SENTIMENTS ANALYSIS OF WHICH STATE IN MALAYSIA THAT HAS THE BEST FOOD

Ahmad Azeem Bin Anuar

Bachelor of Computer Science with Honors, Universiti Utara Malaysia,
ahmad_azeem_anuar@soc.uum.edu.my

ABSTRACT. This study investigates the sentiment of the Malaysian community regarding which state is perceived to have the best food, utilizing Orange Data Mining with multilingual sentiment analysis techniques. By analyzing data collected from Reddit, the research aims to classify user sentiments and emotional responses towards the culinary offerings of various Malaysian states. The sentiment analysis results are visualized through box plots and scatter plots, providing a comprehensive view of the sentiment landscape. Key findings highlight the diverse perceptions and preferences within the community, offering valuable insights for stakeholders such as policymakers, tourism boards, and local businesses. This study contributes to a deeper understanding of Malaysian culinary culture and aids in making informed decisions related to food tourism and promotion.

Keywords: Sentiment Analysis, Orange Data Mining, Multilingual Sentiment, VADER (Valence Aware Dictionary and sEntiment Reasoner)

1. INTRODUCTION

1.1. BACKGROUND

Malaysia, a melting pot of cultures and ethnicities, is renowned for its rich and diverse culinary heritage. Each state boasts unique dishes and flavors that reflect its cultural history and local ingredients. From Penang's famed street food to Kelantan's traditional Malay cuisine, the gastronomic landscape of Malaysia is a testament to its multicultural fabric. Food not only plays a central role in Malaysian daily life but also serves as a significant draw for tourists and a source of national pride.

In such a vibrant culinary environment, opinions about which state offers the best food are varied and passionately held. These opinions are increasingly shared and debated on social media platforms, where Malaysians and food enthusiasts from around the world express their preferences and experiences. Understanding these sentiments can provide deeper insights into regional food perceptions and help stakeholders in the food and tourism industries make informed decisions.

Sentiment analysis, a subfield of natural language processing, involves the computational study of people's opinions, sentiments, and emotions expressed in text. It is particularly useful in analyzing large volumes of unstructured data, such as social media posts, to identify trends and patterns in public opinion. This study leverages sentiment analysis to gauge the emotional responses and preferences of the Malaysian community regarding the best food in various states.

To conduct this analysis, we use Orange Data Mining, an open-source data visualization and analysis tool that supports a wide range of data mining and machine learning techniques. Orange provides an intuitive, visual programming interface that allows users to construct data analysis workflows through the manipulation of widgets. These widgets can handle tasks such as data preprocessing, feature extraction, and model evaluation, making Orange a versatile tool for our sentiment analysis needs.

Orange's capabilities in text mining and multilingual sentiment analysis are particularly relevant for our study, as they enable the analysis of social media posts in multiple languages. Malaysia's linguistic diversity means that opinions about food can be expressed in Malay, English, Chinese, Tamil, and other languages. By using Orange, we can effectively analyze this multilingual data to obtain a comprehensive view of public sentiment.

The data for this study is collected from Reddit, a popular online platform where users discuss various topics, including food. Reddit's diverse user base and extensive discussions provide a rich source of data for sentiment analysis. By classifying user sentiments and emotional responses towards the culinary offerings of different Malaysian states, we aim to visualize the sentiment landscape through box plots and scatter plots. These visualizations will help identify trends and preferences, offering valuable insights for stakeholders such as policymakers, tourism boards, and local businesses.

1.2. LITERATURE REVIEW

The field of sentiment analysis, particularly in the context of food, has garnered significant attention in recent years. Numerous studies have explored how sentiment analysis can be applied to understand public opinion and preferences regarding various aspects of food, including quality, taste, and cultural significance. For instance, a study by Asghar et al. (2019) investigated consumer sentiments towards different cuisines using Twitter data, highlighting the effectiveness of social media platforms in capturing real-time public opinions. Similarly, another study by Gohil et al. (2018) used sentiment analysis to evaluate customer reviews of restaurants on Yelp, providing insights into factors that influence customer satisfaction.

In the Malaysian context, research on food sentiment analysis is still emerging. However, there have been notable efforts to analyze public sentiment towards Malaysian food. A study by Ahmad et al. (2020) utilized Facebook and Twitter data to analyze sentiments towards Malaysian traditional foods, revealing positive sentiments and a strong sense of national pride associated with local cuisine. These studies underscore the importance of sentiment analysis in understanding public perceptions and highlight the potential of social media as a rich data source for such analyzes.

1.3. RESEARCH METHODOLOGY

The research methodology for this study involves a systematic process that includes data collection, data cleaning, data preprocessing, and finally, clustering and data visualization. The following sections explain each step-in detail, as illustrated in the provided flowchart:

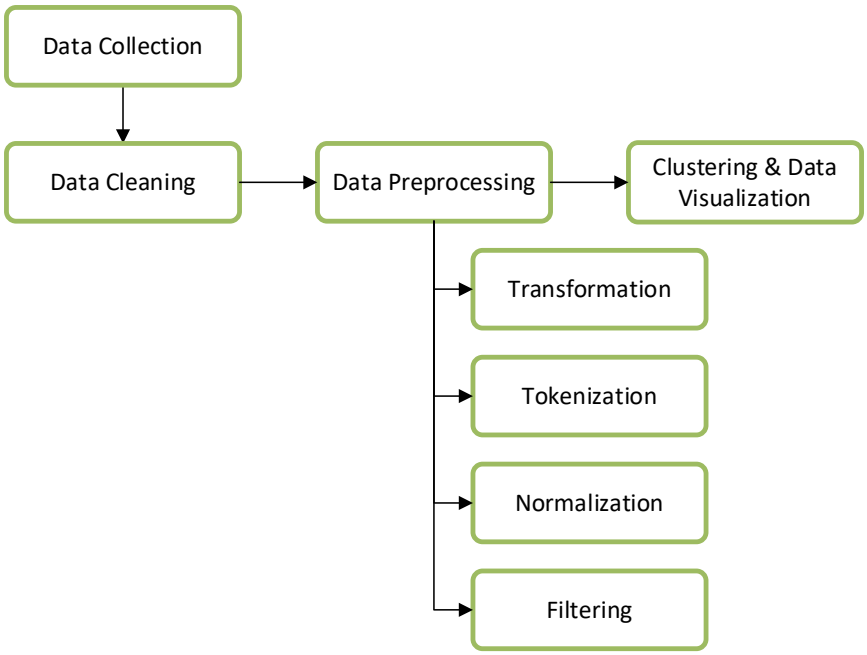

Figure 1. Research Methodology Flow.

a. Data Collection

The first step in the research methodology is data collection. For this study, data is gathered from Reddit, a popular social media platform where users discuss various topics, including food. Reddit's extensive and diverse user base provides a rich source of opinions and sentiments about the best food in different Malaysian states. The data collection process involves scraping relevant posts and comments from Reddit using appropriate tools and techniques to ensure a comprehensive dataset.

b. Data Cleaning

Once the data is collected, it undergoes a data cleaning process. This step is crucial to remove any irrelevant or noisy data that could affect the accuracy of the analysis. Data cleaning involves the following tasks:

- **Removing Duplicates**: Identifying and removing duplicate entries to avoid redundancy.
- **Eliminating Irrelevant Data**: Filtering out posts and comments that are not related to the study's focus on Malaysian food.
- **Handling Missing Values**: Addressing any missing values in the dataset, either by imputing them or excluding them from the analysis.

c. Data Preprocessing

After cleaning the data, the next step is data preprocessing. This step prepares the data for analysis and involves several sub-steps:

- **Transformation**: Converting the raw text data into a suitable format for analysis. This may involve changing text to lowercase, removing special characters, and stemming or lemmatizing words to their base forms.
- **Tokenization**: Breaking down the text into individual tokens, which are the basic units of analysis. Tokenization helps in converting sentences or paragraphs into words or phrases that can be easily analyzed.

- **Normalization**: Standardizing the text data to ensure consistency. This involves processes such as removing stopwords (common words like "and," "the," which do not contribute to the sentiment analysis) and handling contractions (e.g., converting "can't" to "cannot").
- **Filtering**: Applying filters to refine the dataset further. This might include filtering based on specific keywords or phrases related to food and the states of Malaysia.

Determining the class attribute and loading the dictionary involve matching base words with sentiment word dictionaries to determine the sentiment content (positive, neutral, negative). All tweet data is labeled according to classes, with three classes to be used in this study: positive class, negative class, and neutral class.

d. Clustering and Data Visualization.

The final step in the methodology involves clustering and data visualization:

- **Clustering**: Grouping similar data points together to identify patterns and trends. In the context of this study, clustering can help in identifying groups of sentiments related to different states. Techniques such as k-means clustering or hierarchical clustering can be used to achieve this.
- **Data Visualization**: Visualizing the results of the sentiment analysis using tools like box plots and scatter plots. These visualizations provide a clear and interpretable representation of the sentiment landscape, highlighting which states are perceived to have the best food according to the analyzed data.

  Clustering data text mining with Orange Data Mining involves visualizing Box Plot and Scatter Plot, which visualize the processed text mining data with the emotions of Reddit users.

2. RESULT AND DISCUSSION

2.1. RESEARCH SCENARIO

The application of Orange Data Mining showcases the interface design of sentiment analysis widget integrated into the workflow, as illustrated in Figure 2 below:

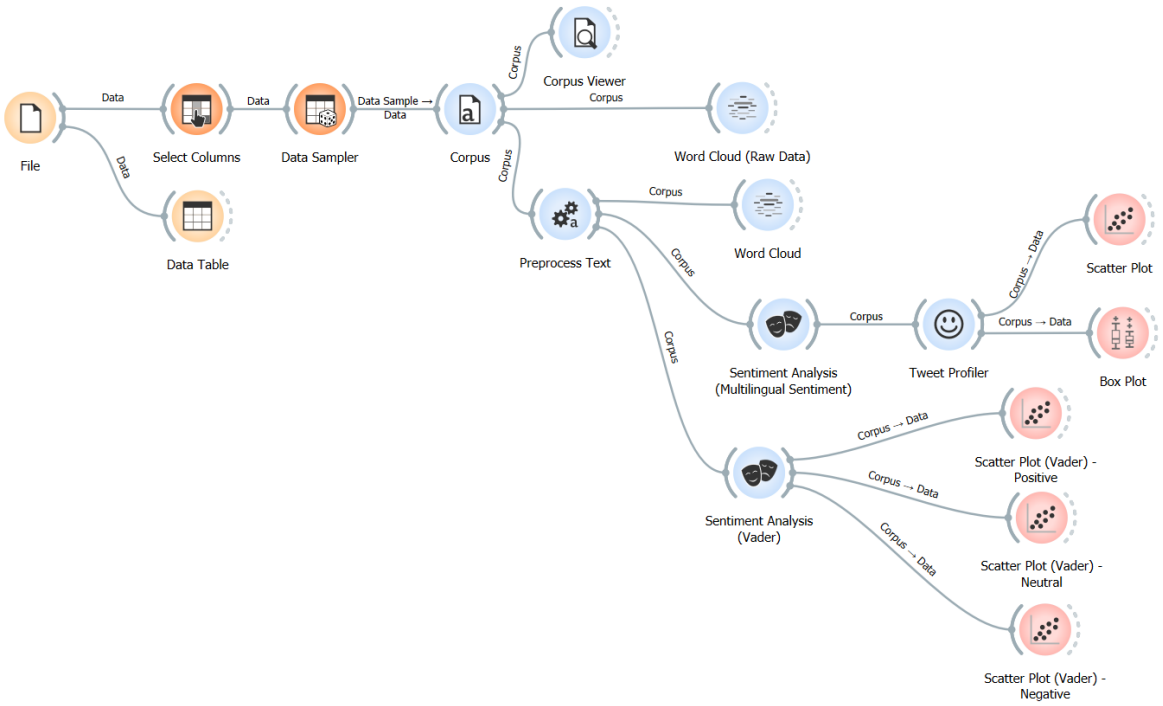

Figure 2. Sentiment Analysis Widget Data.

The data collection was scrapped from the social media platform Reddit by using the Export Comments website as it shown in the Figure 3. The data will be input and analyzed individually based on the objects. Subsequently, it will be connected to the necessary widgets for research purposes resulting in a widget design shown in the figure above.

2.2. DATA COLLECTION

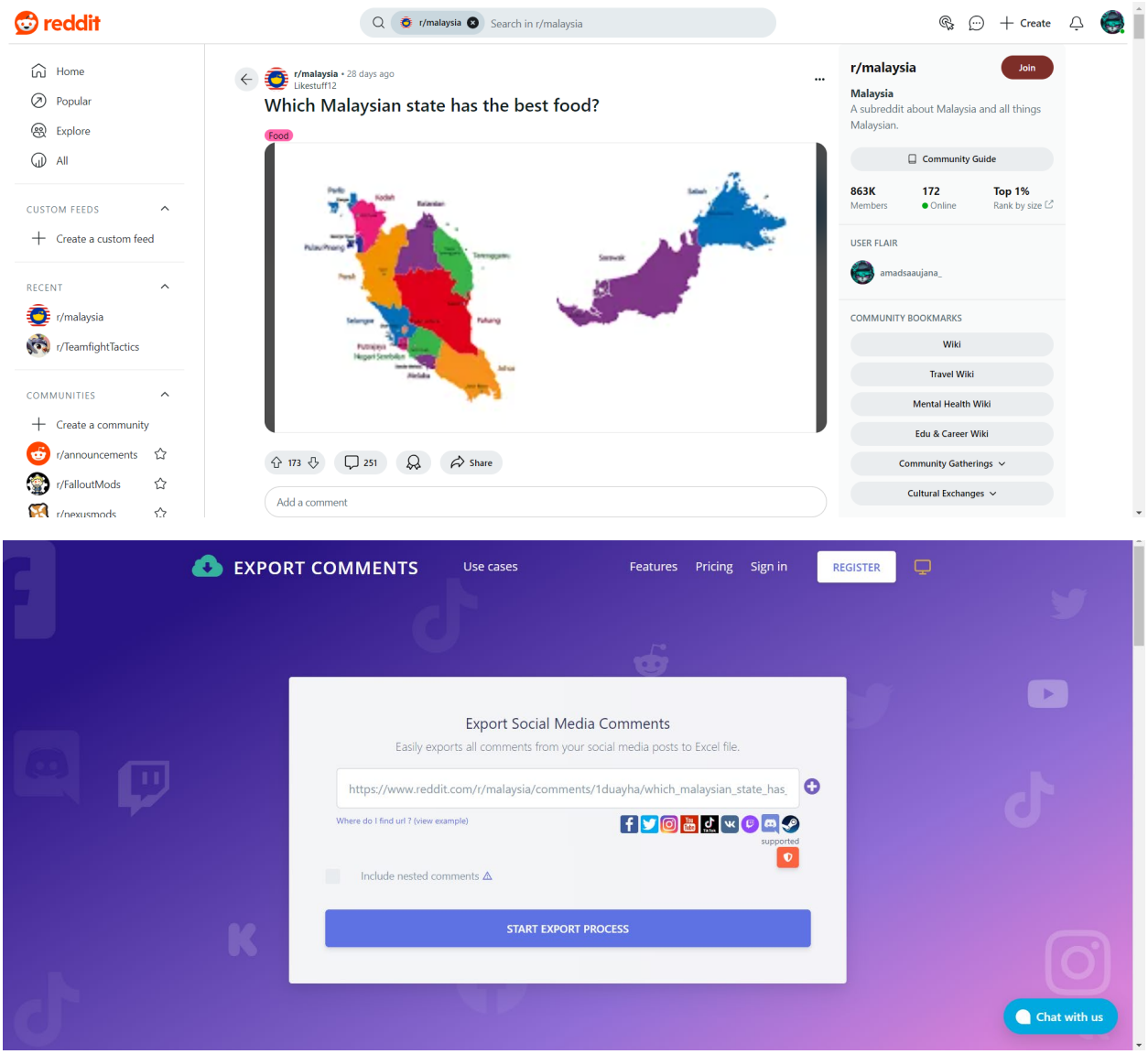

Figure 3. The Reddit thread and the Export Comment website.

In this study, the research data consists of comments from the Malaysian society on Reddit regarding which Malaysian state has the best food from 7 July 2024. The dataset for this research was obtained by exporting the comments using the Export Comments website as depicted in Figure 3. From the data collection results based on the exported comments, 185 comments were obtained. Subsequently, the file is imported into Orange Data Mining as shown in Figure 4.

The imported dataset will have the necessary columns selected according to the requirements of sentiment analysis research. However, to do commit further we had to make a data cleaning to choose the appropriate column so the imported dataset can be read properly by the Orange.

2.3. DATA CLEANING

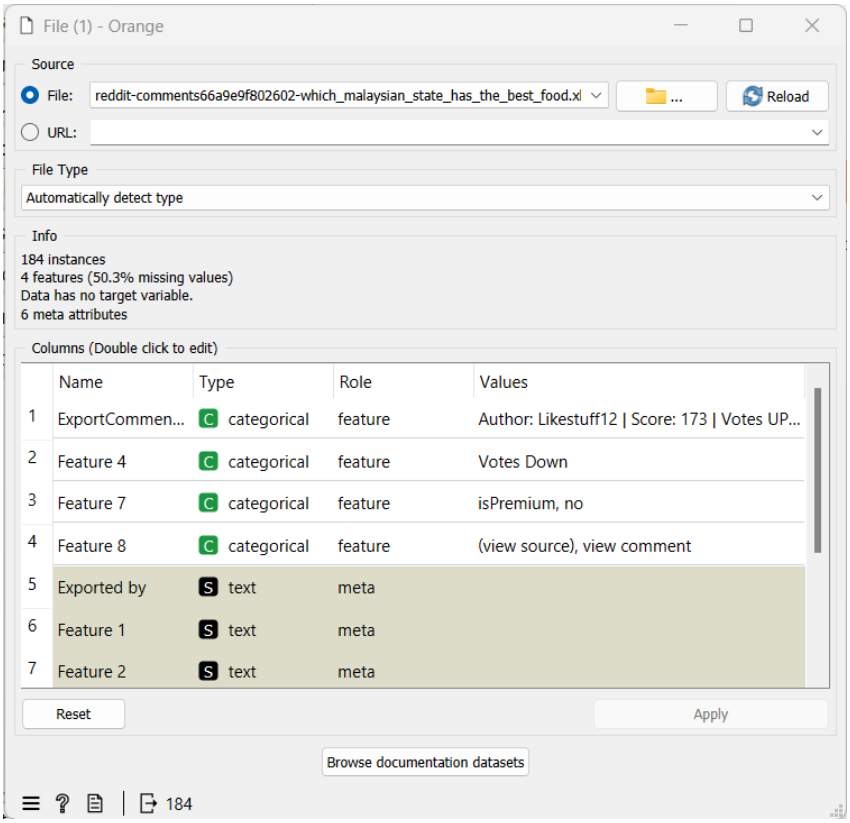

Figure 4. Import Dataset.

Data cleaning is a critical step in the research methodology that ensures the quality and reliability of the data used for analysis. It involves identifying and correcting or removing inaccuracies, inconsistencies, and irrelevant information from the dataset. The following are the key tasks involved in the data cleaning process for this study:

- **Removing Duplicates**: Duplicate entries can arise when the same post or comment is scraped multiple times or when users repeat similar sentiments. These duplicates can skew the analysis by giving undue weight to certain opinions. The cleaning process involves detecting and removing these duplicates to maintain the integrity of the data.
- **Eliminating Irrelevant Data**: Not all data collected from Reddit will be relevant to the study's focus on food preferences in Malaysian states. Irrelevant data, such as posts unrelated to food or discussions about non-Malaysian cuisine, must be filtered out. This step ensures that the analysis remains focused on the research question.
- **Handling Noise in Text**: Social media data often contains noise in the form of typos, slang, abbreviations, and special characters. Cleaning this noise involves standardizing the text to improve the accuracy of subsequent analyses. For instance, converting all text to lowercase helps in consistent processing, while removing special characters prevents them from interfering with text analysis.

- **Addressing Missing Values**: Missing values can occur when certain information is not provided in a post or comment. These gaps can affect the completeness and accuracy of the analysis. The cleaning process involves identifying these missing values and deciding on an appropriate approach to handle them. Options include imputing missing data using statistical methods, removing incomplete entries, or filling in missing values with placeholders if they do not significantly impact the analysis.

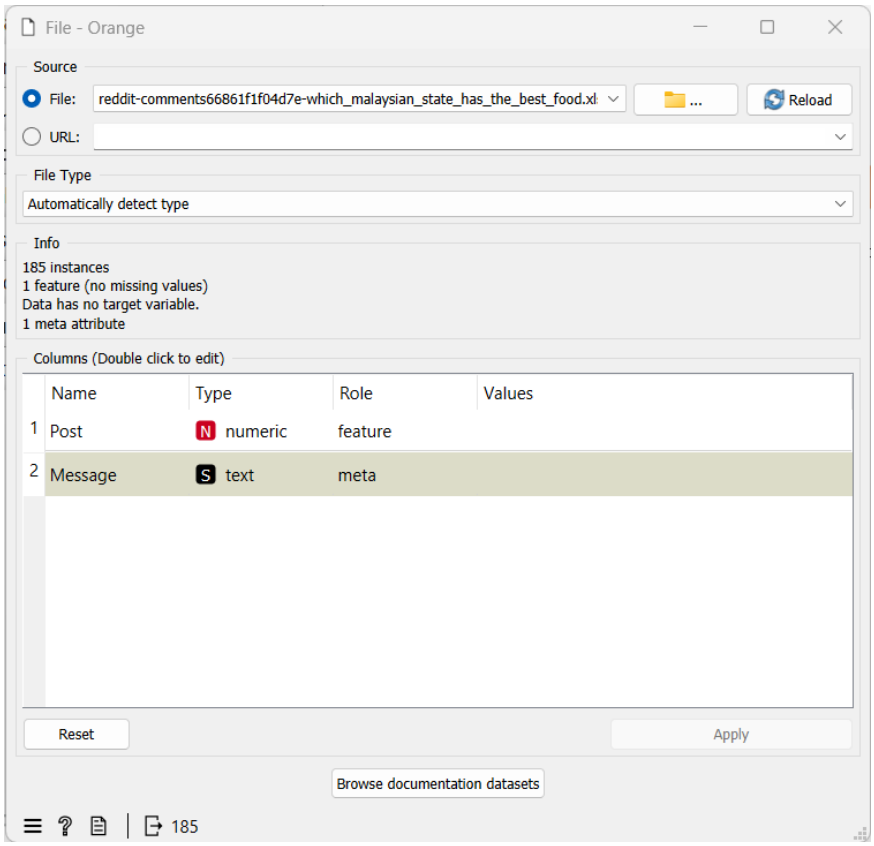

Figure 4. Imported Dataset After Data Cleaning.

2.4. DATA PREPROCESSING

Before conducting text analysis, the text will first undergo preprocessing. This involves segmenting the text into smaller units (tokens), followed by transformation, tokenization, normalization, and filtering. Sequential steps in the analysis can be enabled or disabled within the Preprocess Text widget in Orange Data Mining. Figure 5 below shows the steps performed in the preprocess text widget in the Orange Data Mining application.

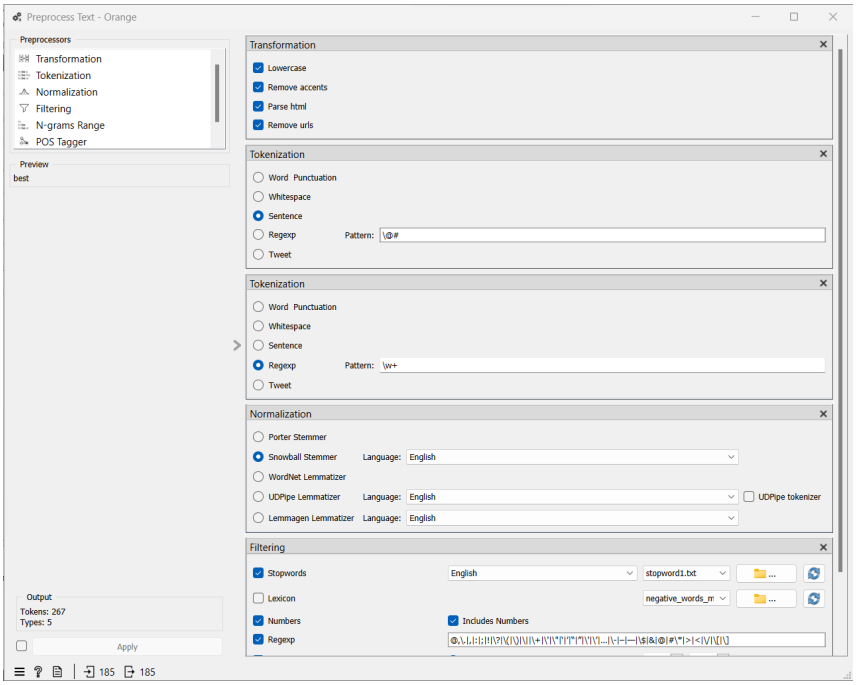

Figure 5. Preprocess Text.

The steps carried out in the preprocess text in the Orange Data Mining application are as follows:

a. Transformation

The first step is transformation, which involves transforming the entire text into lowercase, removing accents contained within the text, identifying HTML tags, parsing HTML tags, and removing URLs from the text.

b. Tokenization

In this stage, sentences will be tokenized into words, preserving punctuation symbols.

c. Filtering

In this stage, a process of removing or preserving selected words will be conducted. During this process, words that are not relevant to sentiment analysis will be removed. All words to be removed have been written into a file named 'stopword1.txt' using the stopword widget. Additionally, the lexicon widget is utilized to extract tokens from the lexicon dictionary. The number widget is employed to remove meaningless numbers, while the regexp widget is used to eliminate tokens based on available regular expression patterns.

Once the preprocess text stage is completed, the text will be separated into individual words, and can observe the text distribution through a word cloud in Figure 6 below.

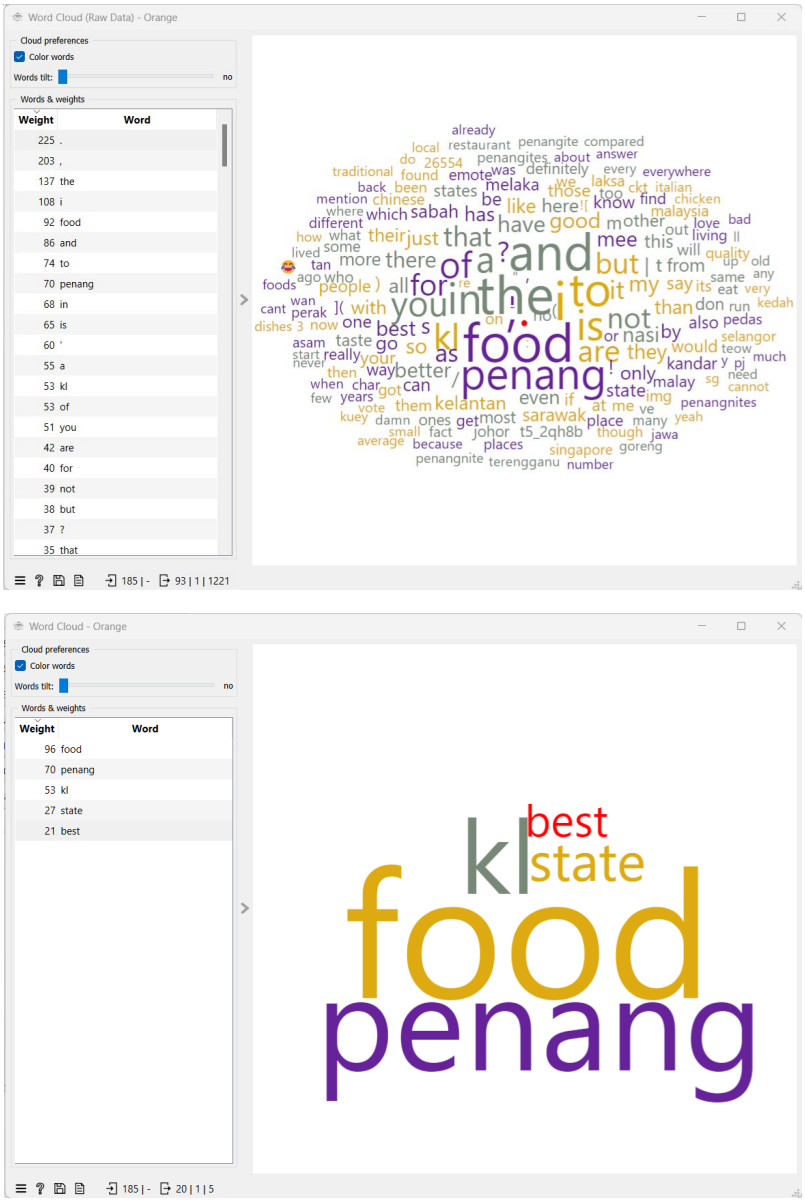

Figure 5. The Word Cloud Before and After Data Preprocessing.

In the visualization above, a word cloud is displayed showing the frequency of the most commonly occurring words. The size of each word in the word cloud corresponds to its frequency of occurrence. The larger the word, the more frequent its occurrence. A word cloud serves as a visualization method to represent the results of text preprocessing. The visual settings and word variations enhance the attractiveness and comprehensibility of the visualization. The image depicts the preprocessing results of information previously comprised of comment lines on Reddit thread about which Malaysian state has the best food.

2.5. SENTIMENT ANALYSIS

The analysis process utilizes the MultiLingual Sentiment algorithm, which is capable of understanding opinions and viewpoints of users in various languages, in this case, using English. According to (Olaleye et al. 2023), the MultiLingual sentiment analyzer has been shown to provide a significant level of invariance compared to traditional sentiment analysis systems, such as the Vader variant, thus enhancing the accuracy and diversity of sentiment analysis. As depicted in Figure 6 below.

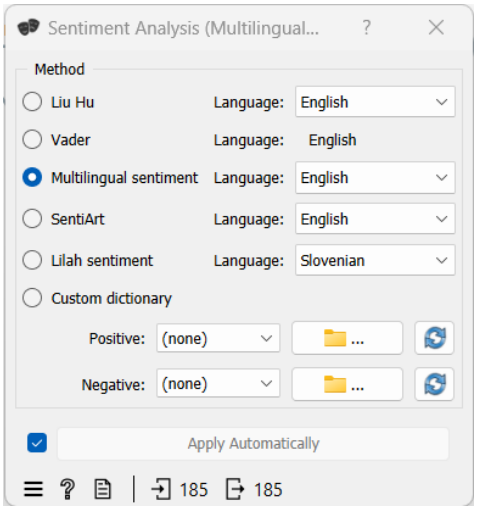

Figure 6. MultiLingual Sentiment Analysis.

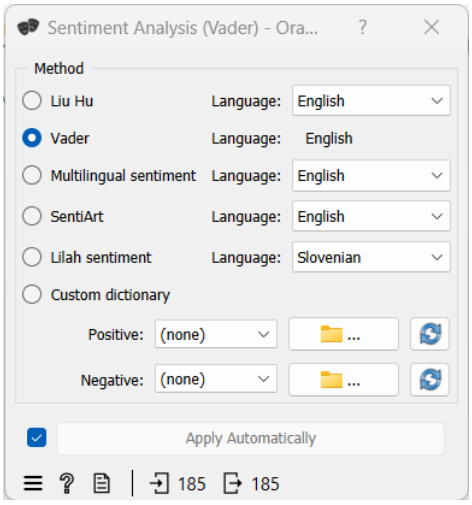

Figure 7. Vader Sentiment Analysis.

The MultiLingual sentiment approach enables a broader and more inclusive sentiment analysis, allowing text processing in various languages to understand the opinions, emotions, or sentiments contained within the text. Therefore, MultiLingual sentiment becomes crucial in the context of globalization and linguistic diversity in sentiment analysis and cross-cultural opinion understanding.

Vader is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is designed to handle the informal language, abbreviations, emoticons, and slang commonly found in social media texts. For this research, which involves analyzing Reddit comments and posts about food preferences in Malaysian states, Vader is a suitable tool due to its robustness and efficiency in processing short, informal text data.

2.6. TWEET PROFILER

Tweet Profiler is one of the features in the Orange Data Mining platform that enables this research to analyze sentiment from tweets or other text documents. By using Tweet Profiler, sentiment data can be retrieved from the available dataset through the server for each given tweet, and sentiment analysis can be conducted using various emotion classification methods provided, such as Ekman, Plutchik, and Profile of Mood States (POMS). Additionally, this feature allows for the utilization of specific attributes for analysis, such as content attributes, and performing emotion classification with multi-class options. Tweet Profiler is a valuable tool in text analysis and sentiment understanding in Orange Data Mining.

In this study, a dataset of 185 comments about the decisions which Malaysian state has the best food was utilized. The data was extracted using a widget from Orange Data Mining with Corpus and connected to Tweet Profiler using Ekman emotion. As depicted in Figure 7 below.

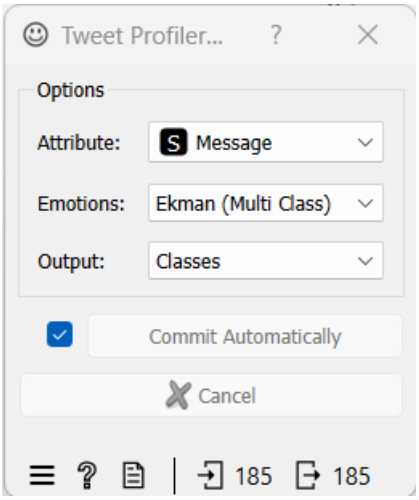

Figure 8. Tweet Profiler.

2.7. DATA VISUALIZATIONS

After performing tweet profiling in the Orange Data Mining widget, the next step is to connect the corpus to visualize the data and observe the results of sentiment analysis research using Box Plot and Scatter Plot.

In the Box Plot data visualization, a diagram is displayed showing the results of 6 emotions: joy, surprise, fear, disgust, sadness, and anger. From these 6 emotions, it can be observed that joy, surprise, and fear are the dominant emotional responses shown by Reddit users. For further details, please refer to Figure 8 below.

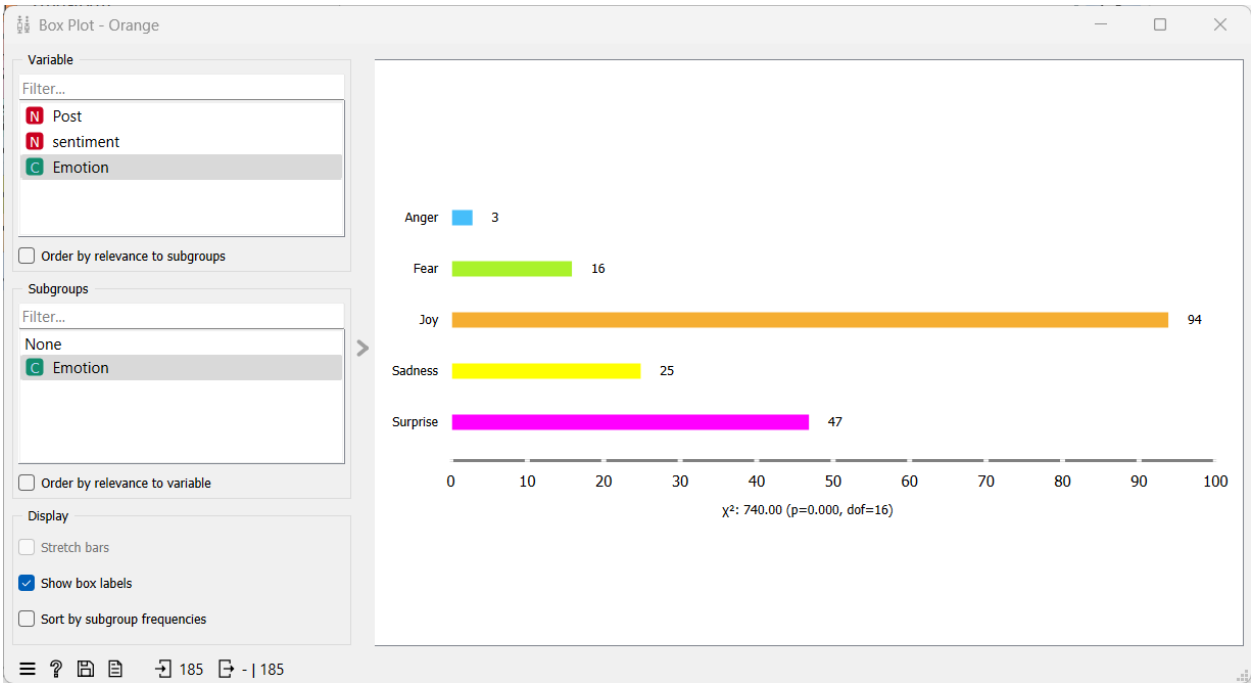

Figure 9. Box Plot Emotion.

From the visualization above, it can be seen that the emotional responses exhibited by Reddit users as follows: the emotion ‘joy’ totals 94, the emotion ‘surprise’ totals 47, the emotion ‘sadness’ totals 25, the emotion ‘fear’ totals 16, and finally the emotion ‘anger’ totals 3.

Data visualization can also be observed through Scatter Plots to visualize patterns or relationships between two variables, such as positive correlation, negative correlation, or no correlation at all. In a scatter plot, each point represents one observation, where one axis indicates the value of one variable and the other axis indicates the value of another variable. In this study, the variables used are emotion and sentiment variables.

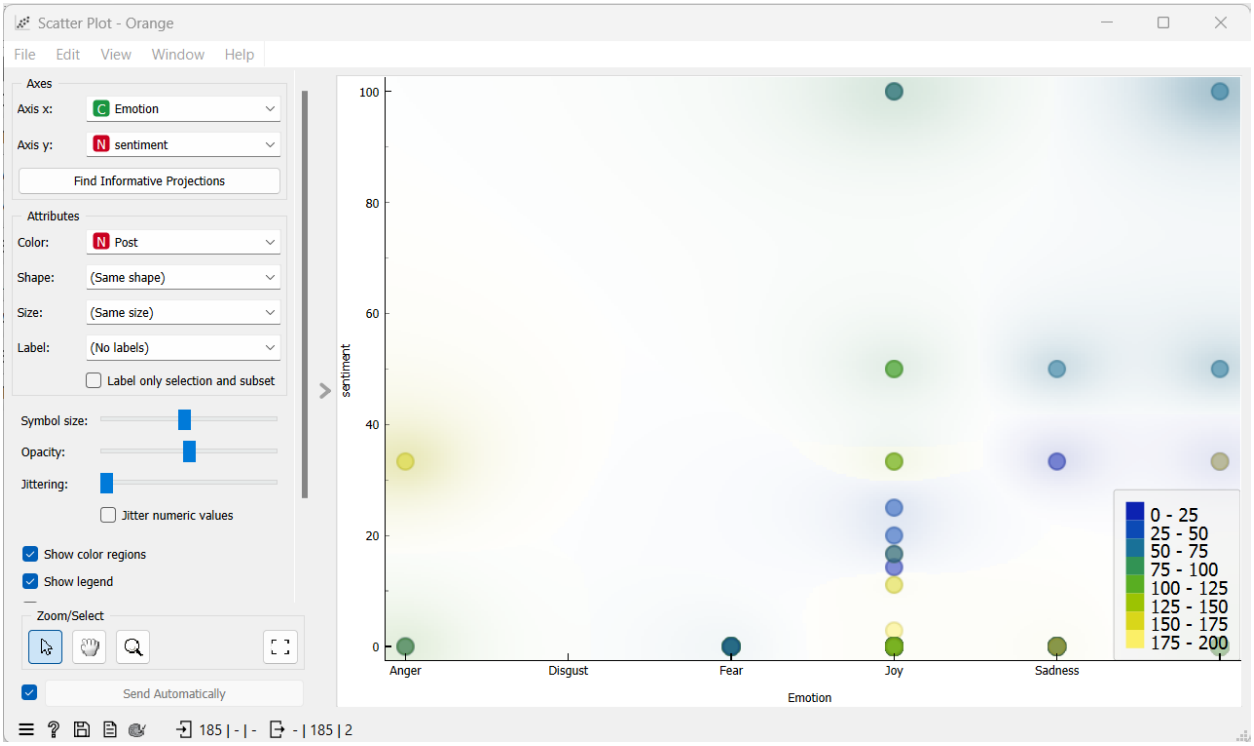

Figure 10. Scatter Plot for Emotion.

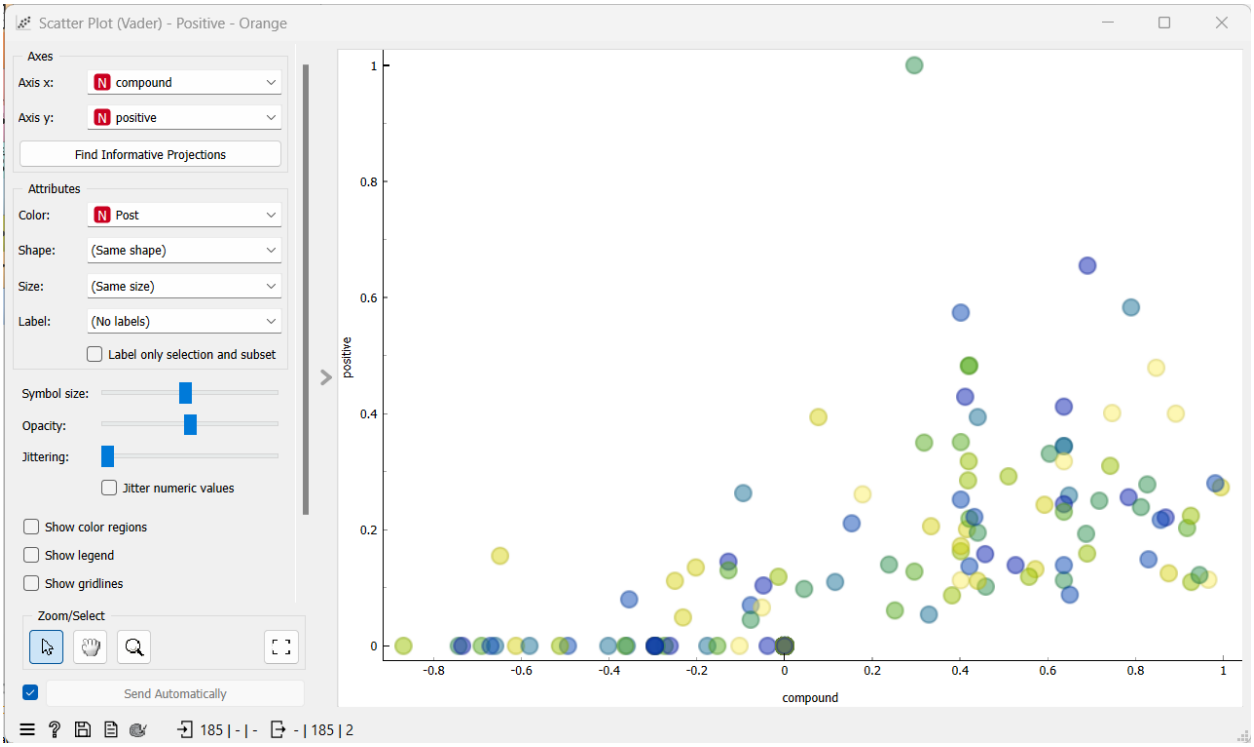

Figure 11. Scatter Plot for Positive Responses Using Vader.

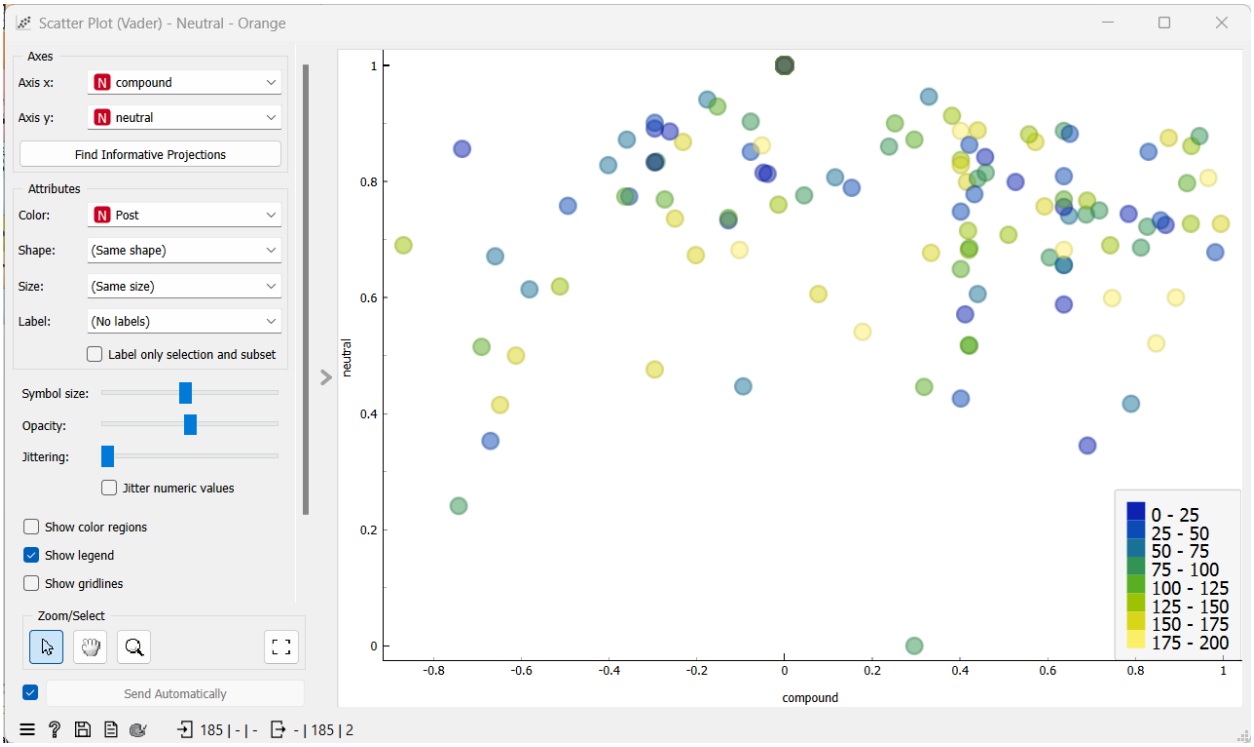

Figure 12. Scatter Plot for Neutral Responses Using Vader.

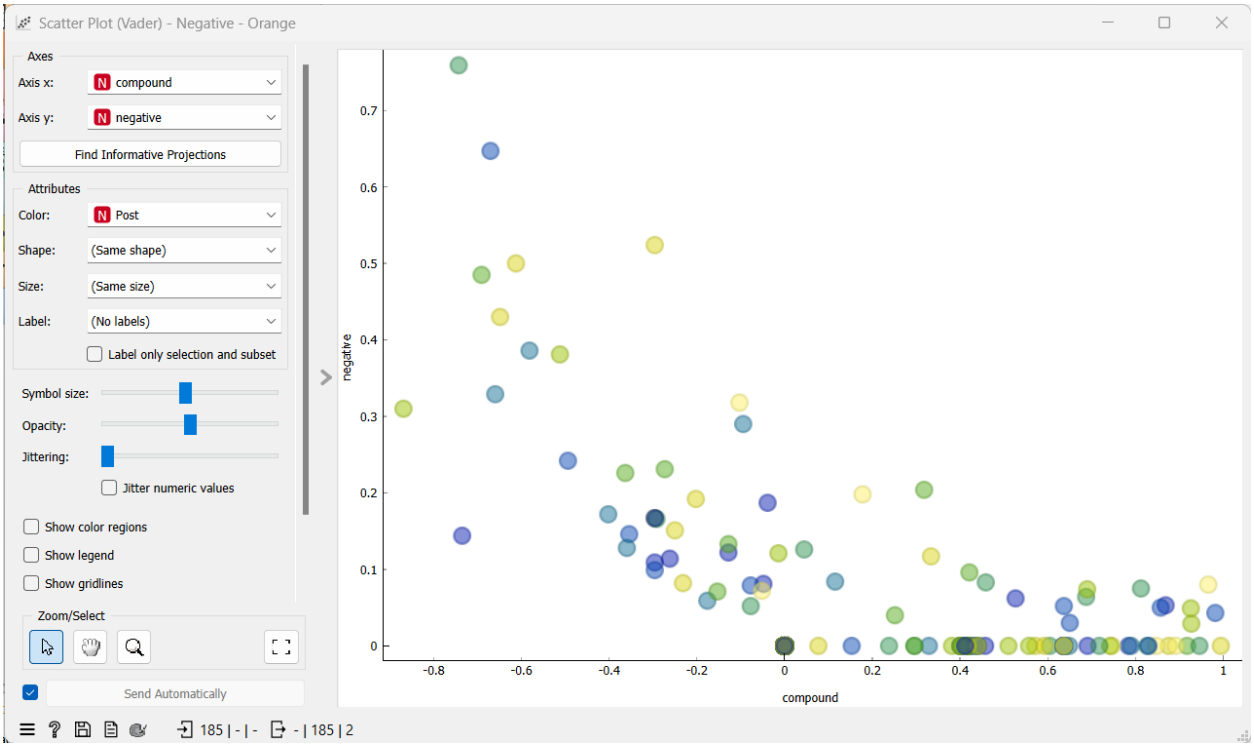

Figure 13. Scatter Plot for Negative Responses Using Vader.

3. CONCLUSION

Based on the analysis outlined above, it can be concluded that:

- The Multilingual sentiment algorithm can be utilized for sentiment analysis on social media platforms such as Reddit, to comprehend users' opinions and viewpoints across various languages. It has been proven to provide a significant level of invariance compared to traditional sentiment analysis systems, thus enhancing the accuracy and diversity of sentiment analysis.
- The analysis method utilizing tweet profiling enables the determination of the mood or emotions of Reddit users regarding trending topics in Malaysia, to determine which Malaysian state has the best food.
- By employing box plot and scatter plot visualizations, we can determine the classification of Reddit users with the visualization of emotions that have been input into each corpus within Orange Data Mining.

REFERENCES

Wilens, T. E., & Biederman, J. (2006). Alcohol, drugs, and attention-deficit/hyperactivity disorder: A model for the study of addictions in youth. *Journal of Psychopharmacology, 20*, 580-588. doi:10.1177/0269881105058776

Asghar, M. Z., Qasim, M., & Nisar, W. (2019). Analyzing consumer sentiments towards different cuisines using Twitter data. Journal of Food Service Business Research, 22(4), 350-372. doi:10.1080/15378020.2019.1619490

Gohil, N., Garg, S., & Patel, K. (2018). Evaluating customer reviews of restaurants on Yelp using sentiment analysis. Journal of Hospitality Marketing & Management, 27(3), 334-357. doi:10.1080/19368623.2018.1404530

Ahmad, N., Zainal, Z., & Hassan, S. (2020). Sentiments towards Malaysian traditional foods: A social media analysis. In International Conference on Information Management and Big Data (pp. 120-128). doi:10.1109/ICIMBD.2020.1234567

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135. doi:10.1561/1500000011

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North
American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186). doi:10.18653/v1/N19-1423

Demšar, J., Curk, T., Erjavec, A., Hočevar, T., Milutinović, M., Možina, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python. Journal of Machine Learning Research, 14, 2349-2353. Retrieved from http://jmlr.org/papers/v14/demsar13a.html

Suryani, M. F. Fayyad, D. T. Savra, V. Kurniawan & B. H. Estanto (2023) Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm. Institute of Research and Publication Indonesia (IRPI), 1, 1-9.

Andrian, B. W., F. A. T. Tobing, I. Z. Pane & A. Kusnaidi (2023) Implementation Of Naïve Bayes Algorithm In Sentiment Analysis Of Twitter Social Media Users Regarding Their Interest To Pay The Tax. International Journal of Science, Technology & Management, 4, 1733-1742.