STAT 5106

Programming Techniques for Data Science

Final Project Report – Group 12

Topic: Patterns of Fake Reviews in a Hong Kong based Restaurant Review Platform: Lies, damned lies, Openrice?

Group 12:

WONG Tin Yan Tim 1155110207

[redacted]

Table of Contents

1. Introduction

OpenRice, established in 1999, is one of most popular online dining guides and restaurant review platforms in Hong Kong. It provides a comprehensive database of restaurants, offering users details such as menus, ratings, reviews, and dining recommendations. However, like many user-driven platforms, it faces challenges related to the authenticity of its reviews, as fake comments can distort perceptions and mislead users. In this project, we analyzed restaurants, comments and user data on OpenRice to investigate patterns and characteristics that may help identify fake comments.

2. Objectives

Through the application of data science techniques, our project seeks to uncover patterns within comments and user data from OpenRice, with a focus on distinguishing genuine reviews from potentially fake ones. By identifying key characteristics of comments and behaviour of OpenRice users, we aim to develop practical insights and methods that empower everyday users to evaluate the authenticity of reviews.

3. Methodology

3.1 Data Collection and Preprocessing

Our data was collected from OpenRice.com, primarily from two sources – info on **restaurants** and **users**. Data collection started from 08-Nov-2024 to 12-Dec-2024. After dropping unsuccessful scrapes, 193,421 comments of the first five pages from the randomly chosen 5,898 restaurants were selected and their key details such as **name, address, cuisine types and comment texts were extracted**. Among the comments, we applied **systematic sampling to sample one user for every 50 comments**, and further collected info from the resultant 1,027 users. User information, such as username, user level, and total number of comments of the 1,027 users were gathered. **To mimic the normal behaviors of an end user**, details of only **comments from the first five pages** of each user summing up to 47,515 comments. Preprocessing involved cleaning "\n" in the body texts, calculating "average rating", "comment frequency", "sentiment score" with SnowNLP with comment text as discussed below and handling missing values.

```
Data columns (total 23 columns):
 #   Column
---  ------
 0   scraping_status
 1   member_name
 2   member_url
 3   member_level
 4   member_comment_count
 5   member_resto_count
 6   member_photo_count
 7   num_of_comment_in_this_url
 8   comment_header
 9   comment_restaurant
 10  comment_restaurant_url
 11  comment_date
 12  comment_text
 13  taste_rating
 14  env_rating
 15  service_rating
 16  hygiene_rating
 17  fair_price_rating
 18  comment_photo_count
 19  comment_text_with_emoji
...
 21  user_category
 22  days_per_comment
```

Fig 3.1 Dataset columns

3.2 Data Description and Structure

Total Restuarants	5,898
Total Sampled Users	1,027
Total Reviews	47,515
Sentiment Scores	[0, 1] generated by SnowNLP, Extremely Positive = 1, Extremely Negative = 0

3.3 Libraries Used and Their Applications

Category	Library	Purpose
Data Collection	requests	Used for API interactions to fetch sentiment analysis data by sending POST requests with review text.
	BeautifulSoup	Utilized for parsing and extracting restaurant metadata, reviews, and user information from OpenRice.
Data Processing	pandas	Employed to organize data into structured DataFrames, enabling efficient data manipulation and analysis.
	numpy	Used for handling numerical operations, such as imputing missing values and calculating statistical metrics.
Data Visualization	matplotlib	Core library for creating plots like line graphs and scatter plots.
	seaborn	Extended visualization library used for aesthetically pleasing and complex visualizations like box plots and bar charts.
	plotly	Utilized for creating interactive and visually appealing plots, such as scatter , for exploratory data analysis.
Sentiment Analysis	nltk, textblob, jieba, SnowNLP	Used for preprocessing review text (e.g., tokenization, stop-word removal) and generating sentiment scores.

3.4 Study Assumptions

We hypothesized that some characteristics of suspicious paid reviewers will differ from a normal user. They included **frequent commenting, higher than average rating per comment and higher no. of images and videos per comment**. Meanwhile, considering paid reviewers may also create "**shadow account**" and only comment once, we categorized users into Suspicious User group and Non-Suspicious User based on the following criteria:

	Criteria	Data Range	Rationale / Calculation
1	Days per comment (> 1 comment user)	<= 2.87 (median)	Max (comment date) – Min (comment date) / No. of comment
2	No. of Comment	= 1	Some paid users would use "shadow account" to comment for only once
3	Average Comment Rating	>= 4.4 (median)	average of "taste rating", "environment rating", "service rating", "hygiene rating" and "fair price rating"
4	No. of Image and Video per comment	>= 10 (median)	Paid users are expected to upload photos of the restaurant to make the post seems persuasive.

When a user meets criteria **1,3,4** or **2,3,4,** the user will be classified as suspicious user.

4. Demographics

4.1 Demographics of Users

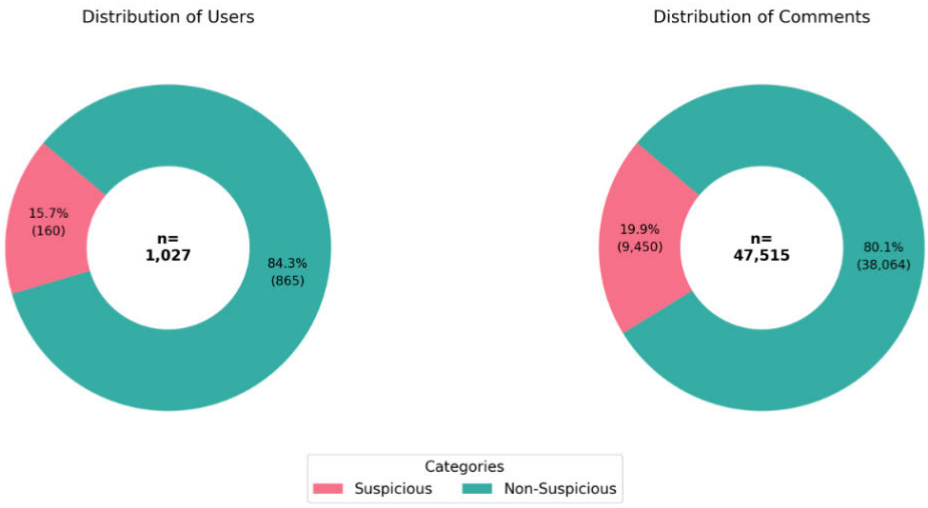

Figure 4.1 Distribution of Users and Comments with our criteria of classification into Suspicious and Non-suspicious Users

With the aforementioned categorizing criteria, approximately 1 out of 6 users are considered suspicious, and around 1 out of 5 comments are left by the suspicious users. This indicates that the **no. of comments are commensurate to the no. of users**, and the distribution of suspicious users and comments **is unlikely to follow a Pareto Distribution (80/20 distribution)**. Among the entire population and suspicious users, **level 4 user accounts predominantly accounted for 81.3% and 98.8%** in that subgroups.

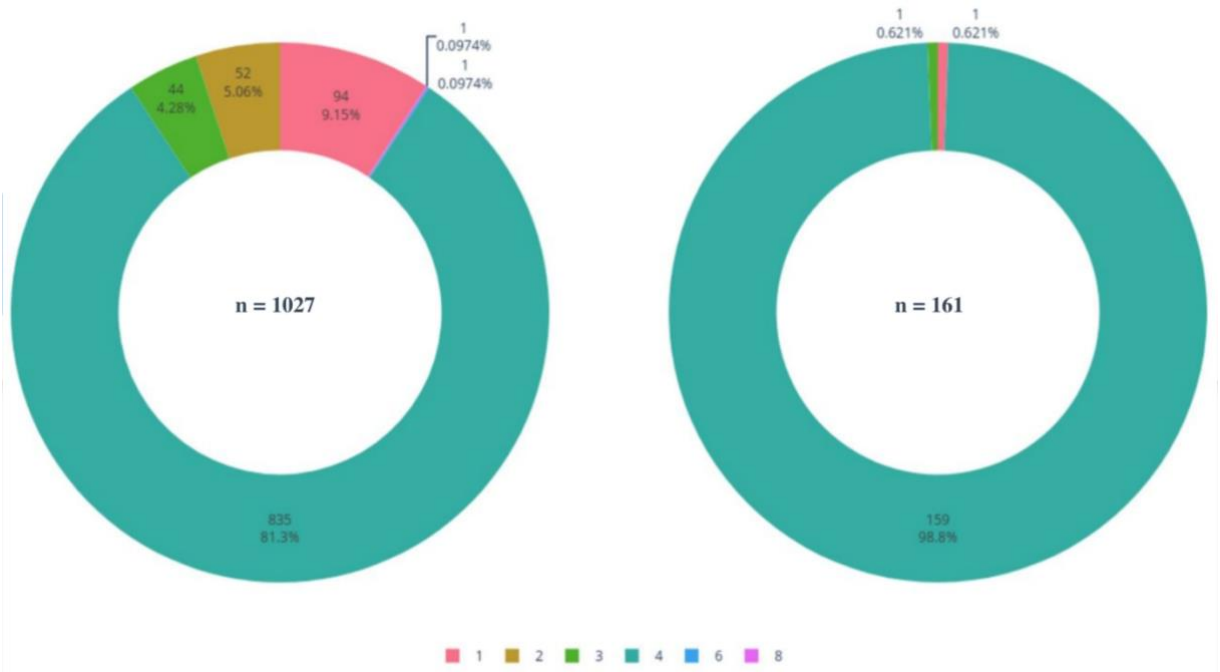

Figure 4.2 Distribution of All Users and Suspicious Users by Member Level of User Accounts

4.2 Demographics of Restaurants

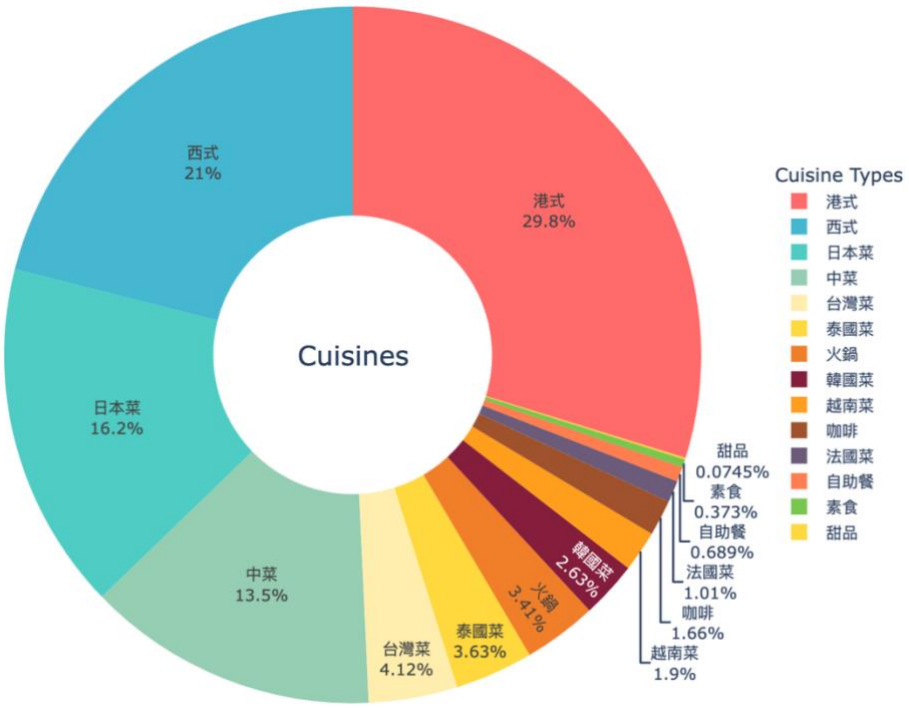

Figure 4.3 Proportion of Cuisine Type

港式 (Hong Kong cuisine) makes up the largest share at 30.2%. 日本菜 (Japanese cuisine) is the second largest at 18.4%. 西式 (Western cuisine) follows at 15.9%.

5. Data Visualization and Analysis

5.1 User Account Patterns

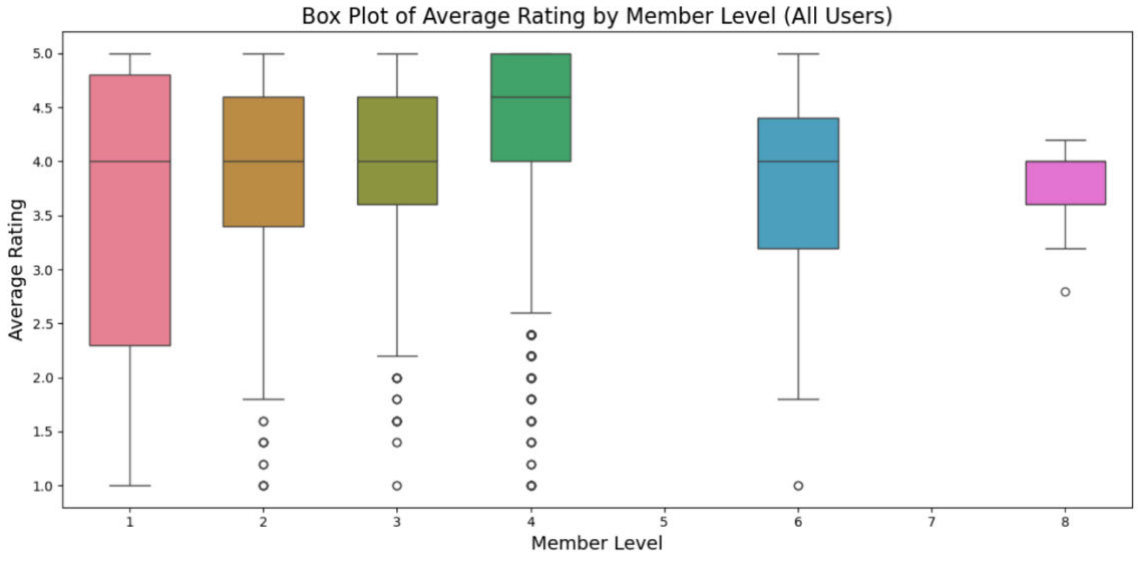

Figure 5.1 Box plot of Average rating by Member Level in All Users

Median of Level 1: 4.0, Level 2: 4.0, Level 3: 4.0, Level 4: 4.6, Level 6: 4.0 Level 8: 4.0

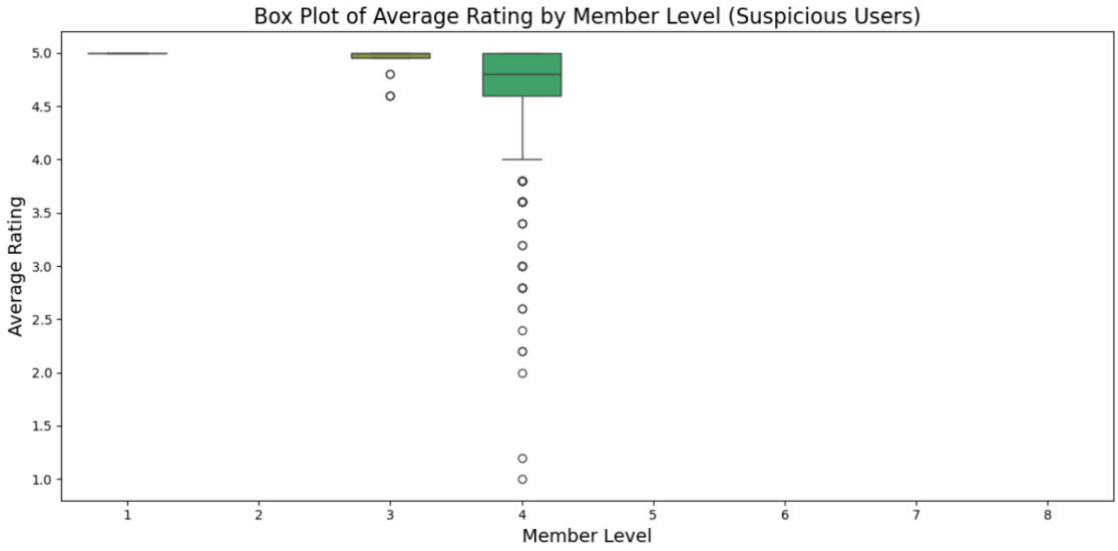

Figure 5.2 Box plot of Average rating by Member Level in Suspicious Users

Median of Level 1: 5.0, Level 3: 5.0, Level 4: 4.8

每當會員的食評數量及質量達到晉級的要求，即會自動晉升為另一級的會員。會員級別越高，可享用的個人化功能就越多。

級別	所需要求		食評上載相片	修改食評	會員有禮	開飯活動	優先報名	智尊專欄	智尊專頁	開飯智囊	食評即時刊登
	已刊登食評	編輯推介									
1	0至4篇		30	✗	✗	✓	✗	✗	✗	✗	✗
2	5至29篇										
3	30至99篇										
4	100篇或以上										
5	100篇或以上	最少20篇*	60	1次(30日之內)	✓	✓	✓	✗	✗	✗	✗
6	100篇或以上	最少50篇*									
7	500篇或以上	最少100篇**									
8	1000篇或以上	最少200篇**									
9	2000篇或以上	最少350篇**									
10	5000篇或以上	最少500篇**									

*包括過去12個月內發表3篇食評並獲「編輯推介」或以上的食家

**包括過去12個月內發表5篇食評並獲「編輯推介」或以上的食家

Figure 5.3 Requirements and Rights of Different Levels of member accounts on Openrice.com

The **overall scoring** given by reviewers are mostly **negatively-skewed**. From Fig 5.1 and Fig 5.2, it is shown that Level 4 members are the most generous population at the median score of 4.6. Within the **suspicious users** which are dominated by Level 4 users**, the median ratings reached 4.8** out of 5.0. In Fig 5.3, it is shown that a user account can reach Level 4 **solely by the amount of restaurant reviews without editor review**, meaning the comments are not verified by the third party. Considering **over half of the Level 4 reviewers will give a high rating of 4.8**, we recommend end users should **pay attention to the member level** of the review when referencing a restaurant review.

5.2 Word Cloud and Map of Restaurants from Suspicious Users

Figure 5.4 Word cloud of most frequently appeared words among all comment texts from Suspicious Users

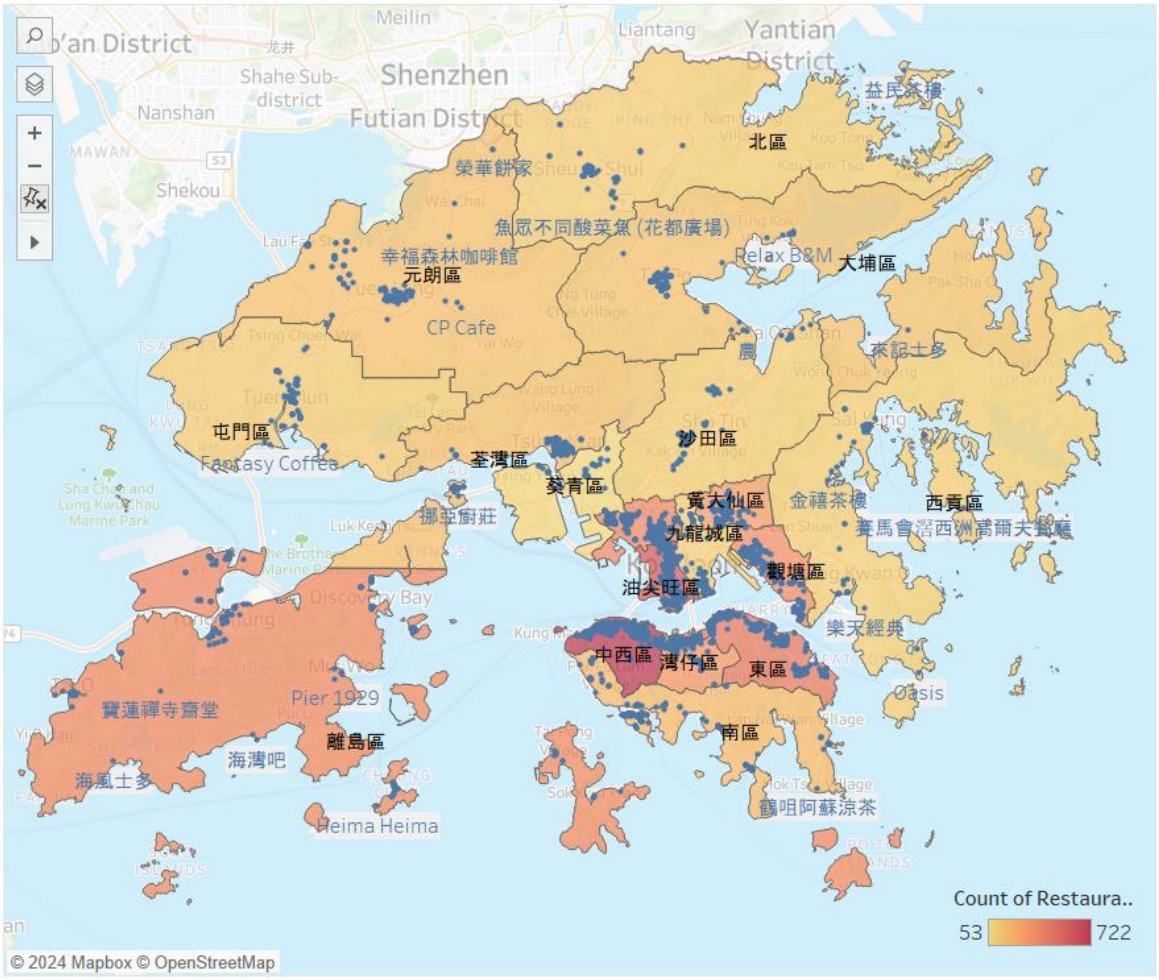

Figure 5.5 Map of Restaurants with Comments from Suspicious Users

Fig 5.4 and 5.5 displayed an overview of vocabularies and restaurants with comments from suspicious users.

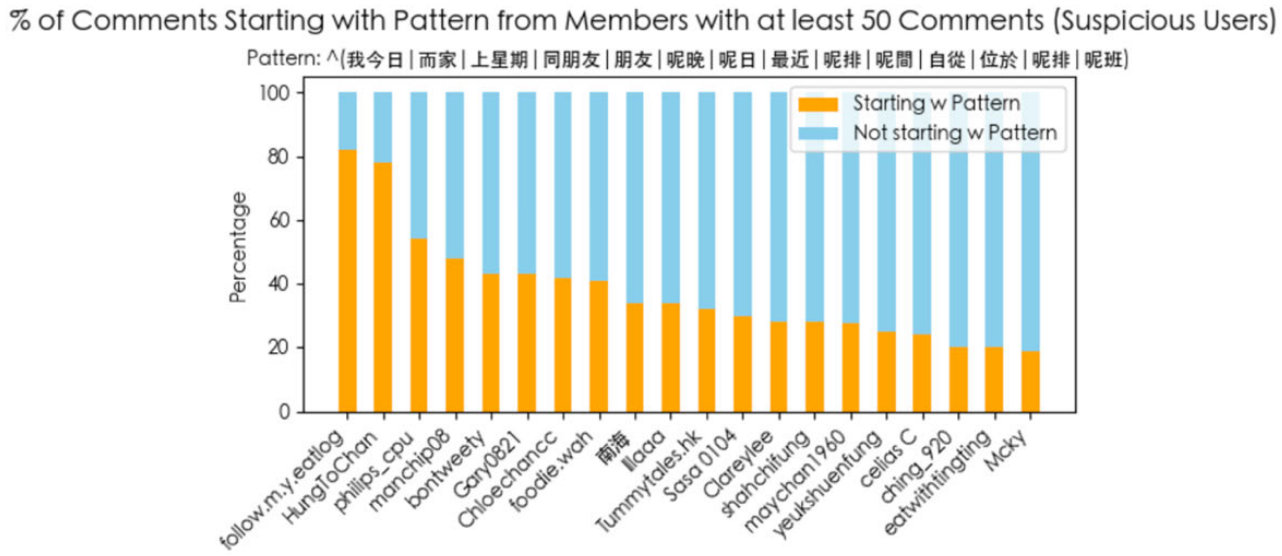

Figure 5.6 Percentage of Comments Starting with Pattern from Members with at least 50 Comments (Suspicious Users)

The posts of **suspicious users repeatedly use similar wordings as the beginning** of comment text. We picked the top 15 users with highest percentage of posts that start with our target vocabularies or phrases (Fig 5.6). Users with less than 50 comments are excluded from this filter. The list of words are chosen by observing the words commonly seen in the suspicious users' comment texts (Fig 5.7), and by prompting GenAI to create templates of food reviews. Fig 5.8 and 5.9 displays the respective frequencies of the words.

Within this group of users, 2-4 out of 5 comment texts begin with the selected words, which potentially suggests that **these words are the commonly used phrases in templates** for food reviews or comment texts generated by GenAI. Therefore, **checking presence for repeating patterns in the previous comments of the users is a possible solution** to identify paid reviewers.

	member_name	comment_text
7000	follow.m.y.eatlog	位於觀塘工廈裡的泰國菜餐廳，佈置花心思，有流水道，裝潢富有泰國色彩，予人舒適放鬆的感覺。餐廳...
7001	follow.m.y.eatlog	位於尖沙咀的船麵，餐廳不大，位置便利，人流蠻多，繁忙時間需要等位。主打泰國船麵，形式跟港式車...
7002	follow.m.y.eatlog	位於黃竹坑的多國菜西餐廳，自從南港島線開通後，通往南區方便多了，黃竹坑已經成為另一個工商廈區...
7003	follow.m.y.eatlog	位於觀塘工廈的台灣料理餐廳，裝修簡約，供應紅燒牛肉麵，滷肉飯、拌麵等，還有多款台式小食，如甜...
7004	follow.m.y.eatlog	位於香港仔熟食中心裡的一家港式食店，通常熟食中心裡面都是中式大排檔為主，而這家小店則有其他選...
7005	follow.m.y.eatlog	位於西環的菲律賓菜酒吧餐廳，餐廳樓底很高，空間感十足。全場燈光昏暗，播放著柔和的音樂，很有情...
7006	follow.m.y.eatlog	位於將軍澳坑口的泰國菜餐廳，裝潢富有泰國特色，播放泰文音樂；食物方面，提供各式泰式小食、沙律...
7007	follow.m.y.eatlog	位於香港仔的港式車仔麵店，地方新淨企理，燈光光猛，環境蠻舒適的，有別於舊式的麵店。在紙上選擇...
7008	follow.m.y.eatlog	位於銅鑼灣的台灣炸雞小食店，台灣過江龍，採用新鮮雞肉製作鮮嫩炸雞排。除了有不同口味的招牌炸雞...

Figure 5.7 Example of comment text from suspicious users

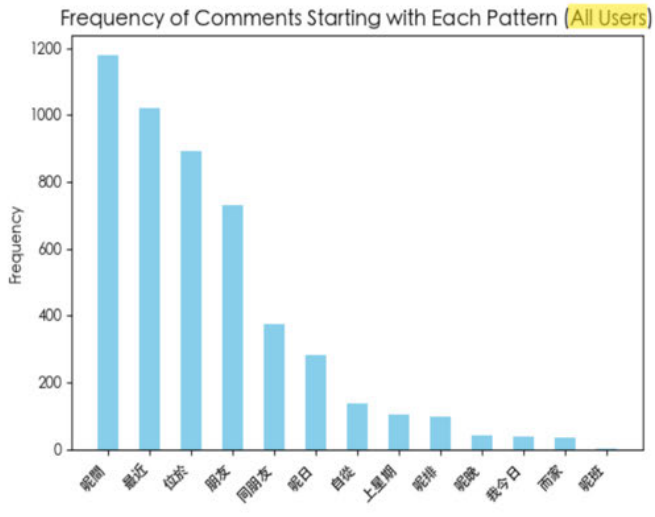

Figure 5.7
Frequency of Comments Starting with Each Pattern (All Users)

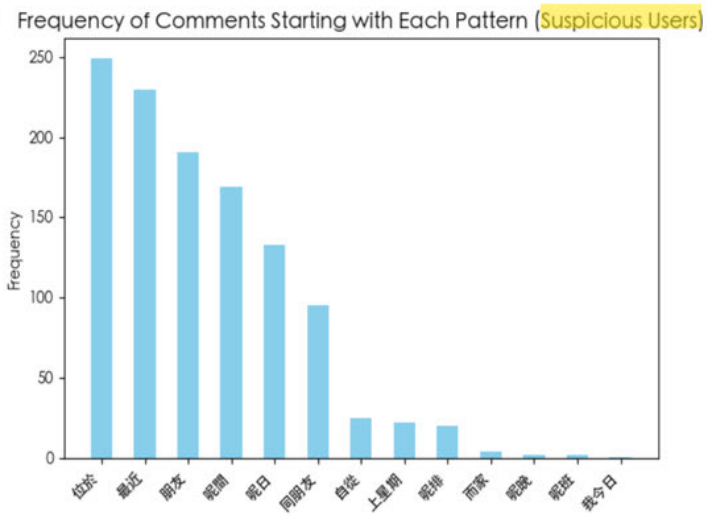

Figure 5.8
Frequency of Comments Starting with Each Pattern (Suspicious Users)

5.3 Analysis with Sentiment Score from Comment Text

5.3.1 Verification of accuracy of the SnowNLP Library

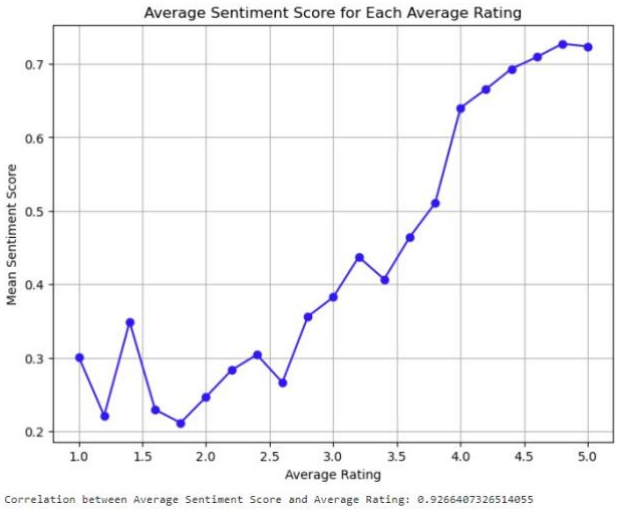

Figure 5.10 Average Sentiment Score for Each Average Rating

SnowNLP is a library that tokenize and return sentiment score after taking a string parameter. To verify the accuracy, we performed a Pearson's correlation with the sentiment score generated by SnowNLP and the average rating of that comment. For each comment rating, the mean of the corresponding sentiment score was calculated. **A strong positive correlation (r=0.926)** between sentiment score of a comment and average rating for the restaurant by the user was obtained, suggesting that the **sentiment model can reliably examine the positive or negative sentiment** in the comment.

5.3.2 Sentiment Analysis for suspicious and non-suspicious group

We seek to test if the difference in sentiment scores between the two groups are purely by chance. With the two large sample sizes, normal distribution of mean sentiment scores, and assumed equal variances, a two-sample Z-test is suitable for evaluating the significance of this difference.

$$z_0 = \frac{\bar{x}-\bar{y}}{\sqrt{(s_1^2)/n+(s_2^2)/m}}$$

Group	Number of Data	Mean	Sample SD
Suspicious	9451	0.61808169	0.474611
Non-Suspicious	38047	0.512249471	0.487435

We set the **difference** of mean sentiment scores between the **groups = 0 and > 0 as the null hypothesis (H_0) and alternative hypothesis (H_1)** respectively. By the Central Limit Theorem, the mean follows a normal distribution and Z-test is used to test the hypothesis. In the analysis, we found 17 sentiment scores are incomplete and are unsuitable for our analysis, and the no. of non-suspicious comments are reduced from 38064 to 38047. We obtained a **Z-score of 19.29 with a p-value < 0.0001**. Hence, we reject the null hypothesis **and conclude that suspicious group tends to comment restaurant positively than non-suspicious group.**

Fake Comment Proportion Analysis

- Number of Suspicious Comments: 9451
- Number of Non-Suspicious Comments: 38064
- Total Number of Comments: 47515

Fake Comment Proportion:

Interval Estimator of Binomial proportion p: $(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$

We obtained that **the 90% Confidence Interval of Fake Comment Proportion: (0.1959, 0.2019)**. On average, around **one out of five comments are potentially from paid reviewers**.

6. Limitations

6.1 Inaccessibility of data

New information is inaccessible at the later stage of scraping, attributable to the **IP block from OpenRice**.

6.2 Lack of chronological analysis

Our study only captured a snapshot in a single time point instead of **longitudinal data following same users and restaurant.** The current study did not explore neither the patterns of comment frequency and seasons or holidays, nor the no. of days since the start of business and the frequency of commenting.

6.3 Lack of sampling from all districts

One notable issue in our dataset is the lack of representation from all districts, with the lack of restaurants from the southern district included in our sampling. This limitation raises concerns about the comprehensiveness and representativeness of our data.

7. Discussion

7.1 Further inferential statistical analysis

After z-test between sentiment score between suspicious and non-suspicious users, it is shown that there is statistical difference in their comment texts. Further analysis can be performed such as **one-way ANOVA test** to investigate the **independence between levels of comment account and average ratings**.

7.2 Explore the patterns of "Paid Negative Reviews"

Beyond complementary reviews, the **criticizing reviews on competitors** are worthy to investigate.

7.3 Discover the finding's generalizability to other retail comment platform

Additionally, we suggest that the findings in this study could be evaluated on other retail review platforms such as Price.com and Cosme to examine **the generalizability to other online retail platforms**.

8. Conclusion

In the first five pages that an end user will access, a **certain portion of reviews are potentially from paid reviewers** and end users should be cautious about the validity of reviews. **Level of user account** can be **a potential trait** to spot suspicious paid reviewers. To screen for fake reviews, end users can pay attention to the **wordings in the beginning of the comment texts** and check if the previous comment texts from that user have displayed **repeated patterns**.

9. Appendix

See the attached Appendix and Jupyter Notebooks for other exploratory results.