Sentiment Analysis and Topic Modelling

For Cililado

Group 16

Jinan, Aaron, Hanif,
Chun Xiang, Darren

Overview

- Introduction
- Problem Statement
- Objectives
- Proposed Solution
- Dataset
- Implementation
  - Libraries
  - Web Scraping
  - Descriptive Analysis
  - Data Cleaning
  - Pre-processing
  - Modelling
  - Results
- Results and Inferences
- Recomendations
- Conclusion

Introduction

- Among the many SMEs in this country is Cili Lado. Cili Lado is an SME based in Melaka initiated by Afiq Noordin who take pride in their true sambal flavoured sambal minang. However, even with the successes it has several challenges in its business operations.

- While it has been in operation since 2018 and has made numerous marketing strategies/effort to promote its sambal, there are still area that can be further enhanced and improved.

Problem Statement

Problem

The company is facing the challenge of optimising its sales strategy and customer engagement processes. Despite receiving generally positive online reviews, there is a discrepancy between sentiment analysis predictions and actual customer ratings, indicating potential gaps in understanding and addressing customer needs. There's also a need to better segment the market, tailor product offerings, and enhance the overall customer experience to improve loyalty and retention.

Objectives

Objectives 1

To accurately gauge customer sentiment, understand purchase behaviors, and refine product offerings.

Objectives 2

To enhance sales performance and customer satisfaction through data-driven insights.

Objectives 3

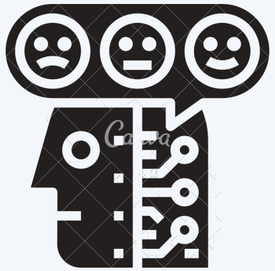
To improve customer segmentation for targeted marketing and personalized customer experiences.

Objectives 4

To highlight the limits of exisiting selling platform and give recommendations.

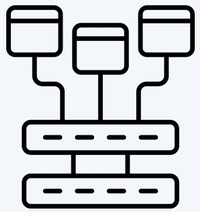
Proposed Solution

Sentiment Analysis



To understand customer bias by categorizing reviews as very positive, positive, neutral and negative by analysing assigned sentiment scores.

Topic Modelling



Deriving topics of focus to assist in targeted campaigns.

Models

LDA

Linear model for classification and dimensionality reduction, often used in the use cases of topic modelling.

Textblob

Natural Language Processing library, often used with NLTK to achieve categorization and classifaction based on sentiments.

VADER

Valence Aware Dictionary for sEntiment Reasoning is a lexicon and rule-based approach for expressing sentiment polarity and intensity. It particularly handles social media data well.

Dataset

	A	B	C	D
1	user	review	rating	date
2	anita_azrina	Taste: delicious Quality: good The parcel arrived safely and very c	5	1688382244
3	m*****a	Taste: Seriously delicious Quality: Top tip Not very spicy, good for	5	1696379181
4	asyrafringo88	Taste: delicious and delicious Quality: the best I originally wasn't	5	1696058776
5	nadzmianuar98	Quality:best quality Taste:taste good & not spicy Fast delivery by	5	1696507977
6	syamil1989	Quality: looks neat Taste: licking toes (chicken feet) Packing is sol	5	1687517621
7	shamsulfa	Taste: delicious and spicy. Quality: the best. first time bought bec	5	1690762728
8	faiz120180	The goods arrived quickly & well packaged 👍, I will continue to	5	1695643297
9	nujaeila	Taste: delicious Quality: good Alhamdulillah the item arrived safe	5	1695978806
10	fluffybunny31	Taste: Delicious Quality: Best Third purchase. Yesterday I had a n	5	1694944174
11	s*****l	Quality: the best taste Taste: not sweet and spicy as usual Second	5	1695895556
12	asyrafringo88	Taste: Good Quality: Great Good delivery package. It is solid and	5	1696144908
13	syamil1989	Taste: Delicious Quality: Best I was skeptical at first about this sa	5	1690330849
14	j*****i	Quality: nice wrapping, he's strong Taste: green, not spicy, red is	5	1688059421
15	fairoz87	Quality: Looks OK but the lid of the red chili is a little dented Tast	5	1687778459
16	j*****e	Quality: good packaging. The best wrapping Taste: haven't tried y	5	1696412148
17	dormameow	Quality: the best.. Taste: delicious. The goods arrived in good con	5	1691759268
18	nikjanahhzhkn	Quality: best 4/5 Taste: moderately spicy Sambal does not have s	4	1687816690
19	muhammadsyarifchesuliman	Taste: 100% delicious Quality: Top tip. Good packaging, no leakin	5	1695263513
20	suesani8177	Taste:Good Quality:Excellent The green sambal is not spicy at all,	5	1696203346
21	a*****n	Taste:pedas n sedap Quality:padu Beli 2 tp live lain2 padu mkn n	5	1695511291
22	zati.amran	Quality: best Taste: delicious green minang sambal is not spicy bu	5	1695161088

The data is scraped from Shopee for the reviews of Sambal Minang as of October 2023. It states information about the username, review posted, rating given for the product, and date of review posted. The date is in Unix Timestamp format and is not directly interpretable.

Implementation

Web Scrapping

Scraping data from Shopee

Descriptive Analysis

Analyzing dataset structure and distribution

Data Cleaning

Removing irrelevant columns and missing values.

Pre-processing

Stopwords, URL, special characters, emoji removal etc

Modelling

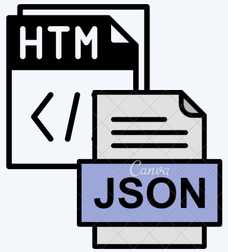
Textblob and VADER for sentiment analysis and LDA for topic modelling

Visualizing Results

Barcharts, histograms and wordclouds

Libraries

The core libraries used in our sentiment analysis and topic modelling

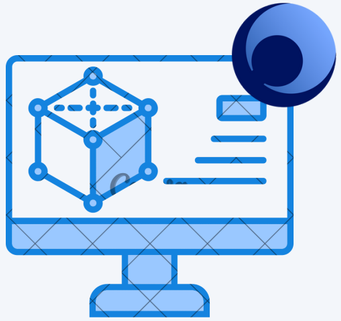


These libraries are used to getting data from webpages using **HTML requests** and storing in **JSON** objects.

Re, HTML Requests, Json
Webscrapping

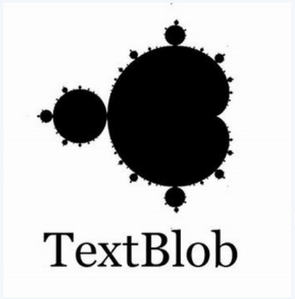
Numpy is used for performing array based calculations while **Pandas** is used to store data in objects called Dataframes.
Matplotlib is used for visualizing data.

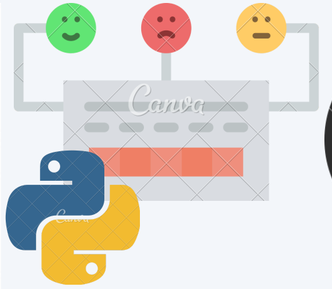
Pandas, Numpy, Matplotlib
Data Analysis and Visualization



Gensim is an unsupervised topic modelling library which is used for topic model and representation of text as vectors.
pyLDAvis allows us to interactively explore topics obtained by an LDA model.

Gensim, pyLDAvis
Topic Modelling

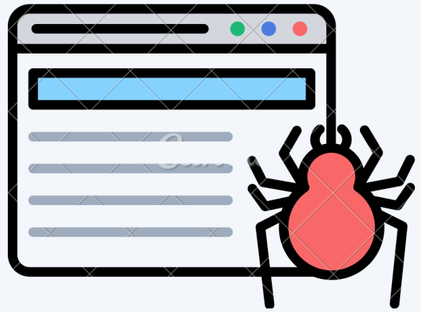




NLTK is a major natural language processing toolkit.
Textblob is commonly used for sentiment classifications and POS tagging.
Worldcloud is used to visuzalise freeform texts in our results.
VADER is also used for sentiment classification that is focused on social media.

NLTK, textblob, Worldcloud, VADER
Sentiment Analysis and Text Processing

Web scrapping



Web scrapping is the process of extracting data from websites and is often automated using web crawlers.

Our data is scraped from the following link:

https://shopee.com.my/Cili-Lado-Sambal-Minang-250-Gram-Halal-i.34398605.8433245961

Libraries like 're' are used to specify regular expressions that help access pages for both the products, i.e. the green sambal minang and red one.

HTTP get request is made to get the SHOPEE API to get the ratings and reviews.

```
while True:
    data = requests.get(ratings_url.format(shop_id=shop_id, item_id=item_id, offset=offset)).json()

    for rating in data['data']['ratings']:
        author_username = rating['author_username']
        comment = rating['comment']
        rating_star = rating['rating_star']  # Extract rating
        rating_date = rating['ctime']  # Extract date
        reviews_data.append({'Author Username': author_username, 'Comment': comment, 'Rating': rating_star, 'Date': rating_date})
```

Finally, dataframe is converted to csv and stored. These reviews are in Malay and for accurate sentiment analysis, they are translated one by one by using Google Translate in English. Alternatively, Google Translate API can be used if higher computing resources are available.

Descriptive Analysis

Descriptive Analysis is done at the beginning to analyze the structure of the dataset. First five data values are printed and the shape of the data frame is accessed to determine the number of rows (863) and columns (4). .info() function is used to check the data types for each column while also checking the null values.

```
df2.head(5)
```

	user	review	rating	date
0	anita_azrina	Taste: delicious Quality: good The parcel arri...	5	1688382244
1	m*****a	Taste: Seriously delicious Quality: Top tip No...	5	1696379181
2	asyrafringo88	Taste: delicious and delicious Quality: the be...	5	1696058776
3	nadzmianuar98	Quality:best quality Taste:taste good & not sp...	5	1696507977
4	syamil1989	Quality: looks neat Taste: licking toes (chick...	5	1687517621

```
df2.shape
```

```
(863, 4)
```

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 863 entries, 0 to 862
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   user    862 non-null    object
 1   review  412 non-null    object
 2   rating  863 non-null    int64
 3   date    863 non-null    int64
dtypes: int64(2), object(2)
memory usage: 27.1+ KB
```

Descriptive Analysis (cont.)

.describe() function is used to display statistical information about text-based columns as well as numeric columns.

```
text_columns = df2.select_dtypes(include=['object'])
text_columns.describe()
```

	user	review
count	862	412
unique	680	406
top	wmhaziq90	Very tasty
freq	18	4

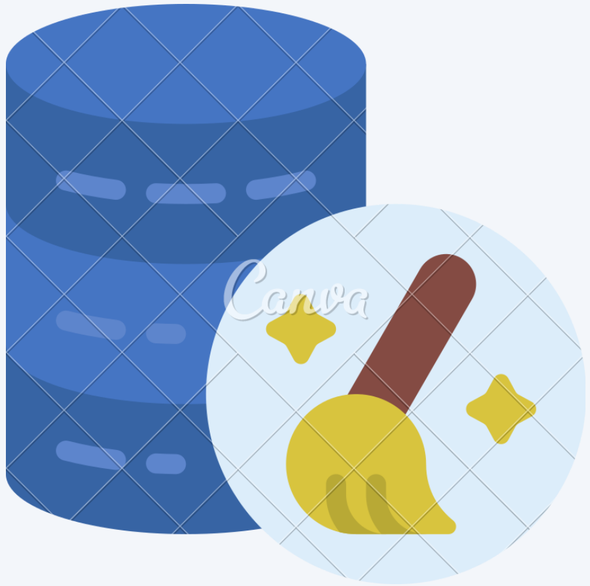
```
df2.describe()
```

	rating	date
count	863.000000	8.630000e+02
mean	4.866744	1.687059e+09
std	0.525096	1.864754e+07
min	1.000000	1.617282e+09
25%	5.000000	1.689416e+09
50%	5.000000	1.693145e+09
75%	5.000000	1.695896e+09
max	5.000000	1.696739e+09

'user' has 862 values and out of them 680 are unique which indicates multiple reviews by same users. 'review' has 412 values and out of them 406 are unique. 'top' shows the top most value in each column.

'rating' has a mean value of 4.86 which gives us the mean rating given to cililado products. The minimum rating is 1 and maximum rating is 5 while the standard deviation of ratings is 0.52.

Data Cleaning



Data Cleaning is the process of transforming raw data by fixing incorrect, incomplete, duplicate or erroneous data.

Our data is cleaned by dropping irrelavant columns and removing missing values which are found using isna() function of pandas. The missing values detected in the review column are filled with a blank space ' ' by fillna() function so as to preserve the ratings information. The dropped columns are 'user' and 'date' as both the features are irrelevant to the process of sentiment analysis.

```
df2.isna().sum()
```

```
review    451
rating      0
dtype: int64
```

```
df2['review'].fillna('', inplace=True)
```

```
df2.isna().sum()
```

```
review    0
rating    0
dtype: int64
```

Pre-processing

```
pattern = re.compile('https?://\S+')
sentence = pattern.sub('', sentence)
sentence = re.sub(r'(^|\s)@(\w+)', '', sentence)

# Replace colons ':' with spaces
sentence = sentence.replace(":", " ")

emo = re.compile("["
    u"\U0001F600-\U0001FFFF"
    u"\U0001F300-\U0001F5FF"
    u"\U0001F680-\U0001F6FF"
    u"\U0001F1E0-\U0001F1FF"
    u"\U00002702-\U000027B0"
    u"\U000024C2-\U0001F251"
    "]+", flags=re.UNICODE
)
sentence = emo.sub(r'', sentence)

sentence = sentence.lower()

sentence = re.sub(r"[,.\"\'!@#$%^&*(){}?/;`~:<>+=-]", "", sentence)
sentence = sentence.replace("\n", " ")

tokens = word_tokenize(sentence)
table = str.maketrans('', '', string.punctuation)
stripped = [w.translate(table) for w in tokens]
words = [word for word in stripped if word.isalpha()]
```

Pre-processing is necessary as the raw data contains

- special characters
- tags
- emojis

that might interfere with model's working.

These are removed by using 're' **.sub**() function as well as the **replace**() function.

In addition, sentences are divided into smaller parts called **tokens** to analyze each word in order to

- make sure all are aphabetical values through .**isalpha**() function
- for detecting **stopwords** tokens and removing them.

Modeling

Modeling is a process where a program that finds patterns and make decisions based on previous data.

```python
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import pandas as pd

# Initialize the VADER sentiment analyzer
analyzer = SentimentIntensityAnalyzer()

# Create an empty list to store sentiment scores
sentiment_scores = []
df3 = df2.copy()

# Iterate over each comment in the 'Comment' column
for comment in df3['cleaned_comments']:
    # Analyze sentiment for the comment using VADER
    sentiment_score = analyzer.polarity_scores(comment)

    # Extract the compound sentiment score, which represents overall sentiment
    compound_score = sentiment_score['compound']

    # Append the sentiment score to the list
    sentiment_scores.append(compound_score)

# Create a new column 'review_sentiment' in the DataFrame and store the sentiment scores
df3['review_sentiment'] = sentiment_scores
```

```python
from gensim.models import LdaModel

# Specify the number of topics
num_topics = 4

# Build the LDA model
lda_model = LdaModel(corpus, num_topics=4, id2word=id2word, passes=10, random_state=42)
```

```python
# applying TextBlob to all cleaned reviews
df2['review_sentiment'] = df2['cleaned_comments'].apply(lambda x: TextBlob(str(x)).sentiment.polarity)
df2['review_sentiment']
```

Process Summary

- Create 3 copies of the preprocessed data.
- Apply TextBlob, VADER and LDA for each.
- Mapped each to a sentiment category.
- Create Visualization from patterns recognized by the models.
- For TextBlob and VADER:
  - Word Cloud with Bi-gram
  - Frequency Bar Plot
- For LDA
  - Intertopic Distance Map

Modelling Process

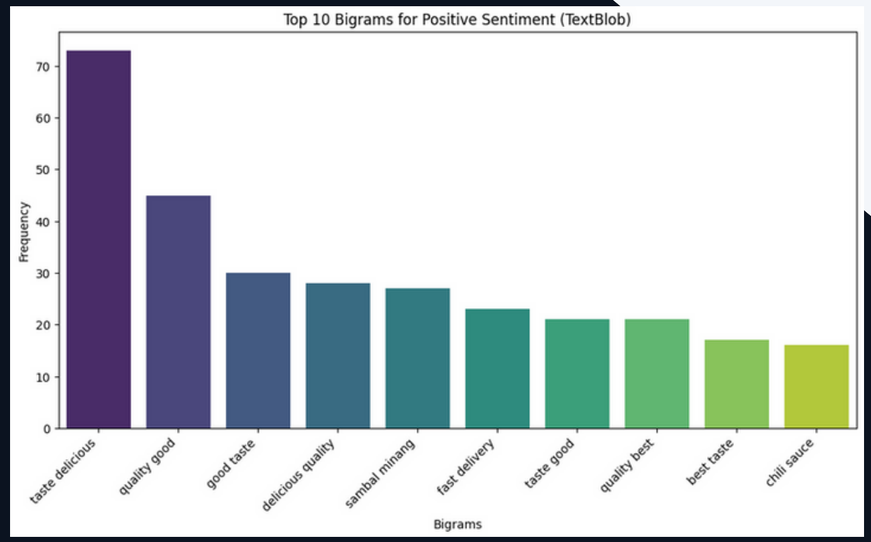
LDA

- Apply LDA model from Gensim for topic modelling
- Use POS tagging to separate just nouns and adjectives from the rest of text
- Create document term matrix for the nouns and adjectives using CountVectorizer
- Convert dtm to tokenized documents to ultimately create the Corpus
- Build LDA model
- Visualize the results through intertopic distance map plotted through pyLDAvis. This gives the frequency all important words in each category topic.

Texblob and VADER

- Create 2 copies of the preprocessed data.
- Apply TextBlob and VADER fpr sentiment analysis
- Obtain sentiment polarity scores for each (between -1 to 1), with -1 being most negative and 1 being positive
- Mapped each score to a sentiment category (i.e. negative, neutral, positive, very positive) to divide scores into distinct intepretable categories
- Visualize the results through sentiment score distribution charts, frequency distribution bar plots for sentiment categories, and worldclouds of bi-grams displaying common words in both positive and negative segment of data

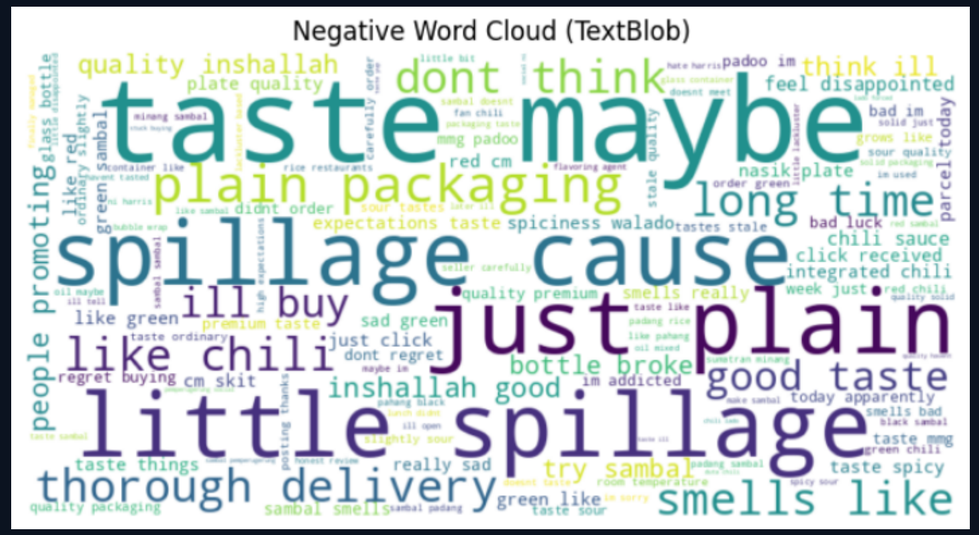
Results using TextBlob

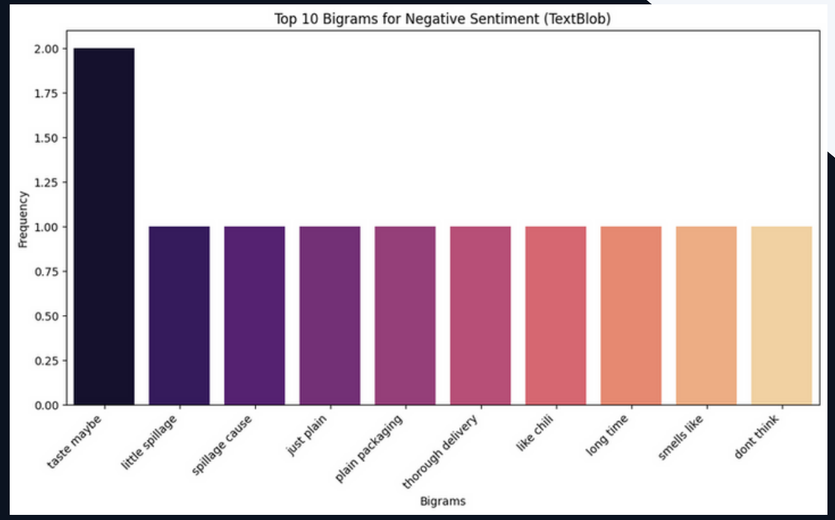


Positive Words and Top 10 Bigram

- The positive word cloud and bigram analysis show a strong customer affinity for taste and quality. Phrases like "taste delicious," "quality good," and "fast delivery" are prevalent, reflecting satisfaction with the product flavors and service efficiency.
- The bigram "best taste" and repeated mentions of "sambal" indicate that the specific product has a favorable reputation among customers, possibly driving repeat purchases.

Results using TextBlob

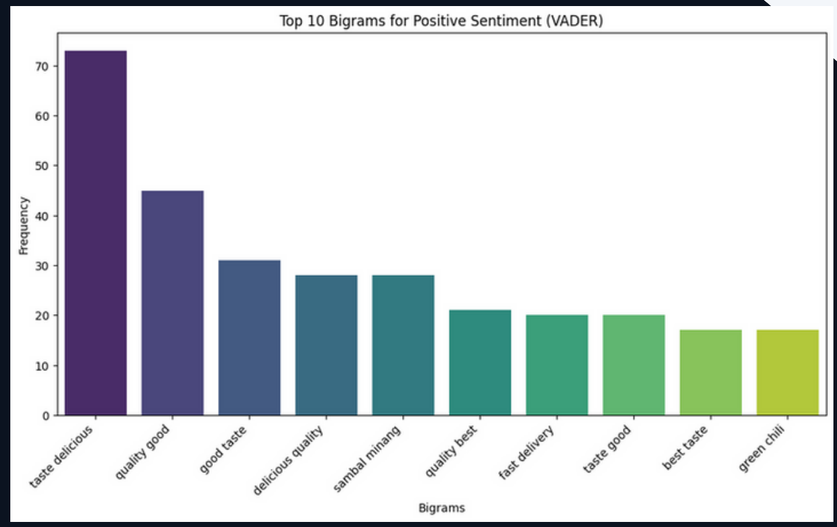




Negative Words and Top 10 Bigram

- On the negative side, the issues seem to be more operational rather than product-related. Terms like "spillage, "bottle broke", "plain packaging," and "little spillage" suggest that while the product itself is well-received, there are concerns about how it's delivered or packaged.
- Phrases such as "taste maybe" "smells bad" and "just plain" hint at a minority of customers who may have had expectations of the product that were not fully met.

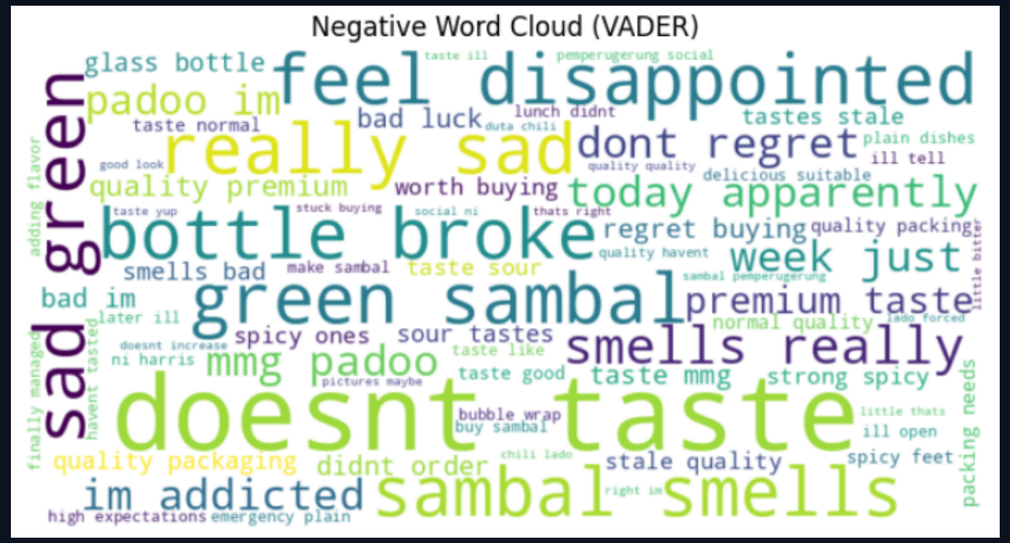
Results using Vader

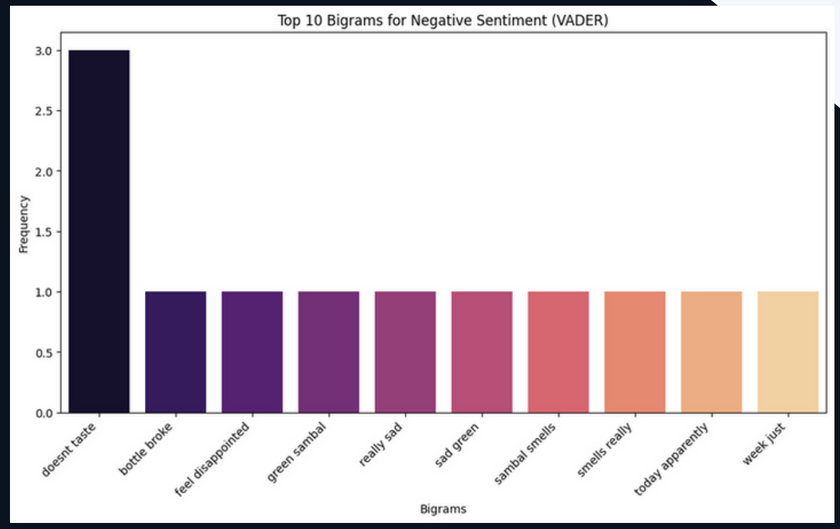


Positive Words and Top 10 Bigram

- The positive word cloud is dense with terms such as "quality," "delicious," "fast delivery," and "good taste," all of which are crucial selling points. The frequency of these terms indicates a high level of satisfaction with the product itself.
- The bigram frequency chart reinforces this, with phrases like "taste delicious" and "fast delivery" suggesting that the product's flavor profile and the speed of service are well-received by customers.

Results using Vader





Negative Words and Top 10 Bigram

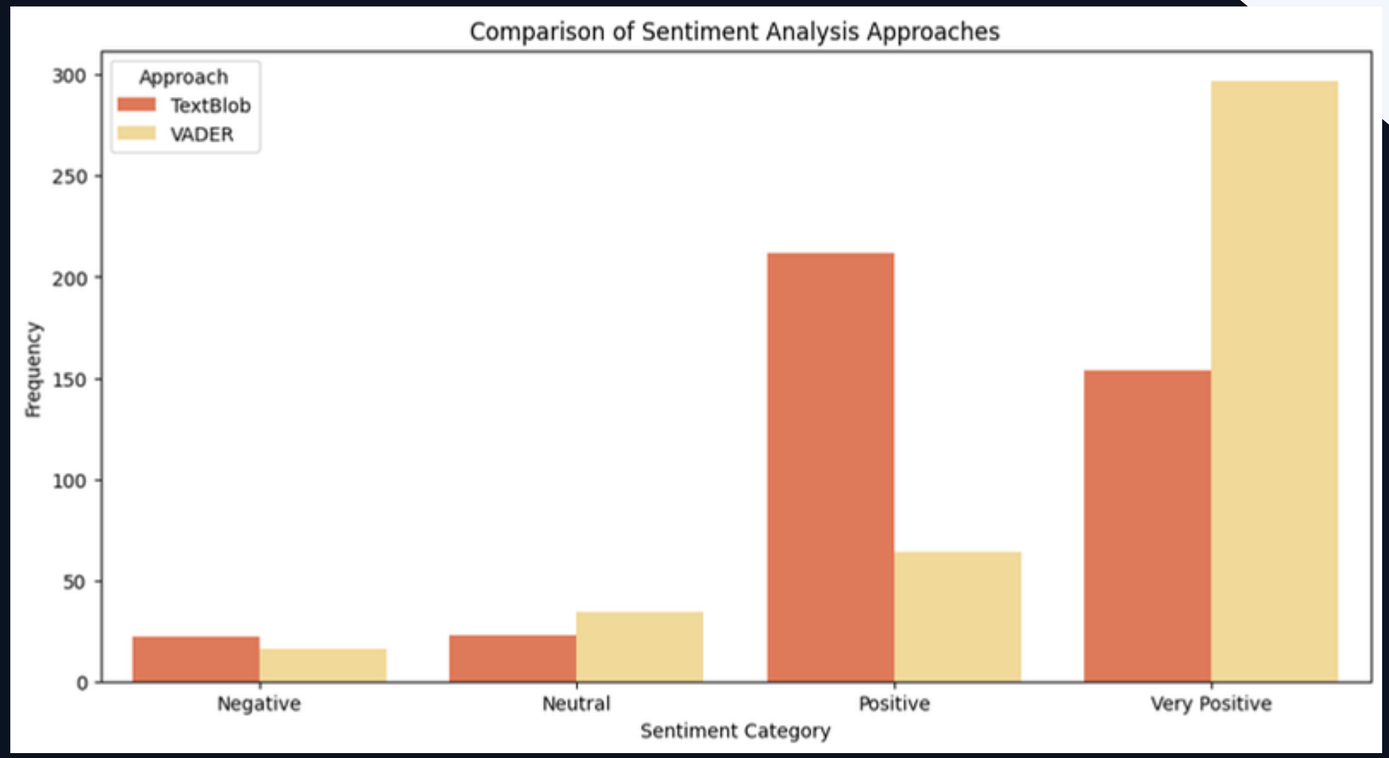
- On the other hand, the negative word cloud shows a prevalence of concerns around "spillage" and "packaging," which could be indicative of issues in the product's delivery and handling.
- Bigrams like “doesnt taste” “sambal smells” “sour tastes” indicate the taste is not as expected for certain customers and the smell of the sambal in particular is not what they prefer.
- The bigrams "bottle broke" and "feel disappointed" highlight specific areas where the customer experience may be falling short, particularly in terms of the product arriving in an expected condition.

Model Comparison

- The overwhelming difference of both approach is how they assign the sentiment score for positive and very positive.
  - TextBlob has more positive, while VADER has more very positive.

- Both model has similar negative and neutral category.

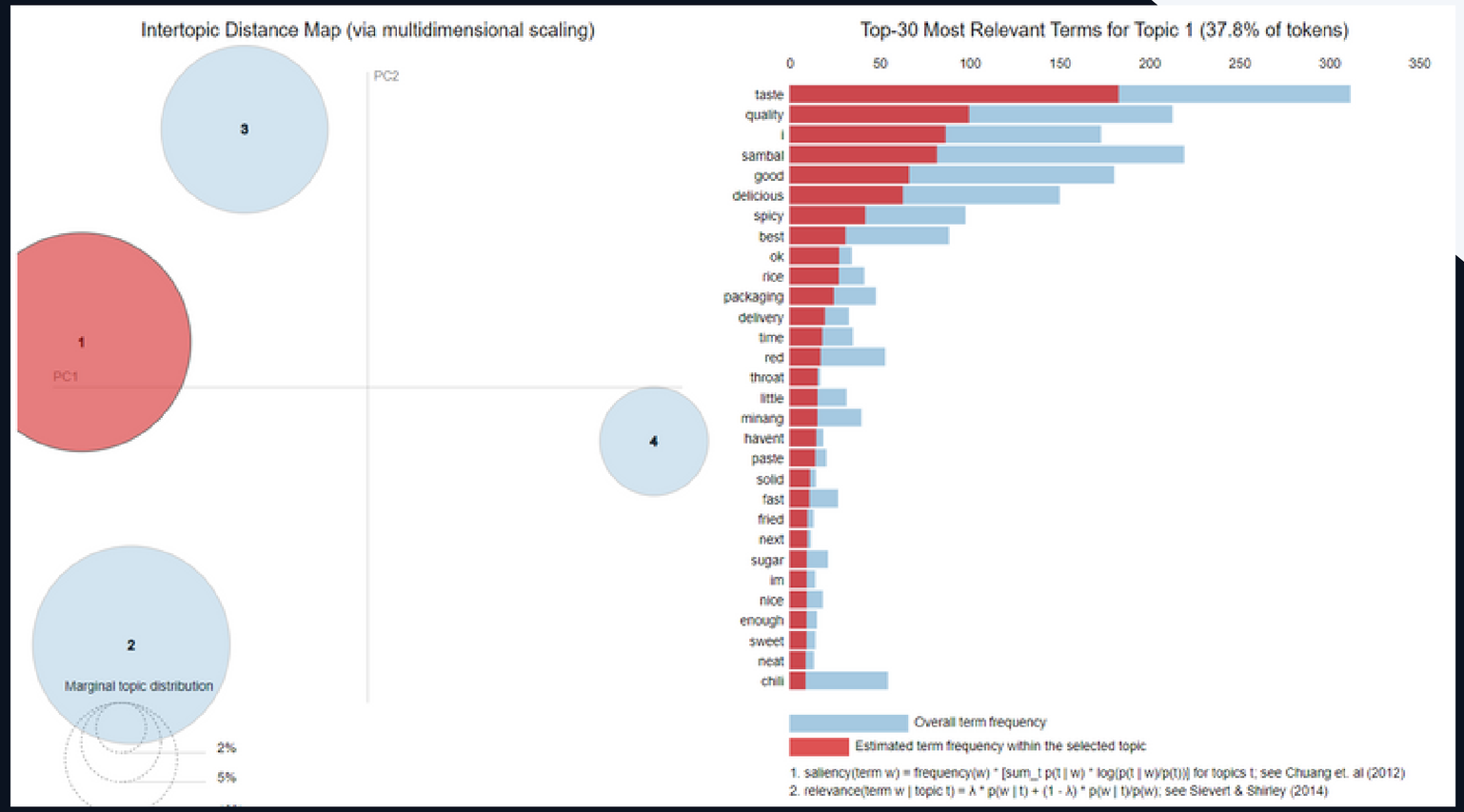
- Both model shows the majority of the comments are positive and very positive.



Results using Intertopic Distance Map

Topic 1 results

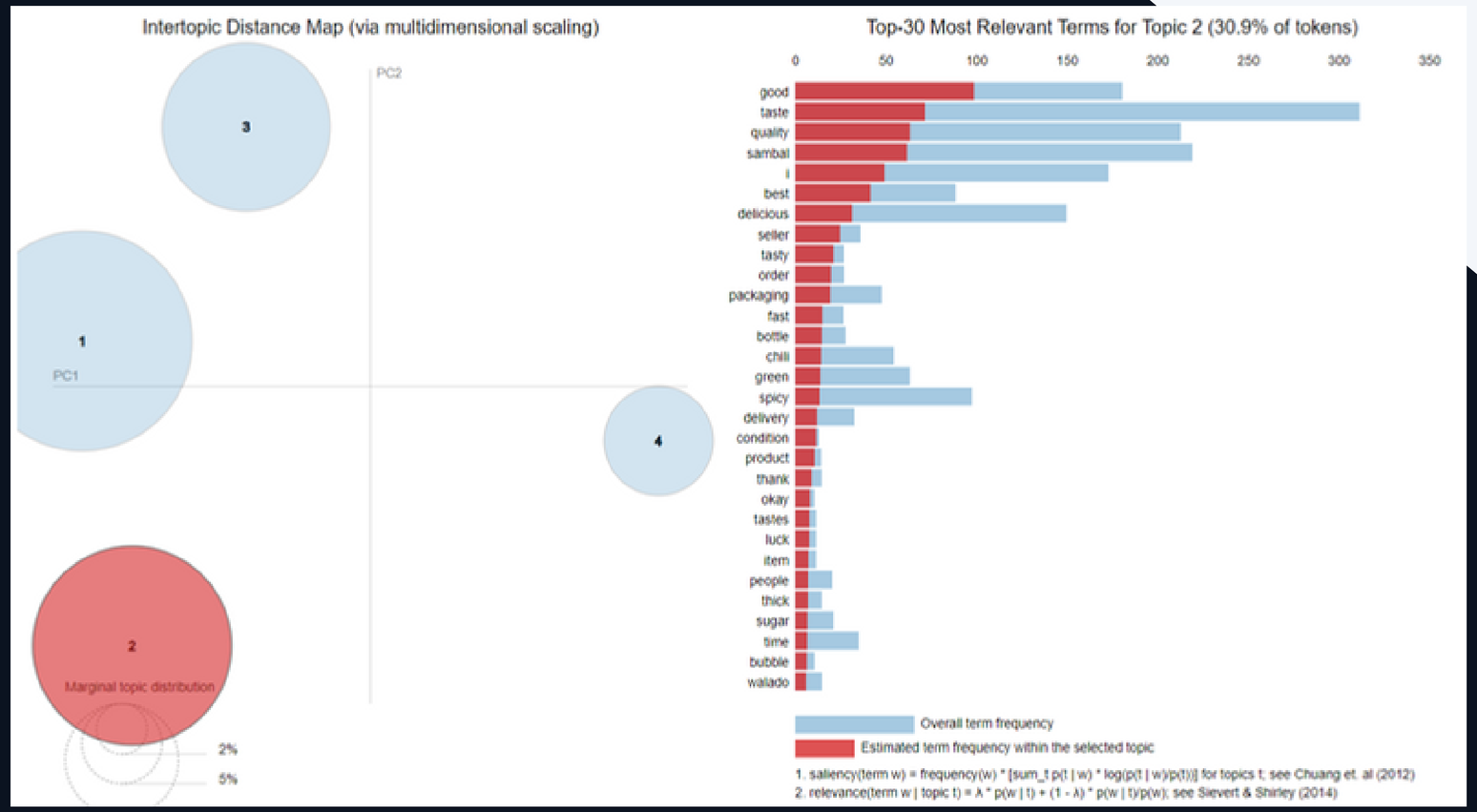
- For Topic 1 as shown in the bar graph to the left a majority of the frequency contains phrases like “taste”, “good” and “delicious”.
- This could indicate that a majority of the customer commented on their satisfaction upon tasting Cillado’s products
- As shown on the intertopic distance map for this topic the topics found by the LDA model shows a distinct and non-overlapping clusters. The model is able to diverse the topic and assign unique terms for each



Results using Intertopic Distance Map

Topic 2 results

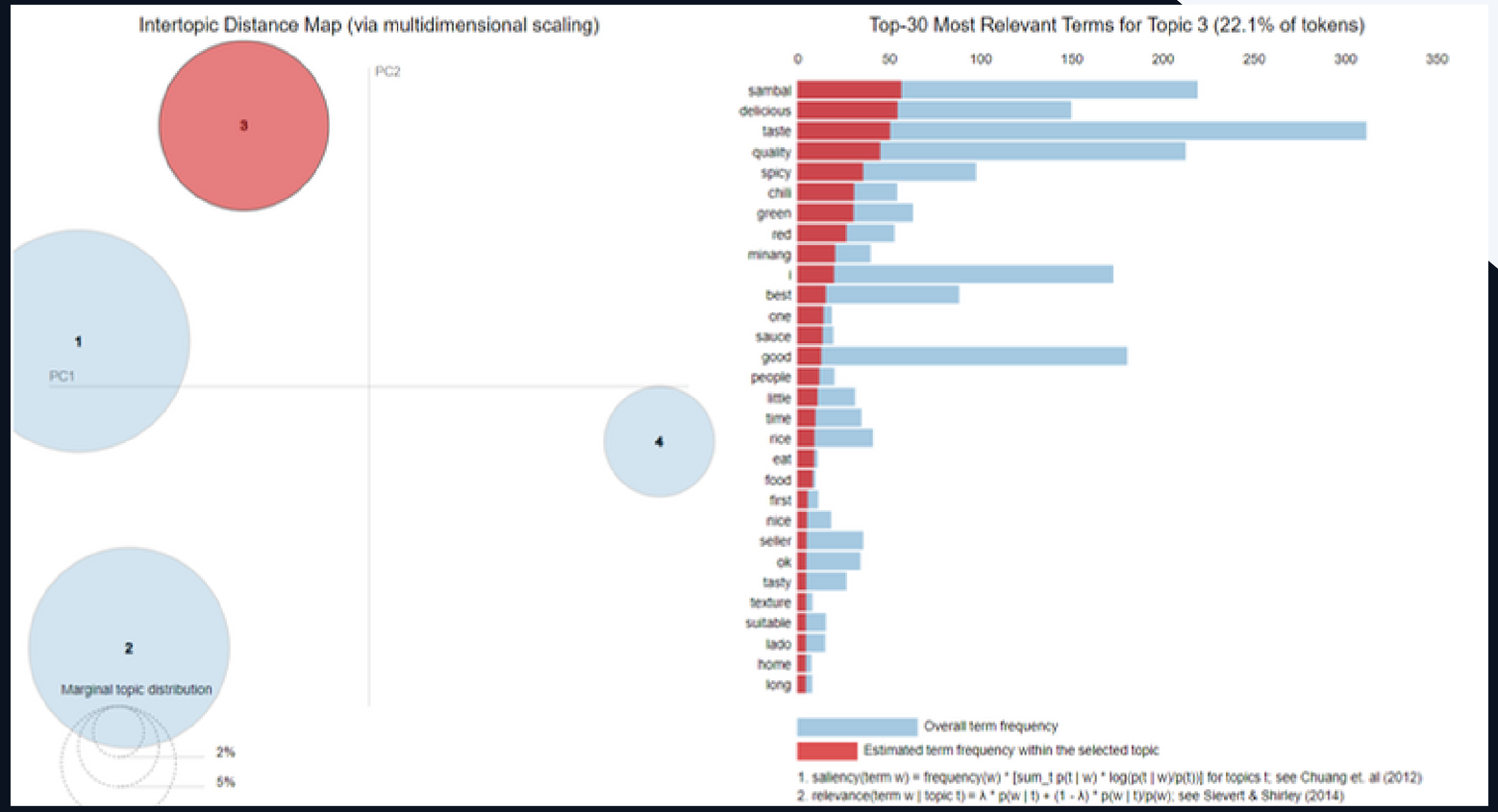
- For topic 2 as shown in the bar graph to the left there also seems to be a high frequecny of phrases like “quality”, “packaging” and “conditions”.
- This could indicate that there is a high number of customers who commented on the overall condition of the product upon receiving it
- As shown on the intertopic distance map for this topic found by the model are unique, distinct and non-correlating. The distance from other clusters could be due to the mention of packaging and shipping related keywords.



Results using Intertopic Distance Map

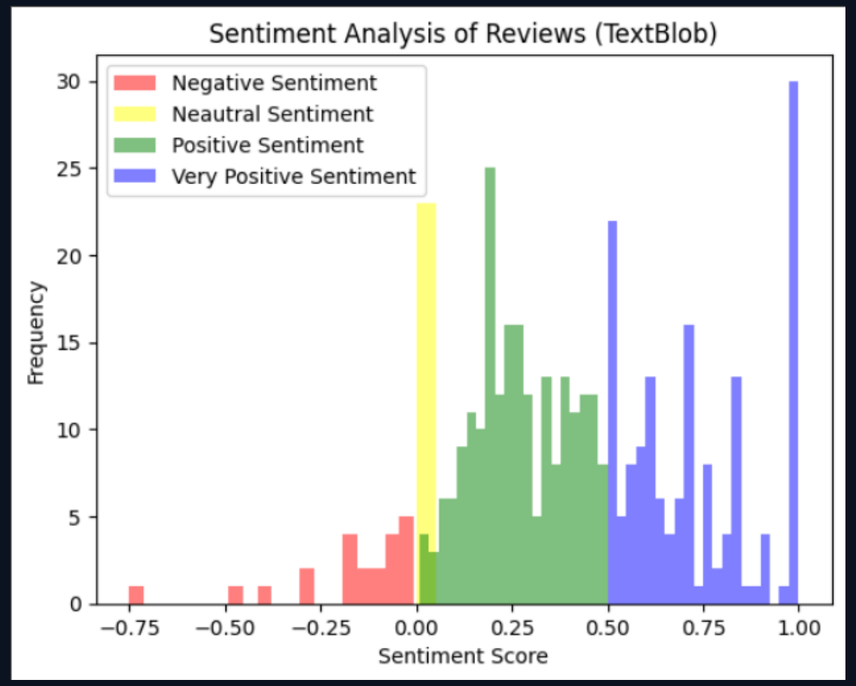
Topic 3 results

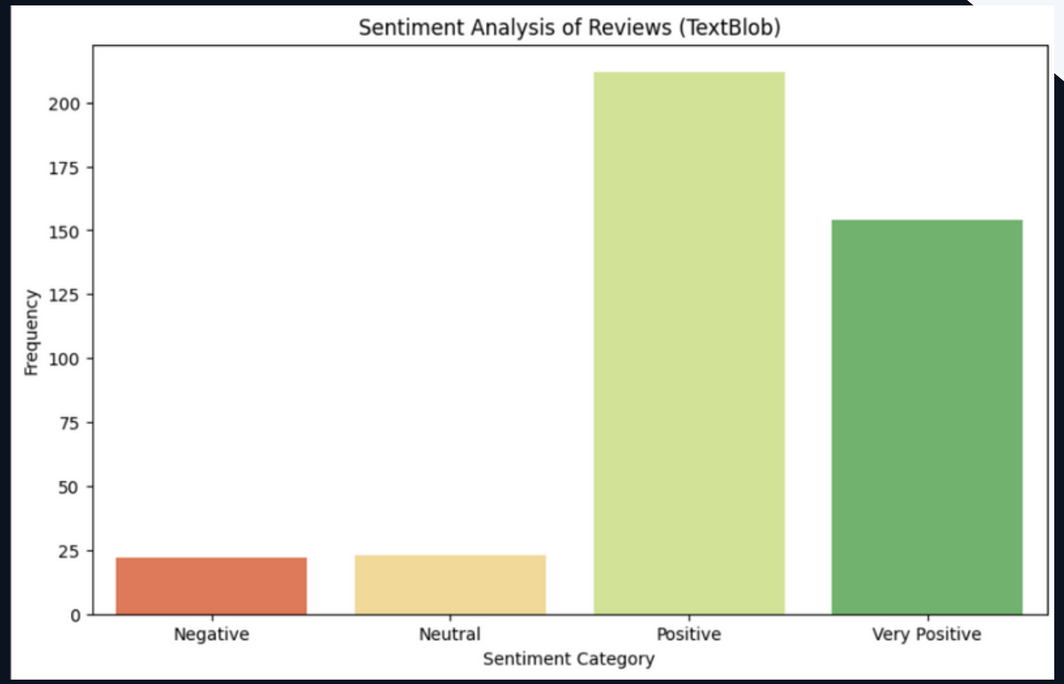
- For topic 3, important terms such as “delicious”, “rice”, “eat”, “texture” and “suitable” are present.
- These can indicate that the sambal is delicious with rice and have a good texture suitable with side dishes.
- The cluster for Topic 3 does not overlap, therefore, it is a distinct topic from others. It is closer to Topic 1, which is most likely due to its thematic relatedness where both topics are talking about the quality of the sambal.



Result & Inferences

Positive and negative sentiments





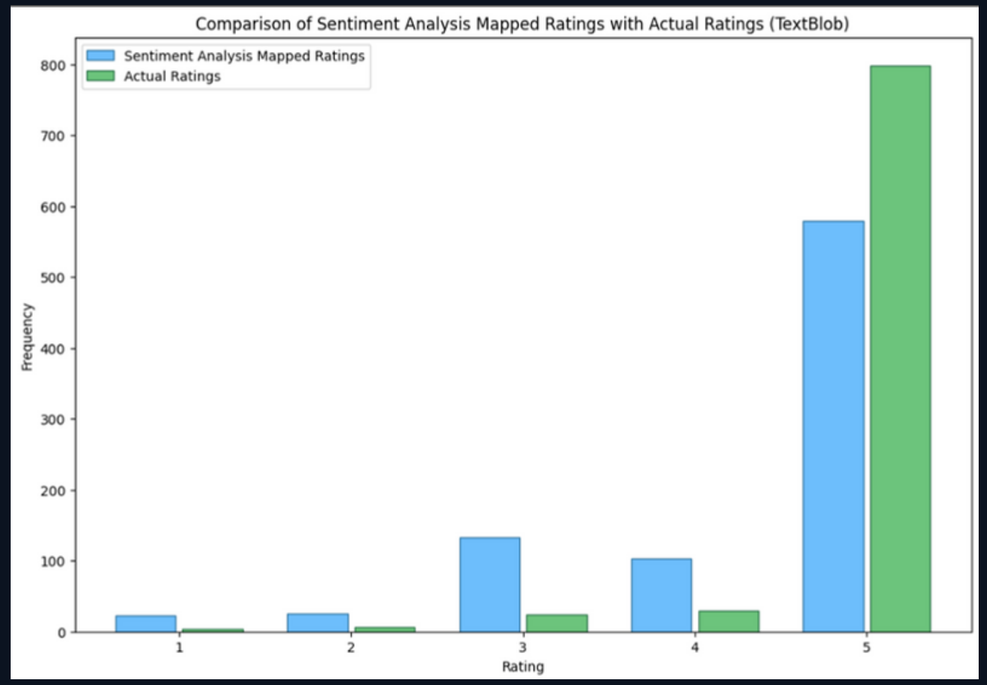
Results from the Sentiment Analysis

- As show above for both the graph there seems to be a high number of frequency and very postive statement compared compared to number of frequency for negative and neutral reviews.

- This could indicate that there are a higher majority of CiliLado customers are extremely happy with the products they received/bought

Result & Inferences

Ratings



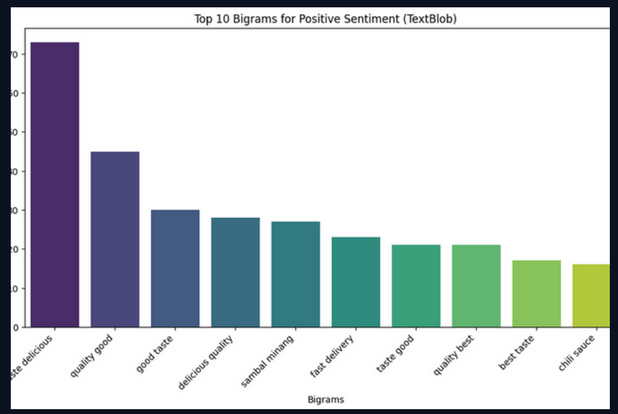
Rating results

- As shown above, the graph show a high number of people who (indetified by our Sentiment analysis) gave a overall postive comments and rating of 5-stars followed by 4-stars

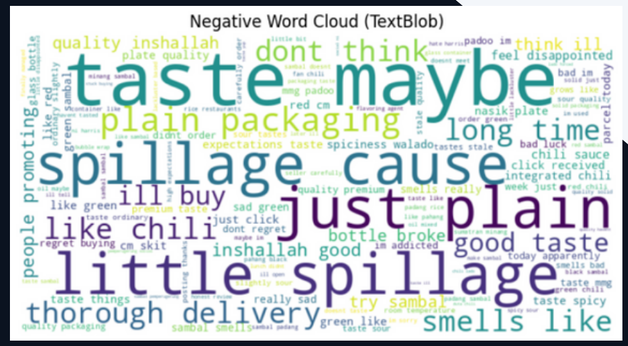
- This further proves that the majority of customers that ordered from Cililado are satisfied with the product bought

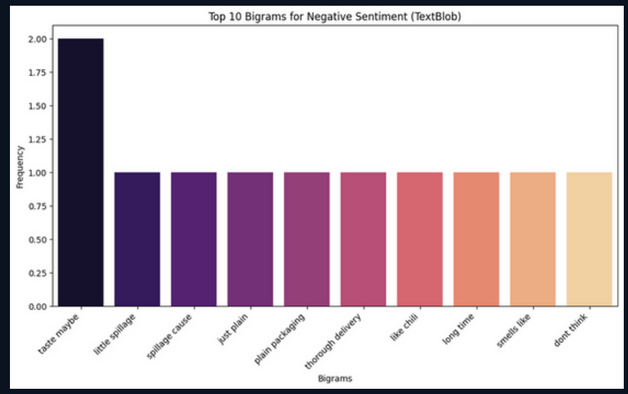
Result & Inferences

Common positive and negative words



Positive and negative reviews left by customer





Negative reviews left by customer

- For positive reviews, phrases such as “delicious taste”, “good quality”, “fast delivery” “best quality”, and many more can be seen left by the customers.
- This indicate that a majority of the positive comments/reviews left by customer are strongly favour the quality and taste of Cililado’s product

- In terms of negative reviews, phrases such as “taste maybe”, “little spillage”, “test plain” “bottle broke” “sambal smells” “sour taste” and others are seen commented by the customer.
- This shows that to some customers the product may not be up to their standard and mey require some revising

Recommendations using the positive and negative words

1. Emphasize the overwhelmingly positive response to the product's taste and quality in marketing and promotional strategies to attract new customers.

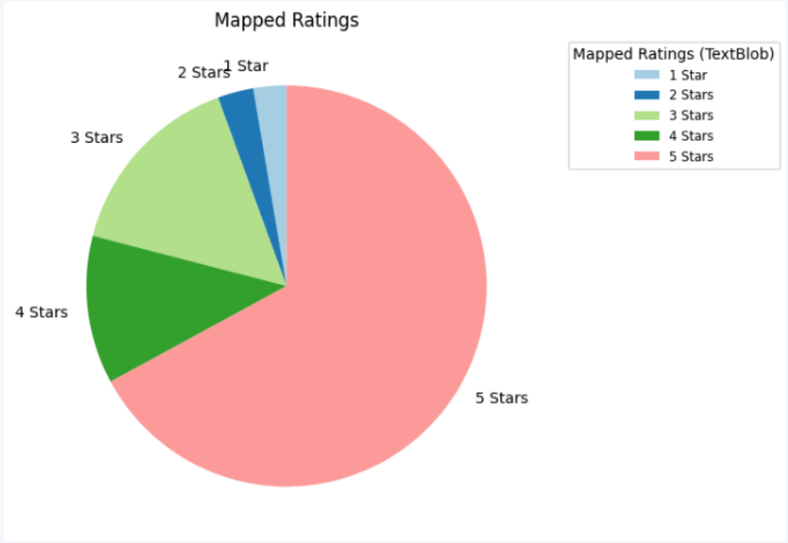
2. Address the packaging and delivery concerns by investing in better packaging solutions to enhance customer satisfaction and reduce negative experiences.

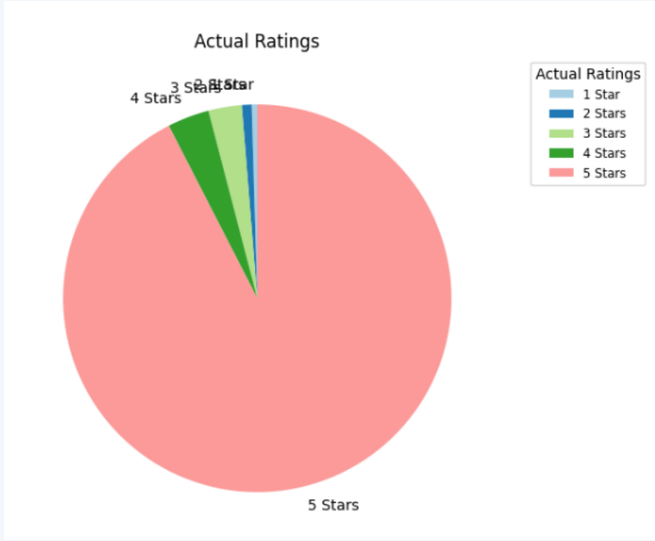
3. Consider introducing a feedback loop where customers can directly report their issues, ensuring that their concerns are heard and addressed swiftly.

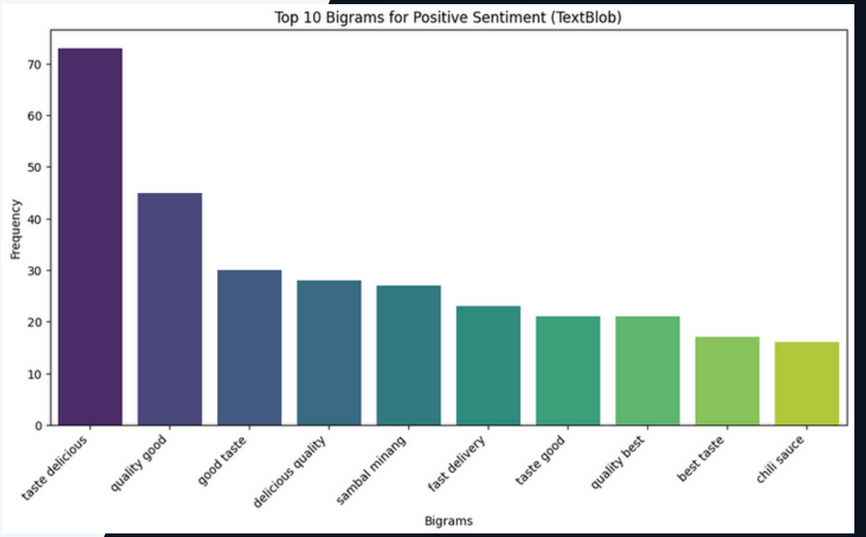
4. Explore opportunities to create product variations that cater to the lesser-satisfied customers, potentially capturing a wider market segment.

Recommendations

Is Shopee a good platform?







1. When comparing the sentiment analysis graphs from earlier to the the pie charts shown above, the 5-star rating decreased, while the others are increased. But, since the majority are still positive, Shopee's review affirms customer's satisfaction.

2. Another reason, as shown in the bar graph one of the more popular positive comments left by customer is "fast deliver". This further proves that Shopee is a good/reliable platform

Conclusion

Through the use of Sentiment Analysis and Topic Modelling, we have identified

- The discrepancy between sentiment analysis and actual customer ratings highlights potential gaps in Cililado understanding of customer needs
- What negative aspects go against customer satisfaction, and what positive aspects meet them
- The overall sentiment distribution of the reviews which will assist in understanding customers and targeted campaign.
- the topics that are persistent in the reviews, which holds importance for targeted campaigns

Through this, we hope that we can aid them overcoming these challenges and ultimately enhancing their overall customer experience as well as fostering customer loyalty and retention for their growth as a business

Thank You

For Your Attention

And wishing Cililado the best on their future ventures!