Exploring Foodpanda consumer reviews using Part of speech and N-gram

Teoh Sin Yee
School of Computer Sciences USM
USM, Penang
sinyee00@student.usm.my

Wan Muhamad Rusyaidi Afifi
School of Computer Sciences USM
USM, Penang
rusyaidiafifi10@student.usm.my

***Abstract*— Foodpanda is a global online food delivery marketplace headquartered in Berlin, Germany. It operates in more than 40 countries across 5 continents including Malaysia. Due to the pandemic outbreak, the usage of Foodpanda apps has rapidly increased. This study aims to sentiment analysis on Foodpanda consumer reviews. A total of 10,000 consumer reviews (June 1st, 2021 to November 26th, 2021) are extracted from Google Play Store. This study applies Natural Language Processing as pre-processing of text in sentiment analysis. N-gram and Part of speech tagging are used to further analyse the reviews. Sentiment analysis on Foodpanda consumer reviews is useful as an opportunity for application improvement, on the other side is also used as a reference for other users before using Foodpanda service. The results provide insights behind the most critical and most positive consumer reviews of Foodpanda.**

Keywords— sentiment analysis, NLP, Part of speech tagging, N-gram

I. INTRODUCTION

COVID-19 drove consumers to dine at home and accelerated the adoption of delivery apps and other technologies, but the trend will likely outlast the virus. The lockdown has affected all forms of business, including the catering and restaurant industry. There have been strict prohibitions on dining out that have resulted in the restaurants losing their business.

Foodpanda as an online-to-offline mobile service has recently gained popularity offering two-way benefits for catering enterprises and customers by providing convenient and efficient online order and offline delivery services. Foodpanda has 674,620 monthly app downloads, according to Apptopia. [1] Statista Reports illustrated that 81 percent of respondents in Malaysia stated that Foodpanda was the food delivery app they used the most. [2]

Foodpanda has raised $318 million of venture capital. Nearly $20 million was raised in initial funding from Rocket Internet and investment AB Kinnevik 2013. During the same year, iMENA Holdings invested approximately $8 million and received another $ 20 million in the year 2014. Goldman Sachs also invested nearly $100 million. All the investments really show confidence from investors and VCs about Foodpanda's success and growth. [3]

In addition, consumer satisfaction is vital in a competitive business environment such as in food delivery business. Satisfaction and consumer experience have an important role in online food delivery services. Currently, these few years are the years we can see Foodpanda growing into a bigger platform which not only delivers food but also groceries and let users pre-order for self-pick-up.

Besides, a Twitter campaign to boycott Foodpanda was held and Foodpanda Thailand reportedly loses over 2 million users after the social media debacle. [4] The campaigns were going on in multiple countries including Thailand, Malaysia, Taiwan. [5] [6] This leads us to investigate why consumers choose to use Foodpanda over other food delivery platforms and why some consumers boycott Foodpanda.

This study uses 10,000 Foodpanda consumer reviews (June 1st, 2021 to November 26th, 2021) extracted from Google Play Store. Furthermore, the dataset is pre-processed with Natural Language Processing (NLP) to improve the dataset. After that, the dataset is used to build a Machine Learning model and further used for Part of speech tagging and N-gram to gain more insights about consumer reviews.

II. RELATED WORK

Integrating N-gram in text exploration analysis has proven to be more effective [7], [8]. Silva J took the approach of classifying users reviews from multiple websites using N-gram. In their work, the authors collected 4520 laptop reviews and 7254 restaurant reviews to be added to their respective domains' libraries. The authors then represent a model where it extracts the sentiment by calculating the frequency of appearance of the textual unit in the sentence. The results found that the proposed method works well in determining the polarity of words based on the domain [7]. Liu J in their works [8] proposed an approach that incorporates N-gram based features into a feature set. The authors found that introducing domain-specific knowledge into textual features extraction can significantly increase classification performance.

For user textual reviews, research using Part of speech (POS) tag embedding techniques was conducted by Da'u A [9]. They applied POS tags features in addition to word embedding features to increase model performance. They used both general and domain-specific embedding to acquire the word's syntactic and semantic information, and POS tag embedding is used to improve the sequential labeling of the aspects [9]. In addition, the work of Hu M [10] and Popescu A [11] uses the POS tagging approach to gather nouns and noun phrases because features mostly consist of nouns. In this case, the authors generated POS tags on every word, regardless of whether it was a noun or a verb. The common feature item sets are then filtered out via association rule mining. As a result, the model was able to perform well when analyzing electronic products.

Our approach has some similarities with Malik H's work [12]. The author classified people's opinions into three categories which are positive, negative, and neutral for sentiment analysis. The stopword removal and lemmatization process were also conducted since they act as noise and do not contribute towards sentiment analysis. By only tagging the filtered sentences, people's emotions can be classified

into more types such as disgust, anger, fear, etc. Each of these types helps the algorithm to generate a more sophisticated analysis [12].

III. METHODOLOGY

A. Data Collections

To investigate consumer reviews on the Foodpanda application, we rely on user reviews from Google Play Store. Foodpanda app reviews on the Google Play store consist of over 2 million user reviews that concern different aspects of the application. We extracted raw data to retrieve all the available attributes from the user reviews such as username, rating, comment, and date of each review. We then continued to select only the latest 10 000 user reviews to ensure that the user review are still relevant. All the reviews selected were written in the English language. The reviews that consist of 4- or 5-stars ratings are considered positive while below 3-star ratings are considered negative. All reviews that were rated as 3 stars rating were removed as it represents a neutral review.

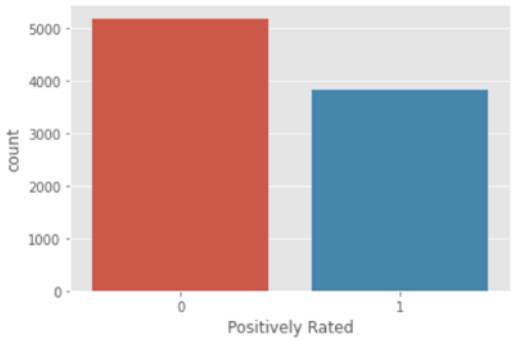
B. Data Preprocessing

Pre-processing is intended to eliminate noise in the dataset. Natural Language Toolkit (NLTK) is used to discard all non-necessary words by four steps: 1) lower case of all words, 2) delete all the punctuations, 3) tokenize the words, and 4) delete all the stopwords such as “the”, “a”, “to”, etc.

C. Features

A machine learning model is built to predict the variable of interest. Part of speech tagging is being done with the help of Spacy model. For each review text, we have features for counts of the number of adjectives, nouns, and verbs. N-gram (n=1,2,3) features are counted and categorized into positive and negative N-gram.

D. Visualization



There are 57.5 % of reviews are categorized as negative and 42.5 are % positive.

IV. RESULTS AND DISCUSSIONS

A. N-gram

After the NLP process for the dataset was completed, obtained top 10 most common appeared words in positive and negative reviews as shown in Table I.

TABLE I. TOP 10 MOST COMMON APPEARED WORDS FOR N-GRAM

#	Top 10 most common appeared words for n-gram					
	unigram_ pos	unigram_ neg	bigram_ pos	bigram_ neg	trigram_ pos	trigram_ neg
1	easy	awful	easy use	waste time	easy order food	worst app ever
2	excellent	poor	fast delivery	cancelled order	thank food panda	worst food delivery
3	nice	sucks	easy order	worst app	keep good work	bad customer service
4	great	disappointing	great app	poor app	app easy use	poor customer service
5	best	expect	nice app	useless app	best food delivery	worst customer service
6	love	hate	best food	unknown error	easy use app	get money back
7	amazing	uninstall	useful app	bad service	overall good experience	order got cancelled
8	fast	disappointed	great service	items ordered	thanks food panda	better use grab
9	improved	lousy	good service	food price	best food app	says unknown error
10	popping	worse	great experience	pick order	really good app	estimated delivery time

Based on the result, we found that unigram can represent the consumers’ feelings. Bigram and trigram are equally good at representing the causes or factors that contribute to consumers’ feelings. The results did not tell the complete story and a further investigation was carried out by scanning through complete review text.

Referring to the positive unigrams, the consumers used ‘easy’ to express that the Foodpanda app is user-friendly and easy to use. We can identify the loyal Foodpanda users through the review because these users’ reviews are complimenting the Foodpanda app has greatly improved over time going. It is a good sign where the ‘improved’ has ranked number 9 among all positive reviews and reflected consumers’ satisfaction.

Word ‘improved’ is being used in sentences including: ‘They have improved over the past two years. Satisfactory experience.’ and ‘Changing my rating to 5 stars as their service has greatly improved.’

To find out the reasons that contributed to consumers’ negative reviews, negative bigram and trigram depicted the part of the stories. Words such as ‘bad customer service’, ‘poor customer service’ and ‘worst customer service’ came three in a row within negative trigram reflected consumers' dissatisfaction towards Foodpanda’s customer service. The consumers are frustrated where the customer service is not efficient in helping them to handle any glitch in the fastest manner.

Word ‘unknown error’ ranked 6th in bigram and 7th in trigram. This reflected a very high potential risk to Foodpanda as it is a business-critical application. If a business-critical application fails or is interrupted, normal operations of the organization cannot proceed as usual. This can lead to short and long-term financial losses, decreased

productivity, loss of brand authority, and loss of customer trust. [13]

By discovering the full review text, we found that some consumers failed to log in to the Foodpanda app and the issue persisted after re-installed the app. Some consumers failed to submit their food orders due to unknown errors.

The 9th negative bigram, 'food price' shown some consumers felt the burden of the expensive food price being listed in the Foodpanda app. The consumers commented that the food price in the app is 2 – 3 times higher than the original food price in the restaurant. The price issue has been addressed in news as well where the restaurants claimed that they are regretted to set a higher mark-up price. The restaurants found it hard to earn a profit if they maintained the same original food price while paying service charges to Foodpanda. [14]

Another interesting insight we can get from the 8th negative trigram, 'better use grab' shows the consumers have lost confidence in Foodpanda and potentially will use the competitor 'Grab' app instead. The app will display the estimated delivery time after ordering successfully. Consumers are expected to get their food on time. However, the consumers are dissatisfied where the estimated delivery time kept changing from time to time especially the delays in delivery.

Improvements

We found there is overlapping of 4th and 7th in negative unigram ('disappointing', 'disappointed') and can be improved by stemming the words. Words 'great app', 'nice app', 'useful app' are close to synonyms. It will be better to cluster the synonyms and use 'great app' to represent all its synonyms. This can be done with the help of NTLK WordNet. [15] With this, we can get more meaningful insights. The same approach can be applied to the set of bigrams: ('bad customer service', 'poor customer service', 'worst customer service'), ('worst app', 'poor app', 'useless app') and ('thank food panda', 'thanks food panda').

B. Part of speech

After processing the data, the 10 most common appeared words in positive and negative reviews based on three features such as adjectives, verbs, and nouns are shown in Table II.

TABLE II. TOP 10 MOST COMMON WORDS FOR POS

	Top 10 most common appeared words for POS					
#	**adjectives_pos**	**adjectives_neg**	**verb_pos**	**verb_neg**	**noun_pos**	**noun_neg**
1	good	bad	use	use	food	order
2	easy	good	order	get	delivery	app
3	great	many	give	cancel	app	food
4	fast	wrong	get	say	order	delivery
5	convenient	poor	make	order	service	time
6	nice	available	deliver	give	time	service
7	many	high	need	deliver	use	customer
8	friendly	online	keep	make	restaurant	restaurant
9	helpful	expensive	go	take	rider	rider
10	available	late	add	try	customer	payment

By scanning through the result, we found that adjectives can represent the consumers' feeling which has a similar function to unigram. The consumers do not only express holistic opinions but often focus on specific features of their interest and the noun comes to play for feature-based sentiment analysis. [16]

Nouns such as order, app, food, delivery, service show different aspects of consumer reviews. Word 'order' is being used most frequently in negative review as noun and verb. 'order' is being used in sentences including: '1 hour and 30mins still counting for my order to be delivered.', 'Inaccurate delivery time, orders are always delayed', 'System suddenly cancels order and refund take 14 working days.'

Besides, the noun 'order' nonetheless ranked 4th in positive reviews. It is being used in sentences including: 'Seamless ordering. I like the options for instructions that you don't always get from the website.', and 'good deal, fast delivery, good service and compensate when there's a mistake in the order'.

From the aspect of order, we understand that the consumers were disappointed by last-minute order cancellations and late delivery. There are many consumers who failed to complete their orders and the consumers were frustrated at the assistance provided by customer service is helpful. Regardless of some dissatisfactions, there are some consumers satisfied with Foodpanda's mechanisms to compensate consumers when Foodpanda mishandled the order. From the aspect of delivery, consumers also complained about the expensive delivery fee and fluctuation of delivery fee during peak hours. Consumers also found the estimated delivery time is not reliable since the actual delivery time is 2 – 3 times the delay of the initial estimated delivery time.

Improvements

Noun can be used as a feature and combinations of nouns and adjective with the help of N-gram can bring deeper insights. For instance, 'expensive food' and 'fast delivery'. Besides, the bigram noun is more suitable used as a feature such as 'customer service'. In this case, since the consumer reviews are focused on 'customer service', using 'service' alone might failed to reflect the real situation that customers faced.

V. CONCLUSION

This paper attempts to highlight the most positive and most critical Foodpanda consumer reviews using N-gram and Part of speech. Both ends of reviews are useful to the company to leverage the service quality and outstand among its competitors.

The present study has some limitations as mentioned below:

- Different reviews contain symbols like (😍, 😡, ☹) which help in presenting the sentiment, but these are not taken into consideration in this study for analysis.
- To give stress on words, some reviews are containing upper case letters and punctuations like (WORST APP ever!). It expressed the utter anger of consumers, but this aspect is not considered in this paper.

- Some reviews are less useful like ('Not optimized .. Give me access I am a developer I will fix it..' and 'very slow connectivity'). These reviews did not point out the issues they faced directly and it could be an invalid issue. The helpfulness of the reviews can be validated by the other users on Google Play Store. Extracted the reviews along with their helpfulness can improve our analysis.
- Some reviews are mixing multiple languages. Further study in Cross-lingual information retrieval can improve our analysis. [17]
- Some reviews have spelling errors. Spelling correction is not considered in our data pre-processing and will extend it in future works.

All the above-mentioned limitations may be considered for future work, to improve the quality of sentiment analysis.

REFERENCES

[1] "foodpanda - Food Delivery - app store revenue, download estimates, usage estimates and SDK data | Apptopia," Apptopia, [Online]. Available: https://apptopia.com/ios/app/758103884/about.

[2] "Malaysia: favorite food delivery apps 2021," Statista, 30 November 2021. [Online]. Available: https://www.statista.com/statistics/1149404/malaysia-favorite-food-delivery-apps/.

[3] A. S. Suleiman, D. M. Kee, A. M. Azmi, D. W. Chan, J. H. Aw, W. R. Khanum and A. G. Utama, "The performance of Foodpanda during the pandemic: A study of consumers' perspective," *Journal of The Community Development in Asia,* vol. 4, no. 3, pp. 36 - 48, 2021.

[4] "Tech in Asia - Connecting Asia's startup ecosystem," [Online]. Available: https://www.techinasia.com/foodpanda-thailand-reportedly-loses-2m-user-accounts-after-social-media-outburst.

[5] "Vendors on Foodpanda protest 'hidden costs'," [Online]. Available: https://themalaysianreserve.com/2021/11/09/vendors-on-foodpanda-protest-hidden-costs/.

[6] "Foodpanda Taiwan gets NT$2 million fine for imposing restrictions on restaurants: Taiwan News," 17 September 2021. [Online]. Available: https://www.taiwannews.com.tw/en/news/4290076.

[7] J. V. J. e. a. Silva, "Algorithm for Detecting Opinion Polarity in Laptop and Restaurant Domains," in *Procedia Computer Science*, 2020.

[8] J. ,. P. Liu, "Automatic Identification of Messages Related to Adverse Drug Reactions from Online User Reviews using Feature-based Classification," Iranian J Publ Health, 2014.

[9] A. S. N. Da'u, "Aspect extraction on user textual reviews using multi-channel convolutional neural network," *PeerJ Computer Science,* vol. 2019, no. 5, 2019.

[10] M. L. B. Hu, "Mining Opinion Features in Customer Reviews," in *ACM Digital Library*, 2004.

[11] O. E. Ana-Maria Popescu, "Extracting product features and opinions from reviews," in *ACM Digital Library*, 2005.

[12] H. S. E. M. Malik, "Mining Collective Opinions for Comparison of Mobile Apps," in *Procedia Computer Science*, 2016.

[13] C. Insights, "Business Critical Applications: An In-Depth Look," [Online]. Available: https://cloud.netapp.com/blog/azure-anf-blg-business-critical-applications-an-in-depth-look.

[14] "Restaurants urge customers to bypass food delivery platforms charging exorbitant fees," The Star, 28 June 2021. [Online]. Available: https://www.thestar.com.my/tech/tech-news/2021/06/23/restaurants-urge-customers-to-bypass-food-delivery-platforms-charging-exorbitant-fees.

[15] "Wordnet with NLTK," Python Programming Tutorials, [Online]. Available: https://pythonprogramming.net/wordnet-nltk-tutorial/.

[16] M. A. M. A. T. I. Cataldi, "Good location, terrible food: detecting feature sentiment in user-generated reviews," *Social Network Analysis and Mining,* vol. 3, no. 4, pp. 1149-1163, 2013.

[17] "A Brief Introduction to Cross-Lingual Information Retrieval," Medium, 07 March 2019. [Online]. Available: https://medium.com/lily-lab/a-brief-introduction-to-cross-lingual-information-retrieval-eba767fa9af6.

Appendix

Python scripts for generating N-gram
https://colab.research.google.com/drive/1XnEegDayXhEYlE7C_FAD4qKs2e5YqOT8?usp=sharing

Python scripts for Part of speech tagging
https://colab.research.google.com/drive/1cu_4uvymSEpKTJCMP5LwOz-Ht5QHXExm?usp=sharing

Raw data (Extracted from Google Play Store)
https://docs.google.com/spreadsheets/d/1jRuNnJBVs9UKpzDY9iwJxvmeZGDeHnHhRs5a3iZsfRY/edit?usp=sharing

Pre-processed Positive and Negative reviews
https://docs.google.com/spreadsheets/d/1CAo1BTTfGkDU-iqe0mYpqDbIQhaN6a2-Rmitu7gf73A/edit?usp=sharing

Aspect and opinion word extraction on Foodpanda consumer reviews using Regular Expression

Teoh Sin Yee
School of Computer Sciences USM
USM, Penang
sinyee00@student.usm.my

Mohamed Abdelnasser Mohamed Elsayed Hassan
School of Computer Sciences USM
USM, Penang
mabdelnasser@student.usm.my

***Abstract* — This paper proposes a new method for aspect and opinion word extraction from Foodpanda consumer reviews using Regular Expressions (RegEx) in combination with Part-of-Speech (POS) tags. The proposed method is a seven-steps process that can identify key aspects of Foodpanda's services and extract the opinions of customers about those aspects. The findings of the study show that the proposed method can be used to improve Foodpanda's services and better understand customer needs. The proposed method has several advantages over other sentiment analysis methods. It is able to identify sentiment-laden phrases in the reviews that other methods may miss, extract actionable feedback from the reviews, and is relatively easy to implement. The proposed method can be used by businesses to improve their products and services.**

Keywords— Sentiment Analysis, N-gram, Part of Speech (PoS) Tagging, Regular Expressions, Aspect Extraction, Opinion Extraction, Text Normalisation

I. INTRODUCTION

Foodpanda, an app-based food delivery service, is becoming more and more popular. It helps both restaurants and customers by making it easy to place orders online and get food delivered. Lately, a lot of people around the world have started using it daily. Particularly in Malaysia, a study found that 81 percent of people prefer Foodpanda for food delivery. Another evidence of its popularity is the huge number of monthly app downloads in 2023, which stands at 959,534 according to Apptopia. [1]

Foodpanda has also managed to attract a lot of investment. In 2013, it got funding of nearly $20 million from Rocket Internet and investment AB Kinnevik. In the same year, iMENA Holdings put in about $8 million, and then added another $20 million in 2014. Later, Goldman Sachs invested nearly $100 million. This shows that investors have a lot of confidence in Foodpanda's success.

A. Motivation

The food delivery industry is highly competitive, with customer feelings and experiences becoming very important for improving and growing services. Companies like Foodpanda have not only been delivering food but also have started delivering groceries and offering pre-order pick-up options. Understanding what customers think about these various services is key to making them better and helping the company grow. In addition, there have been cases where users left Foodpanda because they were unhappy, showing the need for a clear understanding of customer feelings. A past study tried to show these feelings through sentiment analysis of user reviews. However, the methods used had trouble separating reviews that provide useful feedback from those that don't. The current study aims to improve the understanding of user feelings by improving the sentiment analysis method, offering a clearer and more informative view of customer reviews.

B. Dataset description

This study uses a set of 10,000 reviews from Foodpanda customers. These reviews, collected from the Google Play Store, cover the period from June 1st, 2021, to November 26th, 2021. This large collection of reviews offers a deep look into how users feel about Foodpanda and its services.

C. Problem

The earlier study [2] experienced difficulties when using methods such as Part of Speech (PoS) tagging and N-grams in analysing Foodpanda reviews. They had trouble pinpointing the aspect, such as 'customer service' in expressions like 'poor customer service', leading to incomplete sentiment analysis. Additionally, traditional approaches may miss subtle aspects of language in reviews, such as sarcasm or indirect criticism. To address these issues, this study introduces the use of Regular Expressions in conjunction with Part of Speech (PoS) tagging. This combination aims to filter out non-opinionated reviews, improving the precision of sentiment analysis and providing a clearer understanding of consumers' sentiment.

II. BACKGROUND

A. Basic concept of regular expressions

Regular Expressions (RegEx) are sequences of characters forming a pattern, used to find a specific string or group of strings within a larger text. They play a vital role in computer science, particularly in text processing, where they serve as a key tool for searching, matching, replacing, and splitting text based on specific conditions. RegEx patterns can range from simple, such as finding a single word in a document, to complex, like validating the format of an email address or identifying sentences containing particular words or phrases. They follow a specific syntax that allows for the definition of highly precise text patterns. This precision makes RegEx an incredibly versatile tool for a broad array of text manipulation tasks.

In the context of sentiment analysis, RegEx can significantly enhance the quality and accuracy of the results. By defining precise patterns, RegEx can help identify specific sentiment-laden phrases, isolate particular aspects of a service that users mention in their reviews, and filter out unhelpful or non-opinionated reviews.

Considering these benefits, this study intends to employ RegEx in improving the sentiment analysis method used in the previous study. It aims to use RegEx for filtering out non-opinionated reviews and emphasizing those with actionable feedback, ultimately providing a clearer, more detailed, and

informative view of customer reviews and their sentiments towards Foodpanda's services.

B. Related works

Aspect extraction, a crucial part of sentiment analysis, has been tackled using both supervised and unsupervised learning methods. In supervised approaches, Gojali et al. [3] utilized a range of features like sentence position, pronouns, adjectives, numbers, capitals, and adverbs, in binary format. This method achieved a precision of 71.04, a recall of 74.55, and an f-measure of 72.75 in sentiment classification. Meanwhile, Ekawati et al. [4] developed feature sets such as Bag of N-grams, Bag of Head Words, Bag of Clusters, and Bag of k-skip-bigram along with their combinations. Using the CRF algorithm for aspect extraction, this approach resulted in a precision of 84.9, a recall of 74.7, and an f-measure of 79.

On the other hand, unsupervised approaches typically involve matching or automatically searching the list of aspects to extract the aspect in a sentence. Akhtar et al. [5] implemented this by matching words in opinion sentences with a list of aspects based on similarity value. Li et al. [6] automatically extracted aspects by identifying nouns or noun phrases as potential aspects and pruning to extract valid aspects. However, in their study, aspect extraction preceded the extraction of the opinion word.

Several studies focused on obtaining the relationship between the word aspect and the opinion words. Lazhar et al. [7] used dependent parsing for this purpose. Gojali et al. [3] relied on the distance method, allowing a specific distance from the aspect word to obtain the opinion words. This approach sometimes resulted in errors in determining the aspect-opinion word pair correctly. Rana et al. [8] implemented the two-fold rule-based model (TF-RBM) method for aspect extraction, which uses rules derived from user reviews, achieving a precision of 87, recall of 92, and f-measure of 89.

In contrast to the above studies, the current investigation aims to improve the understanding of customer feelings about Foodpanda's various services. The method planned for this study involves using Regular Expressions for refining the sentiment analysis technique. This improved method will focus on filtering out non-opinionated reviews and highlighting those with actionable feedback. Therefore, offering a clearer and more detailed understanding of customer reviews and their sentiments towards Foodpanda's services.

III. PROPOSED SOLUTION

The proposed solution can be divided into seven parts which are:

A. Data collection

Consumer reviews on the Foodpanda application are investigated by relying on user reviews from Google Play Store. The Foodpanda app reviews on the Google Play store are comprised of over 2 million user reviews that pertain to various aspects of the application. The raw data was extracted in order to retrieve all the available attributes from the user reviews, including username, rating, comment, and date of each review. Furthermore, only the latest 10,000 user reviews were selected to ensure the continued relevance of the user review. It should be noted that all the selected reviews were written in the English language.

B. Data preprocessing

Pre-processing is intended to eliminate noise in the dataset. Natural Language Toolkit (NLTK) is used to discard all non-necessary words by three steps: 1) lower case of all words, 2) delete all the punctuations, 3) lemmatize the words.

The reason to lemmatize the word is to avoid duplicates when extracting aspects or opinions. For example, customer services and customer service.

C. Data visualisation

Among 10,000 reviews, the average length of positive and negative reviews varies significantly. Negative reviews have a maximum word count of 304, which is more than double the maximum word count of 116 for positive reviews. This suggests that consumers express themselves more extensively when they have a negative experience.

TABLE I. STATISTICS ON LENGTH OF REVIEWS

	Overall	Positive	Negative
Min	1	2	1
Max	304	116	304
Average	30	20	36

The longest positive and negative reviews, along with their summaries, are tabulated in the Appendix.

D. Data normalisation

The first step involves the examination of the top 10 most commonly occurring nouns, as these nouns will be considered as aspects. The presence of both "customer" and "service" in the list is displayed in Table II. It was observed that certain nouns consist of two words, such as "customer service," while others like "rider" and "restaurant" naturally exist as standalone words. To tackle this, the top 10 words occurring before and after each noun are analysed.

For instance, the word "customer" is often followed by words like "service," "support," and "care." These words share the same meaning but have different representations. Therefore, phrases like "customer support" and "customer care" are replaced with "customer service."

The same analysis is applied to multiple nouns, as shown in Table III and Table IV.

Fig. 1. Word cloud before performing data normalisation

TABLE II. TOP 10 MOST COMMON NOUNS

#	Noun	Frequency
1	app	4368
2	order	3623
3	food	2452
4	delivery	2300
5	service	2078
6	time	1981
7	rider	1334
8	customer	1331
9	restaurant	1253
10	payment	788

TABLE III. TOP 5 WORDS OCCURRING BEFORE AND AFTER 'CUSTOMER'

#	Word Before	Occurrence %	Word After	Occurrence %
1	the	15.10%	**service**	48.21%
2	their	6.98%	**support**	8.93%
3	bad	6.41%	and	2.68%
4	to	3.90%	to	2.44%
5	your	3.90%	**care**	1.95%

TABLE IV. NOUNS THAT INVOLVED IN NORMALISATION

#	Nouns	Replaced with
1	customer support	customer service
2	customer care	
3	delivery guy	rider
4	delivery man	
5	delivery person	
6	delivery charge	delivery fee
7	foodpanda	FP
8	food panda	
9	panda	
10	payment option	payment method

TABLE V. TOP 10 MOST COMMON NOUNS AFTER NORMALISATION (BIGRAM ONLY)

#	Noun	Frequency
1	customer_service	707
2	delivery_fee	460
3	food_delivery	283
4	delivery_time	141
5	credit_card	106
6	delivery_service	141
7	phone_number	87
8	app_update	83
9	payment_method	80
10	experience_app	77

E. POS Tagging

In the POS Tagging stage, word tags/labels are assigned to identify the parts of speech in a text. For this study, 10 word tags are utilized, as listed in Table VI.

TABLE VI. LIST OF WORD TAGS INVOLVED IN THIS STUDY

#	Tag	Description	Examples
1	ADJ	Adjectives	fast, good
2	ADP	Adpositions	on, in, for
3	ADV	Adverb	very, too
4	AUX	Auxiliary verbs	be, have, do
5	CCONJ	Coordinate Conjunctions	and, or
6	DET	Determiners	a, an, the
7	PROPN	Proper Noun	Malaysia
8	PART	Particles	"to" in "to go"
9	NOUN	Noun	customer, rider
10	VERB	Verb	pick, eat

F. Selection of Opinion Sentence from User Review

User reviews typically contain multiple sentences, but not all sentences qualify as opinion sentences. Opinion sentences are those that contain both aspect and opinion words. In this study, the focus is on extracting opinion sentences to obtain valid pairs of aspect and opinion words.

Each sentence undergoes POS tagging and is selected using regular expressions to determine if it qualifies as an aspect-based opinion sentence.

In an opinion sentence, important tags for aspect candidates include nouns (NOUN) and proper nouns (PROPN). Opinion words are typically in the form of adjectives (ADJ). Supporting opinion words in an opinion sentence can be particles (PART), adverbs (ADV), auxiliary verbs (AUX), coordinating conjunctions (CCONJ), adpositions (ADP), and determiners (DET).

Non-opinion sentences are eliminated from the study as they do not contribute much useful information during sentiment analysis. For example, the phrase "fast and quality" is

unclear, as it is unclear what is being referred to as fast and what is meant by quality.

Another example is the following lengthy sentence: "*Haish, foodpanda I already give you 5 stars but now I need to give you 1 star back. I have waited for 2 hours for my order and I want to cancel the order because the order takes too much time, then the help centre is broken, then the order got cancelled automatically, luckily the payment didn't get charged yet by my credit card. I am very disappointed because I waited too long to eat but hmmm.*" This example reflects a challenge in sentiment analysis where reviewers tend to express their emotions in lengthy reviews rather than being concise and to the point. The first sentence in this example is not useful. Therefore, this study will only process sentences that have less than 10 words.

After screening thousands of rows of user reviews, opinion sentences are categorised into three types based on common patterns. In the first type, there is one or more aspects (NN/PROPN) followed by one or more opinion words (ADJ). For example, the sentence "delivery fee unreasonable at all" has one aspect (delivery fee) and one opinion word (unreasonable).

In the second type, the sentence begins with a supporting opinion word (ADV), followed by an opinion word (ADJ), and an aspect (NN/PROPN). For example, the sentence "very fast delivery and respectful rider" has two sets of aspects (delivery, rider) and two sets of opinion words (fast, respectful).

In the third type, the opinion word (ADJ) is followed by the aspect (NN/PROPN). For example, the sentence "unresolved cash payment issue" has the opinion word "unresolved" and the aspect is "cash payment issue".

Table VI provides regular expressions and rule-based aspects obtained from the regular expressions, along with examples of opinion sentences.

Rule 1: One or more Adjectives followed by one of more Noun or Proper Noun, then followed by zero of more Coordinate Conjunctions/ Adjectives/ Noun/ Proper Noun/ Adverb

Regular Expression of Rule 1:
(ADJ+ (NOUN|PROPN)+)((CCONJ|ADJ|NOUN|PROPN|ADV))*

TABLE VII. EXAMPLE OF OPINION SENTENCE AND CORRESPONDING EXTRACTED ASPECT AND OPINION (RULE 1)

#	Opinion Sentence	Aspect	Opinion
1	unresolved cash payment issue ADJ NOUN NOUN NOUN	cash payment issue	unresolved
2	nice app good packing fast delivery ADJ NOUN ADJ NOUN ADJ NOUN	app packing delivery	nice good fast
3	nice app and good customer service ADJ NOUN CCONJ ADJ NOUN NOUN	app customer service	nice good
4	easy ui and low delivery fee ADJ PROPN CCONJ ADJ NOUN NOUN	ui delivery fee	easy low
5	pathetic service always late and wrong order ADJ NOUN ADV ADJ CCONJ ADJ NOUN	service order order	pathetic late wrong

Rule 2: One or more Adverb followed by one or more Adjectives, followed by zero or more Adpositions/ Determiners, then followed by one or more Noun/ Proper Noun, then followed by zero of more Coordinate Conjunctions/ Adjectives/ Noun/ Proper Noun/ Verb/ Adpositions/ Adverb

Regular Expression of Rule 2:
(ADV+ ADJ+)((ADP|DET))* (NOUN|PROPN)+ ((CCONJ|ADJ|NOUN|PROPN|VERB|ADP|ADV))*

TABLE VIII. EXAMPLE OF OPINION SENTENCE AND CORRESPONDING EXTRACTED ASPECT AND OPINION (RULE 2)

#	Opinion Sentence	Aspect	Opinion
1	very fast delivery and respectful rider	delivery	fast

	ADV	ADJ	NOUN	CCONJ	ADJ	NOUN				rider	respectful
2	too	high	of	a	delivery	fee				delivery fee	high
	ADV	ADJ	ADP	DET	NOUN	NOUN					
3	very	high	delivery	fee	and	pathetic	service			delivery fee	high
	ADV	ADJ	NOUN	NOUN	CCONJ	ADJ	NOUN			service	pathetic
4	very	lousy	customer	service	and	pick	up	policy		customer service	lousy
	ADV	ADJ	NOUN	NOUN	CCONJ	VERB	ADP	NOUN		policy	lousy
5	very	cheap	delivery	fee	and	more	promotion			delivery fee	cheap
	ADV	ADJ	NOUN	NOUN	CCONJ	ADV	NOUN			promotion	cheap

Rule 3: One or more Noun/ Proper Noun, followed by zero or more Verb/ Auxiliary Verb/ Adverb/ Particles/ Coordinate Conjunctions, followed by one or more Adjectives, followed by zero or more Verb/ Auxiliary Verb/ Adverb/ Particles/ Coordinate Conjunctions

Regular Expression of Rule 3:
(NOUN|PROPN)+(VERB|AUX|ADV|PART|CCONJ)* (ADJ)+ (VERB|AUX|ADV|PART|CCONJ)*

TABLE IX. EXAMPLE OF OPINION SENTENCE AND CORRESPONDING EXTRACTED ASPECT AND OPINION (RULE 3)

#	Opinion Sentence							Aspect	Opinion
1	customer	service	be	excellent	easy	to	deal	customer service	excellent
	NOUN	NOUN	VERB	ADJ	ADJ	PART	VERB	customer service	easy
2	delivery	fee	unreasonable	at	all			delivery fee	unreasonable
	NOUN	NOUN	ADJ	ADV	ADV				
3	rider	be	all	nice	and	attentive		rider	nice
	NOUN	AUX	ADV	ADJ	CCONJ	ADJ		rider	attentive
4	rider	be	too	much	slow			rider	slow
	NOUN	AUX	ADV	ADV	ADJ				
5	location	not	accurate	sometime	can	not	detect	location	not accurate
	NOUN	PART	ADJ	ADV	AUX	PART	VERB		

TABLE X. NON OPINION SENTENCES

#	Non Opinion Sentence						
1	convenient	cheap	fast				
	ADJ	ADJ	ADJ				
2	location	and	gps	issue			
	NOUN	CCONJ	NOUN	NOUN			
3	love	it	easy	to	get	food	
	VERB	PRON	ADJ	PART	VERB	NOUN	
4	after	fpx	payment	the	app	be	stick
	ADP	ADJ	NOUN	DET	NOUN	AUX	ADJ
5	easy	to	use	simple	fast		
	ADJ	PART	VERB	ADJ	ADJ		
6	convenient	and	on	the	go		
	ADJ	CCONJ	ADP	DET	NOUN		
7	efficient	and	accessible	for	all		
	ADJ	CCONJ	ADJ	ADP	PRON		

G. Aspect Candidate and Opinion Word Extraction from User Review

This section explains the process of extracting aspects and opinions from user reviews and how they are paired.

By default, every noun is paired with the adjective that follows it in sequence. For example, if a review contains the pattern 'NOUN ADJ NOUN ADJ', the first noun will be paired with the first adjective, and the second noun will be paired with the second adjective, and so on.

Referring to Table IX, the third review provides two sets of aspects and opinions: 'rider, nice' and 'rider, attentive'.

If the number of nouns/proper nouns and adjectives is not equal, a cascade-style pairing is applied. In this scenario, when there is one noun and two adjectives, the single noun will be paired with each of the adjectives.

There is a special scenario, 'NOUN PART ADJ' (e.g., "location not accurate"), where the opinion "not accurate" is extracted. The rule is that when the particle (PART) precedes the adjective (ADJ), the particle is extracted together with the adjective. This is because particles are closely associated with the word that follows them, and including the particle ensures the complete context of the sentence. If the particle is not extracted, the opinion may be completely incorrect, mistakenly extracting only "accurate" instead of "not accurate."

IV. CONCLUSION

The separation of opinion and non-opinion reviews is vital. It is important to understand that individuals often express themselves in an uncompact manner. Therefore, prioritising shorter reviews (less than 10 words) tends to convey the core information effectively. The use of part of speech tagging and regular expressions has proven to be remarkably useful in extracting key aspects, such as delivery time, delivery fee, customer service, app performance, payment method, rider, and more.

Lengthy negative reviews (with a maximum word count of 304 in this study) often involve storytelling to express dissatisfaction or describe specific issues.

In this study, improvements were made based on previous research, which include:

- More comprehensive data visualisation, such as analysing the top 10 words occurring before and after specific keywords.
- Text normalisation, where similar phrases like "customer care," "customer support," and "customer service" are grouped together.
- Extraction of aspects and opinions from reviews, such as extracting "delivery fee" and "expensive" from the sentence "expensive delivery fee."

However, the present study does have some limitations, which are outlined below:

- Reviews exceeding 10 words were not considered when developing the rules for matching opinion sentences.
- Words tagged as VERB were not used as opinion words due to time constraints in designing more complex regular expressions.
- Removal of punctuations affected compound adjectives (multiple adjectives that joined with hyphen) lost its context. E.g., time-saving (ADJ) became time saving (ADJ VERB).
- Lemmatisation to all words made some reviews hard to understanding. E.g., ‘rider be too slow’, the original sentence is ‘rider is too slow’; ‘time save’, the original sentence is ‘time-saving’.

Suggestions for Foodpanda consumers

- Long negative reviews often indicate highly unsatisfactory experiences, as individuals usually do not invest extensive time unless deeply dissatisfied.
- Therefore, it is advisable for consumers to directly contact Foodpanda's help desk for prompt action, as the resolution of concerns solely based on Google Play Store reviews remains uncertain.
- Google Play Store reviews typically serve as a quick reference for non-users evaluating the app.

Suggestions for Foodpanda

- To maintain brand quality, implementing filters to identify excessively long reviews and thoroughly investigating such cases would be beneficial.
- Additionally, employing aspect-opinion analysis will provide Foodpanda with detailed insights into various aspects of their service.
- The substantial number of negative reviews related to customer service highlights the need for specific investigation in this area.

Future work

- Rule-based approaches necessitate considerable manual effort for rule creation.
- Perform lemmatisation on noun only to retain review context.
- Retain certain punctuations such as hyphen and quotation mark to reserve review context. E.g., ‘time-saving’ and ‘can’t’,
- Exploring intelligent approaches for more automated and efficient analysis is recommended for future studies.

V. REFERENCES

[1] "App Insights: Foodpanda," [Online]. Available: https://www.crunchbase.com/organization/foodpanda/technology.

[2] S. Y. Teoh and W. M. R. Afifi, "Exploring Foodpanda consumer reviews using Part of speech and N-gram," 2022.

[3] S. G. a. M. L. Khodra, "Aspect based sentiment analysis for review rating prediction,," *Theory Appl. ICAICTA,* vol. 4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, 2016.

[4] D. E. a. M. L. Khodra, "Aspect-based sentiment analysis for Indonesian restaurant reviews," *Theory Appl. ICAICTA,* Vols. Proc. - 2017 Int. Conf. Adv. Informatics Concepts, 2017 .

[5] N. Z. A. K. a. T. A. N. Akhtar, "Aspect based Sentiment Oriented Summarization of Hotel Reviews," *Procedia Comput. Sci.,* vol. 115, p. 563–571, 2017.

[6] L. Z. a. Y. L. S. Li, "mproving aspect extraction by augmenting a frequency-based method with web-based similarity measures," *v,* vol. 51, p. 58–67, 2015.

[7] F. L. a. T. G. Yamina, "Mining explicit and implicit opinions from reviews," *Int. J. Data Mining, Model. Manag.,* vol. 8, p. 75, 2016.

[8] T. A. R. a. Y. N. Cheah, "A two-fold rule-based model for aspect extraction," *Expert Syst. Appl.,* vol. 89, p. 273–285, 2017.

Appendix

[URL to Presentation Slide of this study]

[URL to previous study report]

[Python scripts used for this study]

[Raw data (Extracted from Google Play Store)]

[Pre-processed Positive and Negative reviews]

Bigram word cloud after normalisation (Noun only)

Unigram word cloud after normalisation (Noun only)

Solr based Recipe Search Engine

Teoh Sin Yee
School of Computer Sciences USM
USM, Penang
teohsinyee.cs@gmail.com

Mohamed Abdelnasser Mohamed Elsayed Hassan
School of Computer Sciences USM
USM, Penang
mabdelnasser@student.usm.my

***Abstract*— This report presents the development of ORecipes, a new recipe search engine that is designed to address the limitations of existing recipe search engine. Solr and Flask is being used to support system backend. The existing search engines fail to consider individual user preferences and struggle with accurate synonym recognition. ORecipes utilizes an advanced search method that prioritizes user preferences and optimizes recipe searches accordingly. It also incorporates comprehensive synonym handling to ensure accurate retrieval of relevant recipes. The system is designed to provide users with a specific and personalized search experience, allowing them to exclude or emphasize certain ingredients. The report describes the dataset used for system demonstration, the query description and logic, and the system design, including system preparation, query processing and retrieval, and system performance evaluation. ORecipes has the potential to revolutionize recipe searching by offering more advanced and customized features to enhance the user experience.**

Keywords—Solr, search engine, information retrieval, advanced search, personalized search, synonym recognition, spellcheck, field boosting, search query

I. INTRODUCTION

Searching for a specific recipe on the internet has become a trend and a common activity for many people. They may want to expand and improve their cooking skills or explore new dishes in different cuisines. The existing websites and search engines do not often fulfil the needs and the requirements of users, leading to frustration and missed results. In this report, we are introducing a new recipe search engine called ORecipes that aims to solve two of the main problems that most of current recipe search engines have by offering more advanced and customized searching features for the user to get find their needed recipe faster and easier.

The first issue with existing systems is that they do not take individual user preferences into account when optimising recipe searches. For example, if someone wants to cook a brinjal (eggplant) dish with chicken as an addition, a typical search query like 'brinjal chicken' may not accurately convey the importance of the chicken as an additional ingredient. To address this, ORecipes uses field boosting that assign more weight to brinjal based on the user's preferences. This enables the system to better understand the user's intent and provide recipe results that match their desired combination.

The second issue is the limitations of existing search engines in dealing with synonyms. For example, if someone searches for 'brinjal', they might miss relevant recipes that use the synonym 'eggplant'. Even well-known international recipe websites like Epicurious struggle to support multiple languages and recognise synonyms accurately. [1]

ORecipes intends to transform recipe search by addressing the previously mentioned issues. By implementing advanced search methods, user-assigned weight optimisation, and comprehensive synonym handling, ORecipes provides recipe enthusiasts with a more specific and personalised search experience. This system is for people who need to exclude or emphasise certain ingredients in their recipes. This report explores ORecipes functionalities, and the potential impact it can have on the future of recipe searching.

II. DATASET DESCRIPTION

A. Source of Dataset

The dataset was downloaded from the Kaggle website [3]. It consists of 10 thousand rows that were originally created by scraping the necessary data from the Epicurious website. The original dataset only contained five columns, namely title, Ingredients, Instructions, Image_Name, and Cleaned_Ingredients. Some of these column names have been updated. Additionally, two new columns named Cuisine and Meal were added to the original dataset. The values for these two fields were populated using the available ChatGPT intelligent technology [4], which recognized the meal and its cuisine based on the provided values in other fields. A sample of 50 rows was selected for our system demonstration.

B. Fields description

After cleaning the dataset, updating columns name, and adding two extra columns, the new sample dataset consists of 7 fields as follows:

1. **title**: The title of the recipe, which represents the name or brief description of the dish. (Sicilian Grill-Roasted Chicken)
2. **cuisine**: The cuisine associated with the recipe, indicating style of cooking. (e.g. Italian, French, …etc)
3. **meal**: Specifies the type of meal the recipe is intended for. (e.g. main, side, …etc)
4. **ingredient**: Describes the ingredients required for the recipe.
5. **step**: Provides a series of steps that outline the cooking process.
6. **image_name**: Represents the name of the associated image file that visually represents the recipe.
7. **id**: A unique identifier assigned to each recipe.

All the mentioned fields are indexed except for image_name field, because sometimes it contains a random value which is not relevant to the recipe. Beside that each field in a single valued field except for the ingredient field which is multivalued. For a better illustration for our dataset fields, a

table 'dataset fields and description' is provided in the appendix.

III. QUERY DESCRIPTION

A. *Information need*

TABLE I. INFORMATION NEED OF EACH QUERY

	query	***information need***
1	+(ciken AND (title:brinjal^4 OR ingredient:brinjal^2))	Must contain brinjal as the main dish while also including chicken as an additional component.
2	+pie AND -nuts	Must contain pie that does not use any kind of nuts as ingredients
3	+vege AND -meat	Must contain vegetables as the main dish and should not include any kind of meat as ingredients

B. *Logic of query formation*

To justify the decision for forming the queries and meet the information need, the following reasoning can be provided:

For Query 1, two ingredients are mentioned, with an emphasis on brinjal. Weight is assigned to brinjal to indicate its importance. This decision assumes that if brinjal appears in the title of a dish, it is more likely to be the main ingredient. Therefore, a boost factor of 4 is set for the title field. Additionally, if brinjal appears in the list of ingredients, it could be either a main ingredient or a side ingredient. To accommodate this, a lower boost factor of 2 is assigned to the ingredient field. This reflects your preference for dishes that include brinjal, regardless of its role.

In Query 2, the requirement is to exclude any kind of nuts as ingredients in the pie. This is achieved by using the '-' operator to exclude nuts. The query ensures that the documents retrieved contain pie but do not contain any nuts as ingredients.

Similarly, in Query 3, the goal is to include vegetables as the main dish while excluding any kind of meat as ingredients. By using the '-' operator, the query ensures that the documents retrieved contain vegetables but do not contain any meat as ingredients.

To cater to the various varieties of nuts and meat, a synonym lexicon is utilized for mapping purposes. This allows for flexibility in handling different names and variations of nuts and meat, ensuring comprehensive retrieval of relevant documents.

IV. SYSTEM DESIGN

The system can be divided into five parts, each serving a crucial role in enabling efficient and accurate recipe searches. These parts are:

A. *System Preparation*

Step 1: Customizing the Solr Schema

The *managed-schema.xml* file of the Solr core is tailored to define the fields and their types, enabling indexing, and configuring text processing options such as tokenization and stemming.

The field types for title, ingredient, and step are changed from text_general to text_en. This change is implemented to leverage additional text processing steps specific to the English language. The field type for the image_name field remains as text_general. This field is used to store the image name associated with a recipe but is not utilized for matching purposes. Instead, it helps form the image URL for displaying recipe images in the user interface.

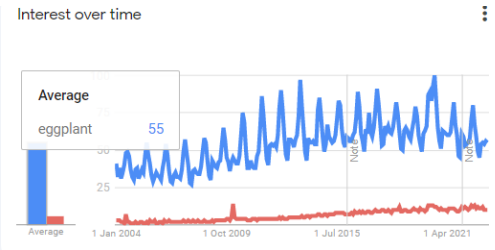
Step 2: Establishing Synonym Expansion

A synonym dictionary is implemented to enhance search results by mapping equivalent terms or phrases. *The dictionary maps equivalent terms and useful to handle ingredient names,* which can vary due to cultural differences or regional preferences.

For example, consider the terms 'brinjal' and 'eggplant'. Based on Google Trend analysis (as shown in Fig. 1), it has been observed that worldwide users tend to search for 'eggplant' 90% more frequently than 'brinjal'. [6] This difference in terminology can lead to disparities in search results. For instance, a popular recipe website like Epicurious might yield no results when searching for 'brinjal', while returning 700 results for 'eggplant'.

By establishing synonym expansion, the recipe search engine addresses these variations and ensures that users can retrieve relevant recipes regardless of the term they use.

Fig. 1. Google Trend Analysis (Brinjal vs Eggplant)



Step 3: Indexing

The recipe data is imported from a CSV file into the Solr index.

B. *System Preparation*

Step 4: Developing the User Interface

A simple user interface is created by using Flask [7] for the backend and incorporates HTML, CSS, and JavaScript for the frontend. Users can search, filter inputs and the matched term in result will be highlighted in color.

C. *Query Processing and Retrieval*

Step 6: Tokenization and Analysis

User inputs can span multiple lines and can be structured in four different formats, which leads to the need for curating four sets of rules for processing and analysis.

These formats allow users to express their information needs with various conditions, including regular queries and field boosting as shown in Table II.

TABLE II. STRUCTURES OF STRING TOKENS FROM USER INPUT

#	***Structures***	***Example***	***Description***
1	No boost	+chicken	Must have chicken.
2	Boost	+title:chicken^2	Must have chicken. If chicken is in title, the document rank score will be doubled.
3	Boost + No boost + …	+(chicken AND (title:brinjal^4 OR ingredient:brinjal^ 2))	Must have chicken and brinjal. If brinjal is in title, the document rank score will be multiplied by four. If brinjal is in ingredient, the document rank score will be doubled.
4	Boost + Boost ….	+(title:brinjal^4 OR ingredient: brinjal ^2)	Must have brinjal and brinjal. If brinjal is in title, the document rank score will be multiplied by four. If brinjal is in ingredient, the document rank score will be doubled.

To demonstrate the tokenization process, let's consider the third structure of the input string as an example. This structure is as follows:

TABLE III. EXAMPLE QUERY

+(chicken AND (title:brinjal^4 OR ingredient:brinjal^2))

Table IV displays the raw user input obtained after clicking the search button. It consists of two conditions. The first condition is '---brinjal--title-4', the second condition is '---brinjal--ingredient-2', and the third condition is '---ciken--#-#'. In the third condition, the boost field and boost factor are not selected, hence represented by the '#' symbol.

After the raw user input is obtained, it will undergo postprocessing using the split and pair functions described in Table V.

The postprocessing procedure involves scanning the input string character by character, identifying groups of hyphens, and extracting the substrings between them. These substrings are then categorized into three groups: 'text,' 'boostfield,' and 'boostfactor,' based on the number of hyphens present in each substring.

Once categorized, the elements from each category are paired together, forming pairs. The pairs are then sorted based on the length of the 'boostfield' element. The sorted pairs are the resulting output of the postprocessing procedure.

For a detailed explanation of the split and pair functions and their respective operations, please refer to Table V. The output of the function is tabulated in Table VI.

TABLE IV. STRING TOKENS FROM USER INPUT

---brinjal--title-4---brinjal--ingredient-2---ciken--#-#

TABLE V. PSEUDOCODE OF SPLIT AND PAIR FUNCTION

```
def split_and_pair(s):

Initialize an empty dictionary result with keys 'text', 'boostfield',
and 'boostfactor'

Set i = 0

Loop through the input string (s)

   if s[i] equals '-': If current character is hyphen
      Initialize if number of hyphen to 1
      Set j to next character position

      Loop until next character position is
      greater than length of input string and
      next character NOT equal to hyphen:

         Increment the number of hyphen by 1

           Set k = j

Loop until k is
greater than length of input string and
next character NOT equal to hyphen:

                  Increment k by 1
                  Set s[j:k] to value

                   if number of hyphen is 3:
                      Append value to result['text']

                   else if number of hyphen is 2:
                      Append value to result['boostfield']

                  else if number of hyphen is 1:
                     Append value to result['boostfactor']


Initialize an empty list 'pairs'

Loop through the length of result[boostfield]:
   Set x = result['text'][i]
   Set y = result['boostfield'][i]
   Set z = result['boostfactor'][i]

   Append the tuple (x, y, z) to pairs

Sort pairs based on the length of the boostfield in ascending order

Return the sorted pairs
```

TABLE VI. TOKENS PAIR FORMED FROM USER INPUT

[('ciken', '#', '#'), ('brinjal', 'title', '4'), ('brinjal', 'ingredient', '2')]

Step 7: Spellcheck

The tokenized query is sent to Solr for spellchecking. Solr provides spellcheck suggestions based on token frequency, effectively assisting in rectifying potential misspelled words within the query.

As shown in Table VII, Solr provided two spell check suggestions. Word 'chicken' has higher frequency than word 'cider', hence it is set to default that the token will be autocorrected to the suggestion that has highest frequency. Table VIII shows the corrected token pair.

TABLE VII. SOLR SPELL CHECK SUGGESTIONS

```
'suggestion': [
{'word': 'chicken', 'freq': 37}, {'word': 'cider', 'freq': 3}]}] …
```

TABLE VIII. TOKEN PAIR AFTER SPELL CORRECTED

```
[('chicken', '#', '#'), ('brinjal', 'title', '4'), ('brinjal', 'ingredient', '2')]
```

D. Query Construction

Step 8: Constructing the Query

After receiving the spellcheck suggestions from Solr, the final query is constructed by incorporating the corrected tokens. Appropriate search operators (e.g., AND, OR, etc) are applied to create a well-structured Solr query string as shown in Table X.

TABLE IX. TOKENS MERGED WITH SEARCH OPERATORS

```
token_merged =
+chicken AND +(title:brinjal^4 OR ingredient:brinjal^2)
```

TABLE X. CONSTRUCTED QUERY STRING

```
final_query =
f'http://localhost:8983/solr/recipe/select?hl=true&hl.fragsize=4
000&hl.fl=title%2Cingredient%2Cstep&q={token_merged}&fl
=*,score&wt=python'
```

E. Query Submission and Result Presentation

Step 9: Query Submission

The constructed query is sent to Solr for execution to retrieve relevant recipe documents based on the query. Before retrieving results, the submitted query is subjected to further tokenization and analysis utilizing Solr's built-in tokenizers, filters, and analysers. This process handles language-specific processing, synonym expansion, stemming, and other pertinent NLP operations.

Step 10: Retrieving and Ranking Results

Relevant recipe documents are retrieved from Solr based on the executed query, incorporating relevancy ranking factors such as field boosts and document scores.

Step 11: Presenting the Results

The search results are presented in the user interface, showcasing key information such as recipe title, ingredient, step, picture, cuisine type and rank score. Additional features such as filters and result highlighting are implemented to enrich the user's browsing experience.

V. RESULTS AND DISCUSSIONS

To evaluate the ranking results, five different metrics are used: precision(P), recall (R), F1-score (F1), precision@k (P@k) and recall@k (R@k). In this case, the value of k is set to 5, which means it considers the top 5 results. This choice is made to represent 10 percent of the dataset.

TABLE XI. PERFORMANCE RESULT

	Query	***Actual Relevant***	**Metrics**				
			P	***R***	***F1***	***P@5***	***R@5***
1	+(chicken AND (title:brinjal^4 OR ingredient:brinjal^2))	7	0.286	1	0.445	0.4	1
2	+pie AND -nuts	2	1	1	1	1	1
3	+vege AND -meat	2	1	1	1	1	1

In the dataset, there are several ingredient phrases such as 'chicken stock', 'chicken broth', and 'tomato sauce'. However, the search system treats these phrases as individual terms rather than recognizing them as a whole. As a result, there is an expectation of many false positives due to the nature of the Solr system.

For instance, when a user searches for chicken recipes, some dishes that do not actually contain chicken as an ingredient might be returned because they include the term 'chicken' in phrases such as 'chicken stock'.

Based on the obtained results, it is found that the first query has the lowest precision and precision@5 because it contains many false positives, i.e., retrieved results that are not relevant.

For example, the top-ranked result might be 'chicken broth', which does not actually use chicken as an ingredient. This issue arises due to the nature of the Solr system, which matches single terms rather than recognizing phrases.

On the other hand, when a user searches for 'pie', there are no phrases in the dataset that contain the term 'pie' along with another word. Consequently, there are no false positives in this case. This could explain why query 2 and 3 achieved perfect scores for all metrics.

Limitations

One limitation of the system is the spell check functionality. While it is designed to help users by automatically correcting misspelled words, it may not always accurately reflect the user's intended search query. For example, when the user inputs 'beef', it is being corrected to 'been', even though the documents in the dataset contain the word 'beef.' This can lead to incorrect search results or confusion for the users.

Additionally, the spell check feature in Solr, may not be comprehensive enough to handle all possible spelling variations. For instance, when searching for 'bef,' it is

expected to return results containing 'beef,' but it shows no matching results.

Improvements

One significant improvement to enhance the user experience is to allow users to input their queries in the form of natural language sentences instead of requiring them to specify the boost field, boost factor, and text column. This approach would provide a more intuitive and user-friendly way to interact with the system. For instance, users could input a sentence like 'brinjal as the main ingredient and chicken as an addition.' By implementing natural language processing techniques, the system would parse the user's input sentence, extracting keywords, entities, and phrases. It would automatically recognize the importance of brinjal as the main ingredient and assign it more weight in the search.

Based on the parsed information, the system would then generate a structured query, inferring the appropriate boost field, boost factor, and text column without requiring explicit user input. This eliminates the need for users to have knowledge of the underlying system's structure.

Moreover, the system would consider the context of the search, considering previous user interactions or session history to further improve relevance. This ensures that the search results align with the user's preferences and previous search behavior.

Overall, this natural language approach empowers users by providing a user-friendly interface and reducing the complexity associated with query structures. It significantly improves the overall search experience and helps users find relevant recipes more efficiently.

VI. CONCLUSION

In this report, we have developed a Solr-based search engine that provides advanced search capabilities for 50 recipe documents. The search engine allows users to include or exclude specific text from the search results, enabling them to refine their queries and obtain more relevant recipes.

To enhance the search functionality, additional fields such as cuisine type and meal type were generated using ChatGPT. This allows users to narrow down their recipe selections based on specific culinary preferences. Although the dataset did not originally include these fields, their inclusion improves the usability and effectiveness of the search engine.

To ensure comprehensive retrieval of relevant documents, a synonym lexicon was manually created to handle variations in ingredient names. This helps to capture synonymous terms and prevent the exclusion of relevant recipes from the search results. However, as the dataset grows larger, an AI-based approach for expanding the synonym lexicon would be more effective in covering a wider range of synonyms.

The search engine implements field boosting to optimize document ranking. Users have the flexibility to specify the boost field and boost factor, allowing them to prioritize certain fields and influence the ranking of search results.

However, this feature may be less intuitive for less tech-savvy users who may struggle with determining the appropriate boost values. To address this, future work will focus on diversifying the user interface by providing options for users to rank their inputs manually, providing a more natural and user-friendly approach.

In terms of user interface design, we have explored and discovered new frameworks such as Elastic UI and Material UI, which offer more visually appealing and engaging designs compared to traditional frameworks like Bootstrap. Utilizing these frameworks can enhance the aesthetic appeal of the user interface, improving the overall user experience.

REFERENCES

[1] Epicurious, “Recipes, menu ideas, videos & cooking tips,” Epicurious, https://www.epicurious.com/.

[2] Welcome to Apache Solr - Apache Solr, https://solr.apache.org/.

[3] S. Goel, Kaggle Food ingredients and recipes dataset with images, https://www.kaggle.com/pes12017000148/food-ingredients-and-recipe-dataset-with-images.

[4] Introducing ChatGPT, https://openai.com/blog/chatgpt.

[5] Elastic UI, https://eui.elastic.co/#/.

[6] Google trends, https://trends.google.com/trends/explore?hl=en-US.

[7] Welcome to Flask - Flask Documentation (2.3.x), https://flask.palletsprojects.com/en/2.3.x/.

APPENDIX

Dataset Fields and Description

Fields	Description	Indexed	Multivalued	Examples
title	The title of the recipe, which represents the name or brief description of the dish	Yes	No	Sicilian Grill-Roasted Chicken, …
cuisine	The cuisine associated with the recipe, indicating style of cooking	Yes	No	Italian, French, ..
meal	Specifies the type of meal the recipe is intended for	Yes	No	Main, side, …
ingredient	Describes the ingredients required for the recipe	Yes	Yes	[‘['4 garlic, finely grated',
step	Provides a series of steps that outline the cooking process	Yes	No	1. Grease the slow cooker with cooking spray.\n2
image_name	Represents the name of the associated image file that visually represents the recipe	No	No	sicilian-grill-roasted-chicken-24270
id	A unique identifier assigned to each recipe	Yes	No	1,2,3….

Query Search Performance

	Actual relevant	***TP***	***FP***	***TN***	***FN***
+(chicken AND (title:brinjal^4 OR ingredient:brinjal^2))	2	2	5	43	0
+pie AND -nuts	2	2	0	48	0
+vege AND -meat	2	2	0	48	0
+chicken	32	33	4	17	0

'synonyms.txt'

```
#Lexicion of Vegetables
vegetable => eggplant, broccoli, chickpea, radicchio, cauliflower, spinach, kale, carrot, cucumber, tomato, radish, lettuce, green bean, pea, cabbage, sprout

#Lexicion of Meats
meat => chicken, lamb, beef, pork, turkey, duck, venison, veal, bison, rabbit, quail, goat, pheasant, ostrich, kangaroo, wild boar

#Lexicion of Nuts
nuts => nuts, cashew, peanuts, almonds, pecans, walnuts, hazelnuts, pistachios, macadamia, chestnuts

brinjal, eggplant, eggplants
tomato, tomatoes
```

'managed-schema.xml' (Modified part only)

```
....
 <field name='id' type='string' multiValued='false' indexed='true' required='true' stored='true'/>
  <field name='image_name' type='text_general' indexed='false'/>
  <field name='title' type='text_en' indexed='true' stored='true'/>
  <field name='ingredient' type='text_en' multiValued='true' indexed='true' stored='true'/>
  <field name='step' type='text_en' indexed='true' stored='true'/>
  <field name='cuisine' type='text_en' indexed='true' stored='true'/>
   <field name='meal' type='text_en' indexed='true' stored='true'/>
....
  <copyField source='image_name' dest='image_name_str' maxChars='256'/>
  <copyField source='title' dest='title_str' maxChars='256'/>
  <copyField source='ingredient' dest='ingredient_str'/>
  <copyField source='step' dest='step_str'/>
  <copyField source='meal' dest='meal_str' maxChars='256'/>
  <copyField source='cuisine' dest='cuisine_str' maxChars='256'/>
....
```

Slide & Screenshots

https://bit.ly/SolrSearchEngineSlides