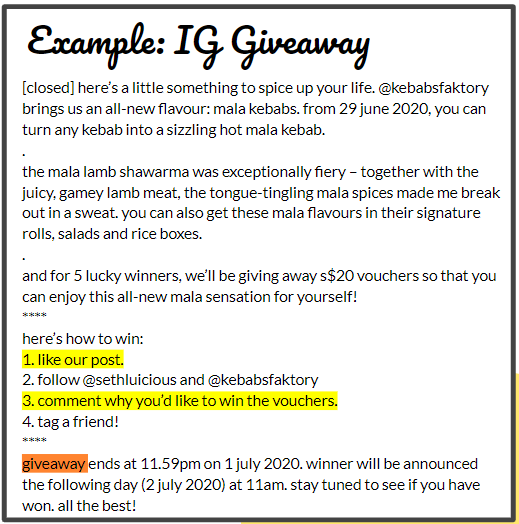
***Figure 8: Final Korean Cuisine Corpus***

### **3.1.3 Identifying under-represented Influencers**

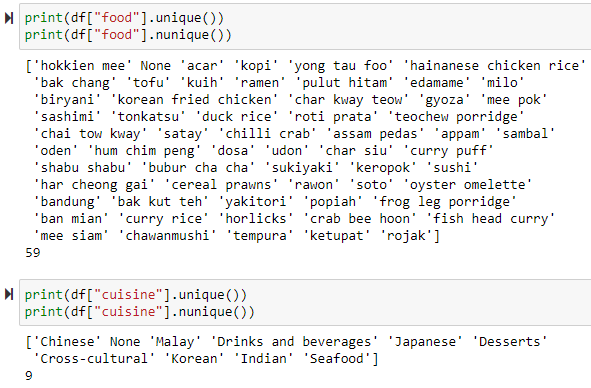
As this corpus is created independently of the Instagram dataset, we considered the possibility of food influencers being under-represented when using our corpus. We wanted to make our labelled data as balanced and representative as possible to draw reliable conclusions.

In order to check if our influencers have an equal proportion of labelled posts, we labelled the posts.

Firstly, we did data pre-processing such as labelling each post with a unique ID, changing all captions to lowercase, and preventing ‘giveaway posts’ from getting labelled by modifying the caption of these posts. *‘*giveaway posts*’* are posts where the influencer asks followers to participate in a lucky draw by liking and/or commenting on the post as in Figure 9. These posts artificially inflate the numbers of likes and comments and hence perceived popularity of the food. We shall exclude them from our analysis.

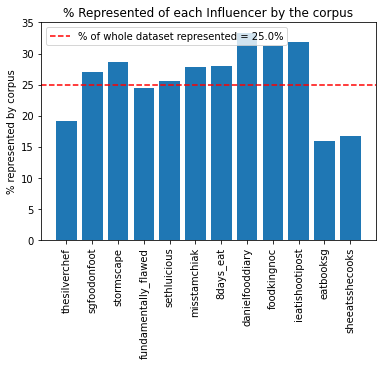
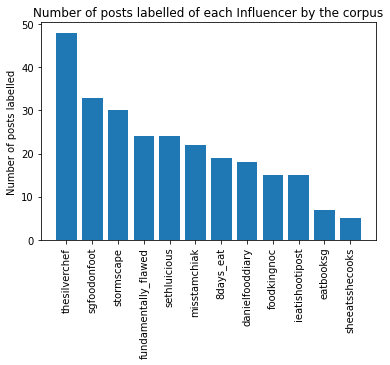


***Figure 9: Example of a Giveaway Post***



***Figure 10: Foods and Cuisines in Labelled Dataset***

Our results shown in Figure 11 showed that there are certain influencers who have a low labelled post to total post ratio. Here, we see that the influencers, *eatbooksg* and *sheeatsshecooks*, are significantly underrepresented by labelling the data from our scrapped corpus.

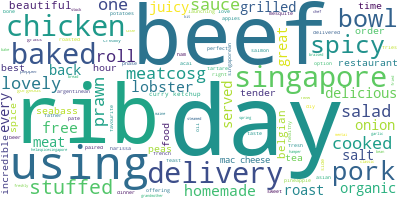
***Figure 11: Labelled Posts per Influencer***

### **3.1.4 Corpus from Pareto Analysis**

From the power law, we hypothesize that there are a small number of food names that appear in a large number of posts. An efficient way to quickly increase the number of labelled posts for these two under-represented influencers will be to identify their frequently occurring food names with pareto analysis. We started that by first cleaning our data with the following steps:

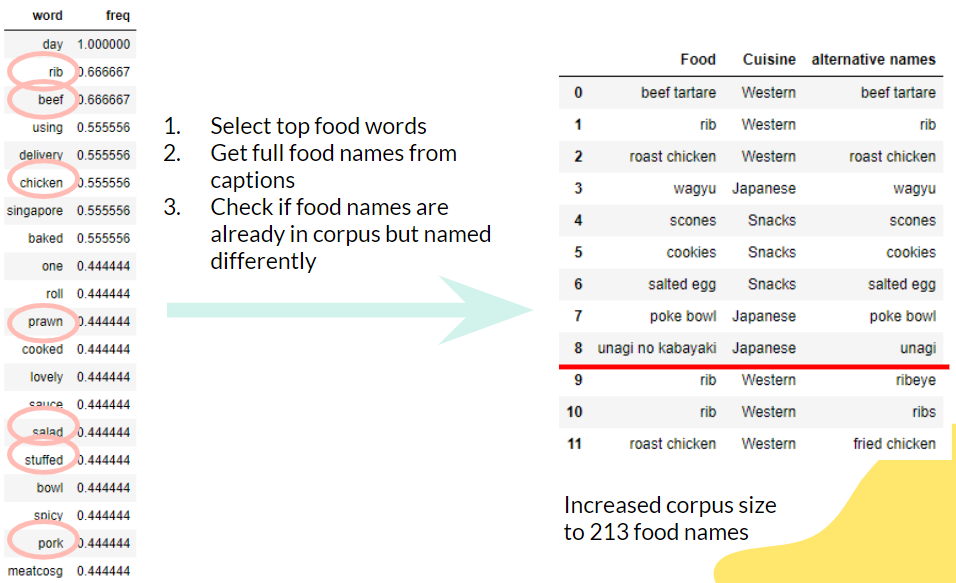
1. Regular expression to remove punctuation
2. Check that all texts are already in lower case
3. Remove stop words with NLTK

From there, we extracted out the posts by the under-represented influencers and did a named entity extraction to further narrow down our search space. We then observed how frequently each entity appears with a word cloud as in Figure 12.



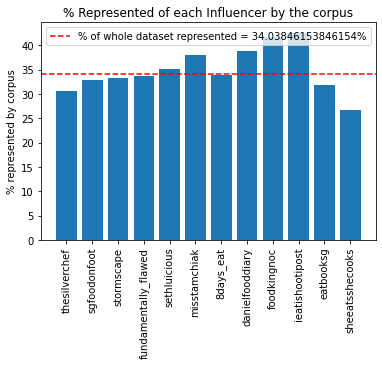
***Figure 12: Word Cloud of Entities***

From the sorted list of entities by their frequency of appearance, we picked the top few entities that may be a part of a food name as shown by the red circles in Figure 13. As food names are n-grams (made up of multiple adjacent words), we then observed the captions where these entities appear to obtain the full n-gram food names. We checked if these food names are already in our corpus but perhaps names differently. Finally, we arrived at 9 new foods and 3 alternative names.



***Figure 13: Identified new Food Names***

Adding these 9 new foods to our corpus, we obtained an outstanding result. We found that these 9 new food items increased our original number of labelled posts by close to 50%, doubling the number of labelled posts for *eatbooksg*, and increasing the number of labelled post for *sheeatshecooks* by around 66.7%.

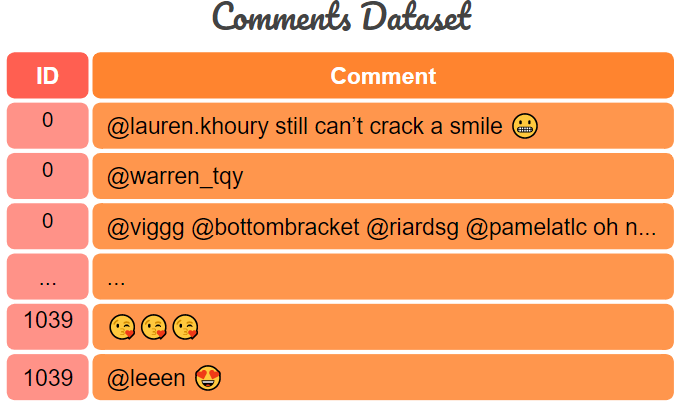


***Figure 14: Labelled Posts per Influencer after adding 9 new food names***

We can now complete the labelling of our dataset with 2 new columns for food name, and cuisine.

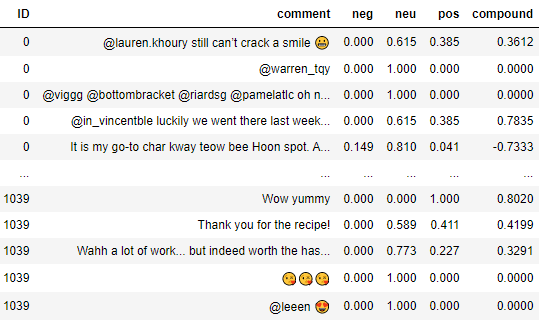
## **3.2 Sentiment Analysis**

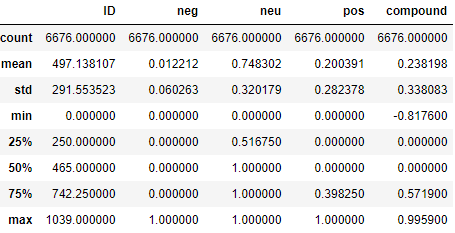
Our existing dataset contains metrics such as number of likes and number of comments for each post. However, a post with many comments may not necessarily mean people like the food mentioned in the post as some comments may be criticising the food. Hence, we conducted sentiment analysis on the comments using our comments dataset as in Figure 15 in order to create a more reflective metric. The ID here refers to the post ID.



***Figure 15: Comments Dataset***

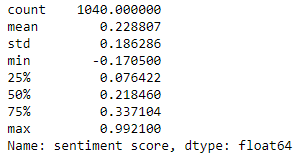
From our sentiment analysis, we gave each comment a compound sentiment score that ranges from -1 for very bad to +1 for very good as show in Figure 16.





***Figure 16: Comments Sentiment Scores***

The sentiment scores for the comments are then averaged to give each post a sentiment score that also ranges from -1 to +1. The summary statistics for our post’s sentiment scores are shown in Figure 17.



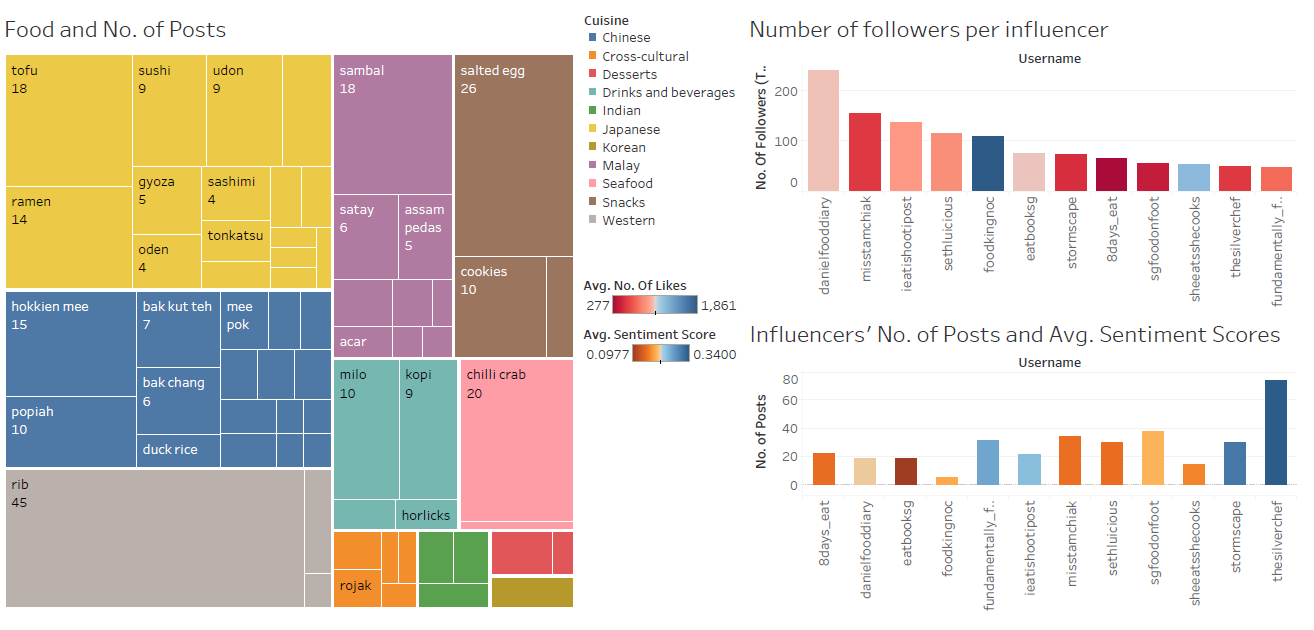
***Figure 17: Post Sentiment Scores***

# **4.0 Data Analysis & Findings**

## **4.1 Dashboard for New and Existing Business Owners**

Two separate Tableau dashboards were created for new and existing business owners. New business owners are able to view which food or influencers are recommended for them to sell and collaborate with in order to maximise their shares in the food industry. On the other hand, existing business owners are able to view which recommended influencers to collaborate with after selecting the cuisine which they are currently selling in the dashboard. This is because we recognise that it would be impractical for existing business owners to sell potentially popular food which are not in their domain.

### **4.1.1 New Business Owners**

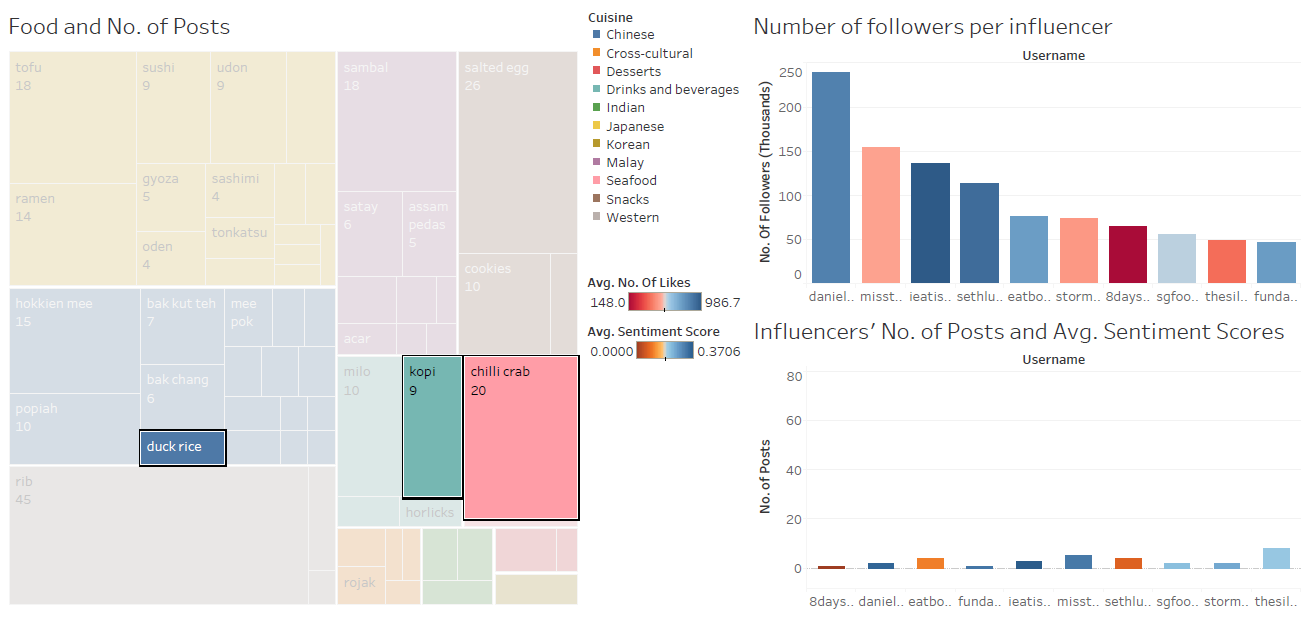


***Figure 18: New Business Owner Dashboard***

At first glance, new business owners would notice that Japanese and Chinese cuisines are popular in terms of the posts that the 12 influencers have published on Instagram, which are indicated in yellow and blue respectively in the heatmap.

New business owners can select the food they intend to sell from the heatmap and see which influencer to collaborate with by looking at the charts on the right side of the dashboard. At the top, the chart is based on the number of followers (bars) and the average number of likes (colours). At the bottom, the chart is based on the number of posts (bars) that the influencers have published and the average sentiment score (colours).

We can see that *danielfooddiary* is the influencer with the highest number of followers while *foodkingnoc* has the highest average number of likes. Furthermore, *thesilverchef* has the highest average sentiment score and has the highest number of posts.

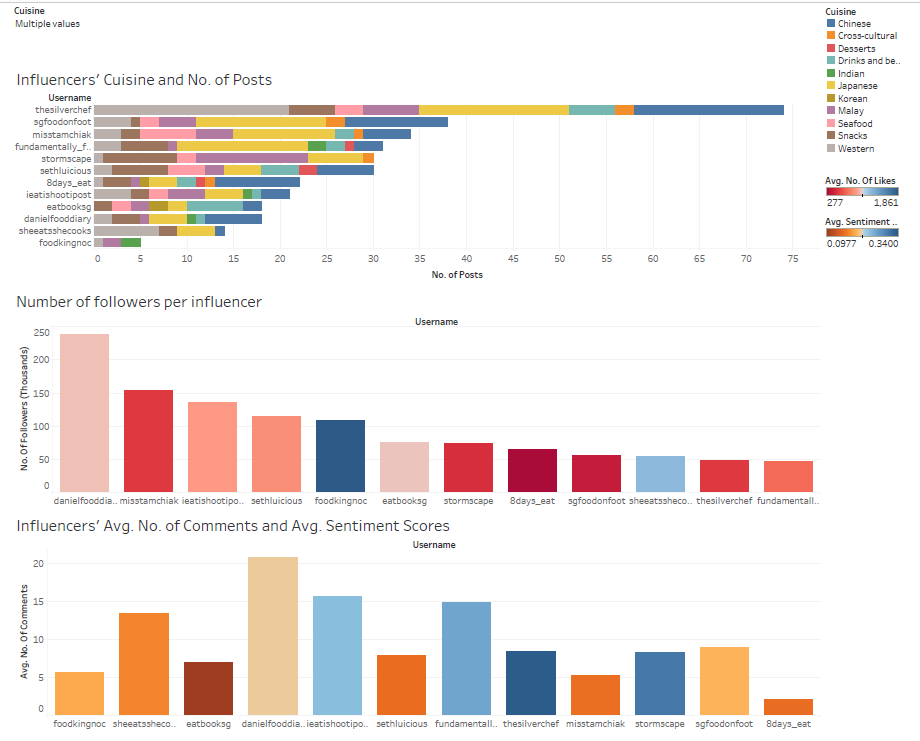


***Figure 19: New Business Owner Selected Duck Rice, Kopi, and Chilli Crab***

If a new business owner would like to sell duck rice, kopi, and chilli crab, he/she can select the food items in the heatmap as shown. Thereafter, he/she can see the influencers’ status on the charts at the right side of the dashboard.

In terms of the food items selected, *ieatishootipost* is the most suitable influencer that the new business owner can collaborate with. He has 136,000 followers and has a very high average of likes. Although he has less posts about duck rice, kopi, and chilli crab than the other two influencers (*misstamchiak* and *the* *silverchef)*, hehas a very high average sentiment score.

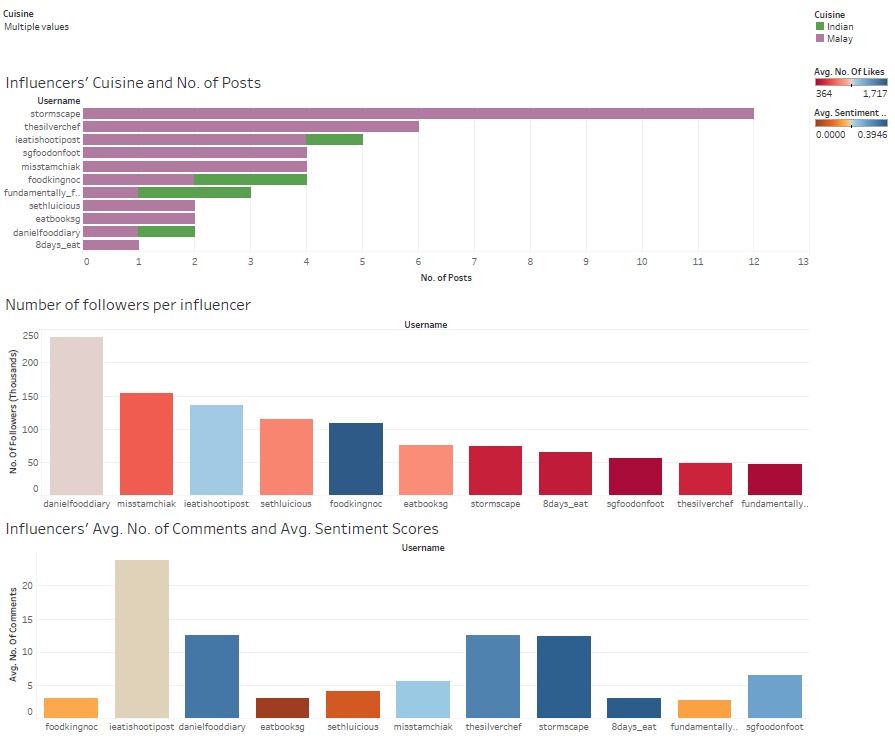
### **4.1.2 Existing Business Owners**



***Figure 20: Existing Business Owner Dashboard***

In the dashboard for the existing business owner, we notice that, among the 12 influencers, *thesilverchef* has the highest number of posts, which mainly focus on Western, Chinese and Japanese cuisines. Overall, the majority of the influencers post more on Chinese and Japanese cuisines except for one influencer, *stormscape*, who focuses more on snacks, Malay, and Japanese cuisine.

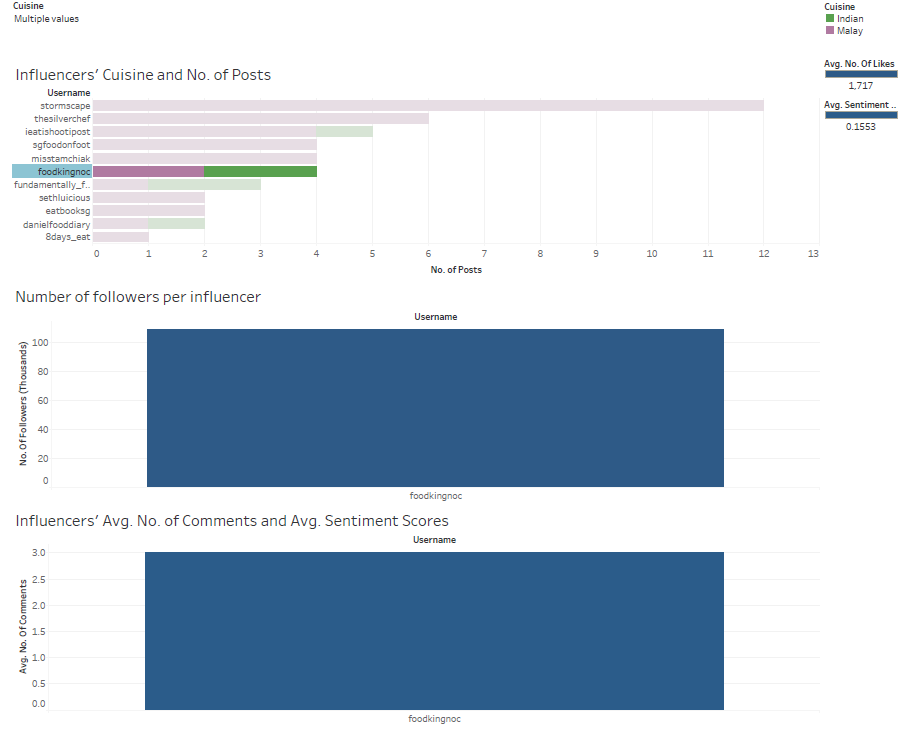
In terms of engagement, we can see that *foodkingnoc* has a very high average number of likes, and he also has the second highest average number of comments amongst the rest of the influencers. However, his average sentiment score is below average.



***Figure 21: Existing Business Owner Selected Malay and Indian Cuisines***

If an existing business owner sells Malay and Indian cuisines, he/she may specify the cuisines in the dropdown menu. The dashboard will then show the influencers which post about Malay and Indian cuisines.

*foodkingnoc* has the highest average number of likes, while *danielfooddiary* has the highest number of followers among the 12 influencers. *danielfooddiary*’s average number of likes is within the average range. *ieatishootipost* has the highest average number of comments, and both *stormscape* and *8days\_eat* have the highest average sentiment score among the rest of the influencers.



***Figure 22: Existing Business Owner Selected foodkingnoc After Selecting Malay and Indian Cuisines***

If the existing business owner would like to find out more about *foodkingnoc* since it published 50% of Malay cuisine and 50% of Indian cuisine, he/she can select the influencer in the bar chart.

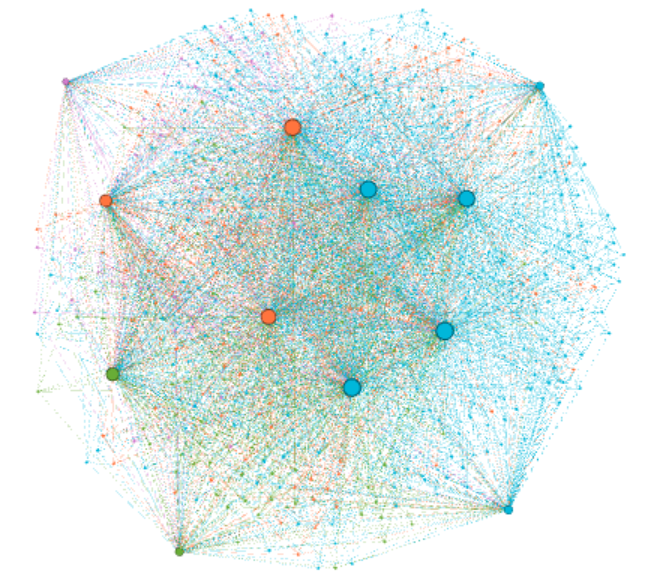
It shows that *foodkingnoc* has a high average number of likes and average sentiment scores. However, he has an average of 3 comments which is very few but, in terms of followers in the previous figures, *foodkingnoc* is ranked as having the 4th highest number of followers among the 12 other influencers.

The existing business owner may consider collaborating with *foodkingnoc* to promote his/her Malay and Indian cuisine business.

## 

## **4.2 Social Network Analysis**

We are interested to find out about the network of food influencers and their followers. In order to do this, we prepared a csv file of the mapping between the food influencers and their followers. We then imported this file into Gephi and obtained a social network graph. However, due to the massive number of nodes, the social network graph was not clearly visible, so we filtered out some nodes. The nodes in the following graph represents the top 12 food influencers in Singapore and people who follow at least 7 of these influencers.



**misstamchiak**

**sgfoodonfoot**

**stormscape**

**danielfooddiary**

**ieatishootipost**

**sethluicious**

**eatbooksg**

**thesilverchef**

**8days\_eat**

**foodkingnoc**

**fundamentally\_flawed**

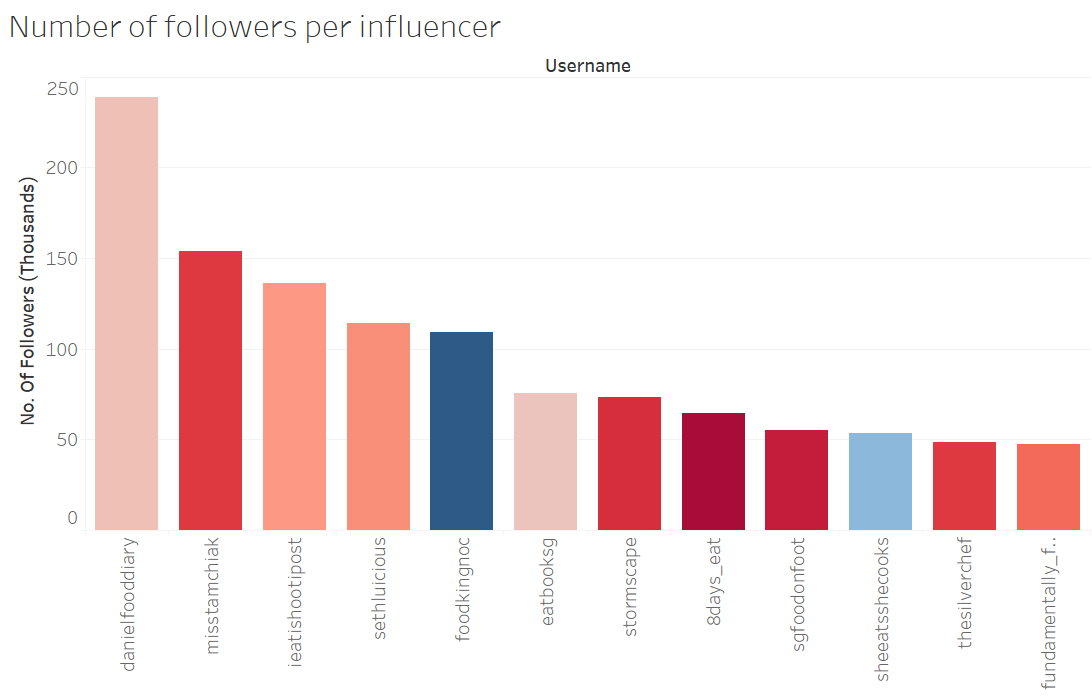
**sheeatsshecooks**

***Figure 23: Social Network Graph of Top 12 Food Influencers & Their Followers***

This particular graph is chosen because we can see disparate communities starting to form at this stage. What is interesting is the fact that the influencers themselves are grouped into different communities. Thus, we proceeded to investigate the similarities between the influencers in each community.

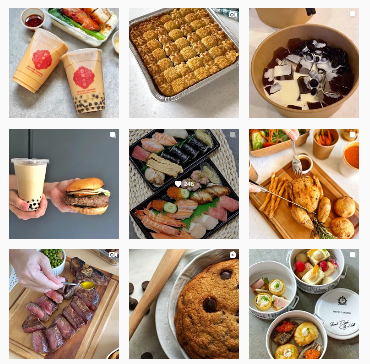
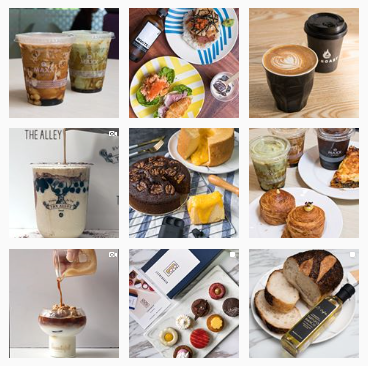
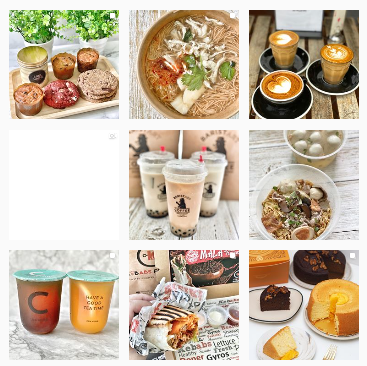
For the purple community, we found out that *sheeatsshecooks* posts a lot about raw meat as it partners with butcher @meatcosg. Perhaps her followers are interested in raw meat, but many people may not share this same interest.

For the green community, the two are the least popular of all the top 12 food influencers as shown in Figure 24.



***Figure 24: Bar Chart of No. of Followers per Influencer***

For the orange community, these influencers are different from the rest because they tend to post about light refreshments such as beverages and cakes instead of heavy meals.



***Figure 25: Orange Community’s Instagram posts***

Lastly, the blue community is grouped together because they are popular and have many followers. Perhaps the followers like to follow popular foodies as seen in Figure 24.

All in all, the fact that there are many followers who follow multiple food influencers indicates that the food community is close-knit, and they are fans of food instead of the food influencers. This observation is unlike the K-pop fandom, whereby fans pledge their loyalty to a particular group at a time.

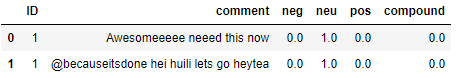
Due to the closeness centrality of the network, food trends can easily spread from person to person. This close-knit community provides evidence that the food industry is a good industry for business owners to consider entering or developing in.

# **Tools & Packages Used**

1. **Instagram Scraping**: Python (Selenium WebDriver, BeautifulSoup, Pandas) and Phantombuster.
2. **Web Scraping**: Python (Pandas)
3. **Data Visualisation**: Tableau and Gephi

# **6.0 Limitations & Future Work**

The main limitation of our project is a slight inaccuracy in sentiment scores due to misspelled words.



***Figure 26: Example of Inaccurate Sentiment Score***

While we have added commonly used abbreviations, this type of misspelling is yet to be resolved. In the future, we can explore options to correct this.

Another possible extension for our project is to get a more comprehensive dataset of labelled posts would be to train a binary classifier to classify words into “food” or “not food”. We found an open source model *en\_core\_web\_sm* which is able to classify n-grams (multiple words that form a name) into the shown classes. It will be possible to do transfer learning on this open source model to create our food/ not food classifier. This will be possible when there is a sufficient quality or quantity of training data.

# **7.0 Team Contribution**

|  |  |  |
| --- | --- | --- |
| **Name** | **Score** | **Contribution** |
| Christal Poon | 5/5 | * Designs * Poster * Report * Slides |
| Gracia Yuwono Kwantalalu | 5/5 | * Instagram Scraping * Social Network Analysis * Report * Slides * Medium Article |
| Kok Jim Meng | 5/5 | * Tableau Dashboards * Food/Cuisine and Influencers Analysis * Report * Slides * Medium Article |
| Neo Yu Yao Terence | 5/5 | * Corpus Creation * Sentiment Analysis * Website * Report * Slides * Medium Article |

# **8.0 References**

* <https://www.enterprisesg.gov.sg/industries/type/food-manufacturing/industry-profile>
* <https://en.wikipedia.org/wiki/Category:Asian_cuisine_by_country>
* <https://en.wikipedia.org/wiki/Category:Singaporean_cuisine>
* <https://en.wikipedia.org/wiki/Singaporean_cuisine>
* <https://monkeylearn.com/text-classification/>
* <https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/>
* <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>
* <https://simpleanalytical.com/how-to-web-scrape-wikipedia-python-urllib-beautiful-soup-pandas>
* <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da>
* <https://towardsdatascience.com/custom-named-entity-recognition-using-spacy-7140ebbb3718>
* <https://www.soompi.com/article/794453wpp/why-there-can-only-be-one-multifandom-in-korea>
* <https://towardsdatascience.com/step-by-step-tutorial-web-scraping-wikipedia-with-beautifulsoup-48d7f2dfa52d>

# **9.0 Appendix**

**Website:**

* <http://gg.gg/IS434FoodieBuddie>
* <https://terenceneo.github.io/Social-Analytics/>

**Medium Articles:**

**GitHub Repository**

* <https://github.com/terenceneo/Social-Analytics>

**Problem Statement Intro Video:**

* <https://youtu.be/adwsMpEWbAE>