Sentiment Analysis on Restaurant Reviews: Using Machine Learning Approach

Ho Yin Tam

230018755

MSc Data Science

ho.tam@city.ac.uk

Problem statement and motivation:

With the rapid growth of technology and the internet, different social media evolved, such as Facebook, Instagram, and TikTok. These allow customers to view the comments left by other customers about the food quality and service of the restaurants when deciding the restaurant to dine in. Customers themselves can also leave comments about the taste of cuisine and service level and share their dining experience after the meals. In addition, the manager of the restaurants can improve the food quality and service according to the feedback of the customers. This eventually led to a growth of business revenue by offering higher quality food and service to the customers.

The win-win scenario for the business and customers shows the importance of the online reviews of the restaurant. However, reading thousands or millions of comments about restaurants is time-consuming and labour-intensive. To solve this matter, applying machine learning techniques is a way.

This paper will explore several supervised machine learning algorithms, such as logistic regression, Naïve Bayes, and support vector machines, to classify if a review is positive or negative. The confusion matrix and different metrics, including accuracy, precision, recall, and F1 score, will be used to evaluate the performance of these models.

Research hypothesis:

The first research hypothesis of this report is that the support vector machines model will have the highest performance metrics including accuracy, precision, recall, and F1 score compared to the logistic regression and Naïve Bayes. The attributed reason is that the support vector machines can perform well and capture complex relationships between features and labels in a high dimensional space. The second hypothesis is that the performance of the models will be improved when the feature extract technique switches from count vectorizer to term frequency-inverse document frequency vectorizer.

Related work and background:

Serval studies have compared the effectiveness of various machine learning algorithms for sentiment analysis on restaurant reviews. The first paper ‘Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques’ (Krishna et al., 2019) investigates the sentiment classification in restaurant reviews using a variety of machine learning algorithms including Naïve Bayes, support vector machines and decision trees. The findings showed that the support vector machines have the highest accuracy with 94.56%. The second paper ‘Machine Learning for Sentiment Analysis and Classification of Restaurant Reviews’ (Morgado-Dias et al., 2024) also applied different machine learning algorithms including K-nearest neighbours classifier, logistics regression, support vector machines, gaussian Naïve Bayes, multinomial Naïve Bayes to the restaurant reviews. The results showed that the support vector machines have the highest accuracy of 78%. The third paper 'Sentiment Analysis of Reviews' (Vaithyanathan, 2002) provides a broad perspective on sentiment analysis using distinct machine learning algorithms including Naïve Bayes, Bernoulli classifier logistic regression, and stochastic gradient descent. The finding shows that the support vector machines have the highest precision. For these three papers, the dataset used is totally different. However, they all indicate that support vector machines performed the best on the sentiment analysis of restaurant reviews. Therefore, they act as a great reference as the support vector machine algorithms are employed in this report.

Beyond constructing machine learning algorithms and evaluating the performance of the models, extracting the features from restaurant reviews is crucial for the success of sentiment analysis. One common technique for feature extraction is the Count Vectorizer method. This method transforms textual data into a numerical representation by counting the occurrences of each word within a review. This creates a bag of words model in which the word order is not considered but the frequency of each word becomes a feature. The paper "Extractive Summarization of Reviews Using Opinion Words and Sentence Importance" (Yang et al, 2016) explores the effectiveness of Count Vectorizers in sentiment analysis tasks, highlighting their simplicity and efficiency. Another popular feature extraction technique is the term frequency-inverse document frequency (TF-IDF). It assigns weights to each word based on its importance within the reviews. Words that appear frequently in a single review but rarely across all reviews might be more indicative of sentiment and receive a higher TF-IDF score. This helps the model focus on the most relevant words for sentiment classification. The research "Opinion Mining and Sentiment Analysis" (Pang and Lee, 2008) and "Sentiment Analysis and Opinion Mining" (Liu, 2012) discuss the application of TF-IDF for sentiment analysis, demonstrating its effectiveness in capturing the importance of words within a specific context.

Accomplishments:

- Task 1: Load the restaurant reviews dataset from Hugging Face and explore the nature of the dataset – Completed
- Task 2: Clean the dataset by checking any null values and removing duplicate reviews - Completed
- Task 3: Perform text-preprocessing by tokenization, lowercasing, punctuation removal, stop words removal, and lemmatization - Completed
- Task 4: Visualize the most frequent words by word cloud, the outliers of review length by boxplot, and the distribution of review length by histogram - Completed
- Task 5: Compare and critically evaluate the differences in review length before and after text preprocessing – Completed
- Task 6: Build and train the logistic regression, multinomial Naïve Bayes, and support vector machines models using count vectorizer as the baseline models - Completed
- Task 7: Test and evaluate the performance of these baseline models using various performance

metrics including accuracy, precision, recall, and F1 score - Completed

- Task 8: Visualize the performance of the baselines models by confusion matrix and receiver operating characteristic (ROC) curve with an area under curve (AUC) value - Completed
- Task 9: Build and train the logistic regression, multinomial Naïve Bayes, and support vector machines models using term frequency-inverse document frequency vectorizer as the improved models - Completed
- Task 10: Perform the grid search to tune the hyperparameters of the model and ten-fold cross-validation to prevent overfitting - Completed
- Task 11: Examine the best-trained model by the test set using different performance metrics including accuracy, precision, recall, and F1 score - Completed
- Task 12: Visualize the performance of the best-trained models by confusion matrix and receive operating characteristic (ROC) with an area under curve (AUC) value - Completed
- Task 13: Compare the model performance between the baselines and improved models – Completed
- Task 14: Perform in-depth error analysis to figure out the examples being misclassified - Completed

Approach and Methodology:

The target of this report is to compare and critically evaluate the performance of logistic regression, multinomial Naïve Bayes and support vector machines with two different feature extraction techniques which are count vectorizer and term frequency-inverse document frequency vectorizer. The first approach is to understand the basic statistics of the restaurant reviews dataset. To have a better intuition, data visualization is required such as the most frequent words by word cloud, the outliers of review length by boxplot, and the distribution of review length by histogram. Then text preprocessing is performed for tokenization, stop words removal, punctuation removal and lemmatization by the NLTK library. Once the text is pre-processed, the next task is to convert the text into matrix form using a count vectorizer or term frequency-inverse document frequency vectorizer. Then the matrix serves as the input of the logistic regression, multinomial Naïve Bayes and support vector machines. The scikit-learn library is employed for these feature extraction and model training.

Dataset:

The restaurant review dataset is publicly available and extracted from the website https://huggingface.co/datasets/pachequinho/restaurant_reviews. It contains 1000 rows and 2 columns. The columns are the reviews of the customers and whether they like the food or not. Customers who like the food are encoded as 1 while people who do not like the food are labelled as 0.

In the data exploratory process, no missing data is found. However, four reviews are duplicated and hence removed as they provide the same information. Removing the duplicate rows or observations can reduce the bias of the machine learning model. Figure 1 shows the bar chart of the number of positive and negative reviews. The restaurant reviews dataset is said to be a balanced dataset as the number of positive reviews and negative reviews are 499 and 497 respectively which account for 50.1% and 49.9% respectively.

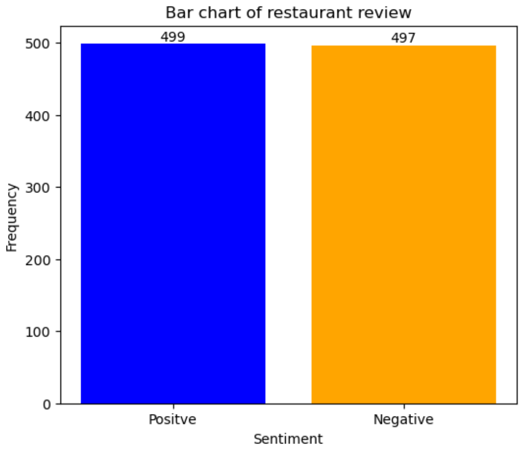

Figure 1: Bar chart of restaurant reviews

Figure 2 and Figure 3 display the word cloud representing the most common words that appeared in the positive reviews and negative reviews respectively. It is normal to observe the presence of words including ‘great’, ‘friendly’, and ‘amazing’ in the word cloud of positive reviews, while the words, such as ‘bad’, ‘disappointed’, and ‘worst’ emerge in the word cloud of negative reviews. The attributed reason is that these words carry either positive or negative connotations. Another notable discovery is that the words ‘food’, ‘place’, and ‘service’ appear very frequently in both positive and negative reviews. This could indicate the importance of food quality, environment, and service level in determining customer satisfaction or disappointment. Furthermore, this increases the difficulty of the machine learning model in accurately classifying a review as positive or negative.

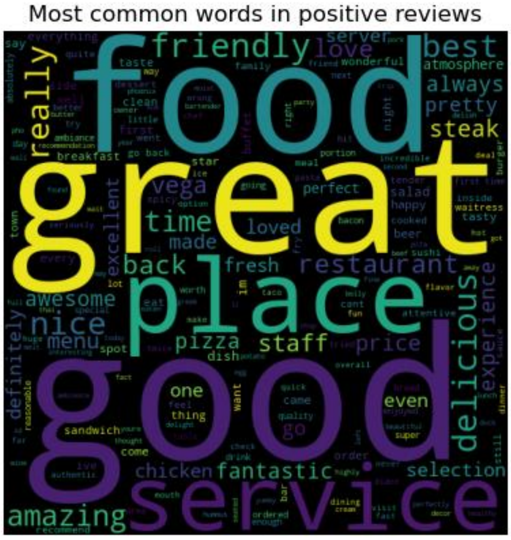

Figure 2: Positive word cloud

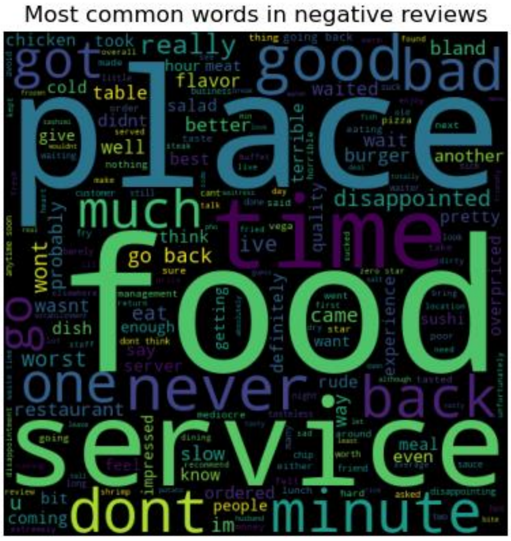

Figure 3: Negative word cloud

Figure 4 and Figure 5 present the boxplot of the review length of positive and negative reviews before and after text preprocessing respectively. It is clear that the mean of the review length of both positive and negative reviews before text preprocessing is approximately 10 words. The average review length of both positive and negative reviews after text preprocessing was trimmed down close by 100% to 5 words. This suggests that the reviews before text preprocessing contain half of the stop words. Another worth noting observation is that the outliers of review length before and after text preprocessing do not deviate much from the mean. Therefore, these outliers of the reviews are not removed.

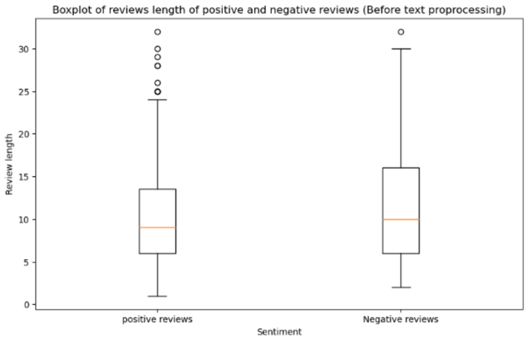

Figure 4: Boxplot of review length (Before text preprocessing)

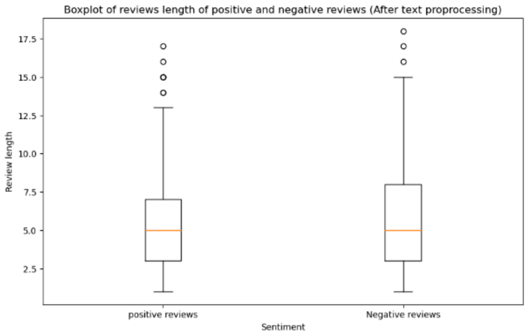

Figure 5: Boxplot of review length (After text preprocessing)

Figure 6 and Figure 7 exhibit the distribution of review length of the positive and negative reviews before and after text preprocessing respectively. Both positive and negative reviews before and after text preprocessing have a similar shape with the right-skewed distribution. The most frequent review length of both positive and negative reviews before text preprocessing is around ten words. After text preprocessing, the most frequent review length of both positive and negative reviews is approximately three to four words. This further reinforces that the text preprocessing successfully discarded a significant amount of text including stop words.

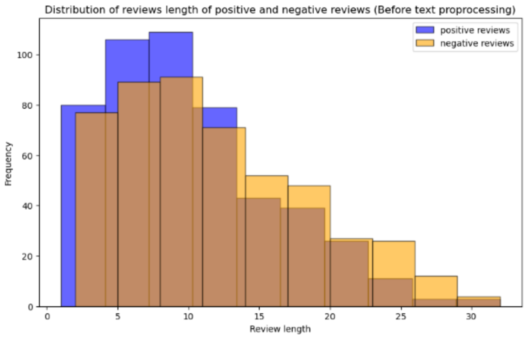

Figure 6: Distribution of reviews length (Before text preprocessing)

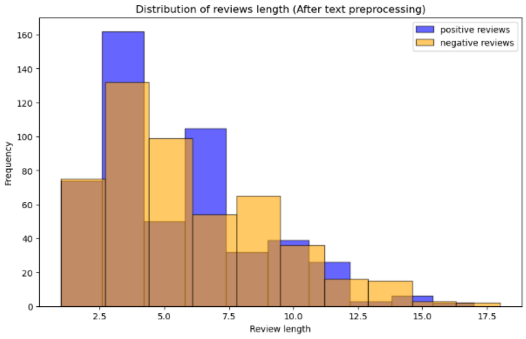

Figure 7: Distribution of reviews length (After text preprocessing)

Data preprocessing:

In the text preprocessing, a customized function named ‘text_preprocessing’ is created. The function consists of replacing the characters that are not uppercase, lowercase or whitespace with an empty string. This provides the ease of removing the special characters ‘…’. Also, the function encompasses tokenization which splits the reviews or sentences into words, and lowercasing which aids the consistency and normalization of the text. Furthermore, the function removes punctuation and English stop words as they do not carry meaningful information. Last but not least, the function comprises lemmatization which reduces the words into their base form since it considers the context and part of speech of the word.

Baselines:

As the dataset is balanced, a random guess could be a simple and easy-to-approach baseline, indicating a probability of 50% that a review is positive or negative. However, a probability of 50% or a baseline with an accuracy of 0.5 would not be a satisfactory result. Therefore, three baseline machine learning models are used with the count vectorizer.

The first baseline is logistic regression with the count vectorizer. Logistic regression is a supervised machine learning model that can be used in classification. The algorithm exploits the sigmoid function to predict the probability of a binary outcome. The advantage is that the algorithm is easy to implement and interpret. Yet, the disadvantage is that it cannot capture complex relationships between features and labels.

The second baseline is the multinomial Naïve Bayes with count vectorizer. Naïve Bayes is a supervised learning algorithm that can be used in classification. It exploits Bayes' Theorem in which the algorithm assumes all the features or words are conditionally independent of each other. The advantages are that it is simple, easy to understand and fast. It also performs well in text classification. However, the disadvantage is the unrealistic assumption that the occurrence of each word is independent of the other words.

The third baseline is the support vector machines with the count vectorizer. Support vector machines are supervised learning algorithms that can be used for regression and classification. The algorithm attempts to maximize the margin between the hyperplane and the nearest point of any class. The advantage is that it performs well in high-dimensional space. Nonetheless, the disadvantage is that the computation cost is expensive, and the training time is long.

Count vectorizer is used to convert the preprocessed text into matrix form which can then be the input of these models. It was chosen as the feature extraction technique for the baseline models since it is simple and easy to understand.

Results, error analysis:

This section focuses on the comparison of adopting the count vectorizer and term frequency-inverse document frequency (TF-IDF) vectorizer. Compared to the count vectorizer, another feature extraction technique named term frequency-inverse document frequency (TF-IDF) is used to convert the processed text into the matrix which then serves as the input of the algorithms. For logistic regression and support vector machines, grid search is employed to explore different combinations of hyperparameters while ten-fold cross-validation is applied to avoid overfitting. For multinomial Naïve Bayes, only ten-fold cross-validation is exploited.

Logistic regression:

In the training process of the logistic regression model, grid search is employed to explore different combinations of hyperparameters including the inverse of regularization and the regularization penalty term. The grid search result shows that the best hyperparameters are the value 1 of the inverse of regularization and L2 of the regularization penalty term. The best-trained model with these hyperparameters achieves an accuracy score of 0.79 which is slightly higher than the baseline mode with 0.76. Then the best trained model is evaluated using the test set.

Figure 8 displays the confusion matrix of logistic regression using the term frequency-inverse document frequency (TF-IDF) vectorizer. Table 1 shows the evaluation metrics including accuracy, precision, recall and F1 score of the logistic regression model using count vectorizer and frequency-inverse document frequency vectorizer. Though the performance metrics between them are very close, it is clear that the logistic regression using a frequency-inverse document frequency vectorizer performs slightly better.

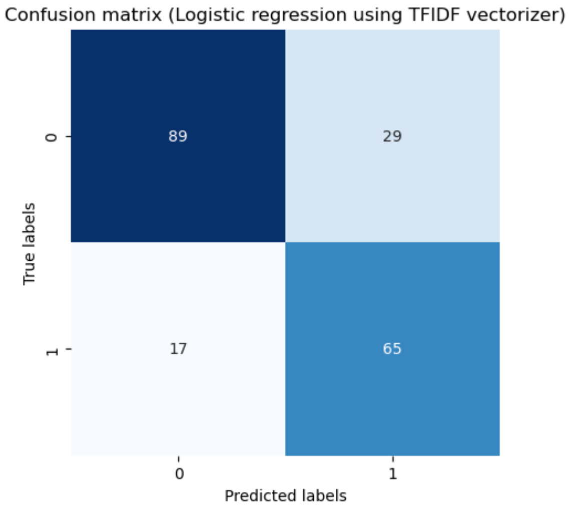

Figure 8: Confusion matrix (Logistic regression with TF-IDF vectorizer)

	Logistic regression with count vectorizer (baseline)	Logistic regression with TF-IDF vectorizer
Accuracy	0.76	0.77
Precision	0.68	0.69
Recall	0.78	0.79
F1 score	0.73	0.74

Table 1: Performance metrics of logistic regression with count vectorizer and TF-IDF vectorizer

Figure 9 shows the receiver operating characteristic (ROC) curve with the area under curve (AUC) value of the logistic regression using the TF-IDF vectorizer.

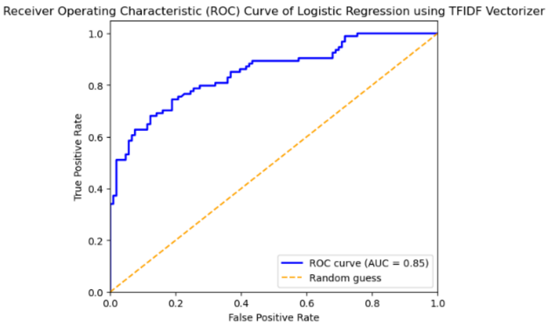

Figure 9: Receiver operating characteristics (ROC) curve of logistic regression with the term frequency-inverse document frequency (TF-IDF) vectorizer

Error analysis:

The following examples are the restaurant reviews that are misclassified as positive but are actually negative. In the investigation, the removal of stop words such as ‘no’ and ‘not’ may lead to the sentence only being left with some positive words including ‘right’, ‘nice’, and ‘impressed ‘This may lead to the bias of the model.

- ‘not even a "hello, we will be right with you.’
- ‘It was a pale color instead of nice and char and has NO flavor.’
- ‘I checked out this place a couple years ago and was not impressed.’

In contrast, the following example is the restaurant reviews that are misclassified as negative but are actually positive. After the removal of the stop words ‘no’, the review contains the negative word ‘disappointed’. This may lead to the bias of the model.

- ‘I went to Bachi Burger on a friend's recommendation and was not disappointed.’

Naïve Bayes:

As Naïve Bayes has few hyperparameters to tune, no grid search is required. Term frequency-inverse document frequency (TF-IDF) is used to convert the pre-processed text into the matrix form which can then be the input for the multinomial Naïve Bayes classifier. Ten-fold cross-validation is applied to prevent overfitting.

Figure 10 displays the confusion matrix of Naïve Bayes with the term frequency-inverse document frequency (TF-IDF) vectorizer. Table 2 compares the performance metrics including accuracy, precision, recall and F1 score of the Naïve Bayes with count vectorizer and frequency-inverse document frequency (TF-IDF) vectorizer. All performance metrics of Naïve Byes with term frequency-inverse document file (TF-IDF) vectorizer are higher than with count vectorizer. The difference in precision is the most significant with 5.8% in which the precision using the term frequency-inverse document file (TF-IDF) vectorizer is 5.8% higher than using the count vectorizer. This indicates that the Naïve Bayes with the term frequency-inverse document frequency (TF-IDF) vectorizer is more likely to identify the true positive restaurant reviews out of the total number of positive predictions.

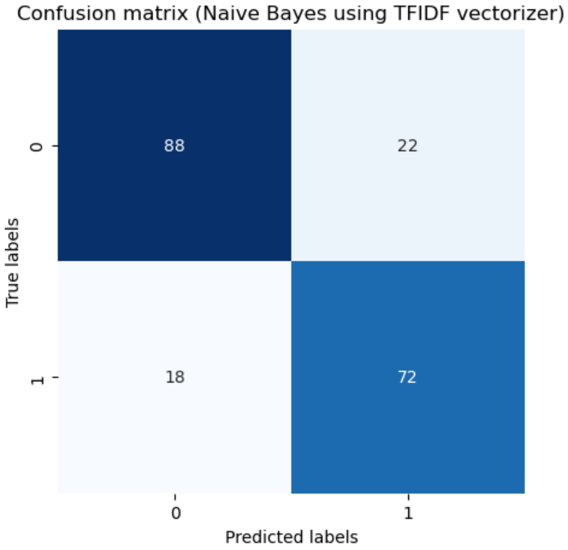

Figure 10: Confusion matrix (Naïve Bayes with TF-IDF vectorizer)

	Naïve Bayes with count vectorizer (baseline)	Naïve Bayes with TF-IDF vectorizer
Accuracy	0.79	0.80
Precision	0.72	0.77
Recall	0.80	0.80
F1 score	0.76	0.78

Table 2: Performance metrics of Naive Bayes with count vectorizer and TF-IDF vectorizer

Figure 11 plots the receiver operating characteristics (ROC) curve of Naïve Bayes with the term frequency-inverse document frequency (TF-IDF) vectorizer. Both Naïve Bayes with the count vectorizer and the term frequency-inverse document frequency (TF-IDF) vectorizer have the same area under curve (AUC) value of 0.85. This indicates that they perform well in classifying the restaurant reviews into positive classes or negative classes.

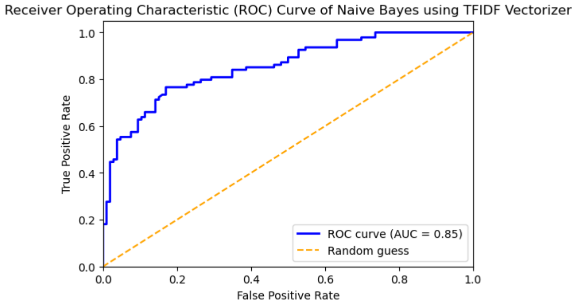

Figure 11: Receiver operating characteristics (ROC) curve of Naïve Bayes with the term frequency-inverse document frequency (TF-IDF) vectorizer

Error analysis:

The following examples are the restaurant reviews that are misclassified as positive but are actually negative. In the investigation, the words ‘sandwich and ‘buffet’ appeared more frequently in the positive reviews in the training set. This may lead to a higher posterior probability in which they are classified as positive reviews given these words appeared.

- ‘The problem I have is that they charge $11.99 for a sandwich that is no bigger than a Subway sub (which offers better and more amount of vegetables).’
- ‘The ambience here did not feel like a buffet setting, but more of a douchey indoor garden for tea and biscuits.’
- ‘The Buffet at Bellagio was far from what I anticipated.’

On the other hand, the following examples are the restaurant reviews that are misclassified as negative but are actually positive. Through detailed analysis, the words ‘chicken’, ‘salad’, and ‘burger’ appeared more frequently in the negative reviews in the training set. This may lead to a higher posterior probability in which they are classified as negative reviews given these words appeared.

- ‘High-quality chicken on the chicken Caesar salad.’
- ‘I got to enjoy the seafood salad, with a fabulous vinaigrette.’
- ‘I went to Bachi Burger on a friend's recommendation and was not disappointed.’

Support vector machines:

In the training process of the support vector machines, grid search is employed to explore different combinations of hyperparameters including the kernel function, degree of polynomial kernel function and regularization parameter. The grid search result shows that the best hyperparameters are the polynomial kernel function, degree of 2 of the polynomial kernel function and regularization parameter of 1. The best-trained model with these hyperparameters achieves an accuracy score of 0.79, slightly higher than the baseline mode with 0.76. Then the best trained model is evaluated using the test set.

Figure 12 displays the confusion matrix of support vector machines using the term frequency-inverse document frequency (TF-IDF) vectorizer. Table 3 shows the evaluation metrics including accuracy, precision, recall and F1 score of the support vector machines model using count vectorizer and frequency-inverse document frequency vectorizer. The accuracy and precision of the support vector machines with the term frequency-inverse document frequency is higher than the count vectorizer. This further reinforces that the support vector machines with the term frequency-inverse document frequency using grid search and ten-fold cross-validation enhance the model performance.

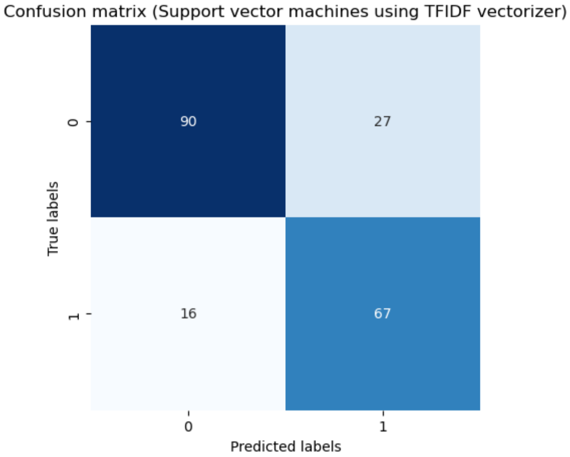

Figure 12: Confusion matrix (Support vector machines regression with TF-IDF vectorizer)

	Support vector machines with count vectorizer (baseline)	Support vector machines with TF-IDF vectorizer
Accuracy	0.76	0.79
Precision	0.63	0.71
Recall	0.81	0.81
F1 score	0.71	0.76

Table 3: Performance metrics of support vector machines with count vectorizer and TF-IDF vectorizer

Figure 13 depicts the receiver operating characteristics (ROC) curve of support vector machines with the term frequency-inverse document frequency (TF-IDF) vectorizer. The support vector machines with the term frequency-inverse document frequency (TF-IDF) vectorizer have a slightly higher area under curve (AUC) value of 0.85 than the count vectorizer of 0.83. The difference is small. Hence, both perform well in classifying restaurant reviews into positive classes or negative classes.

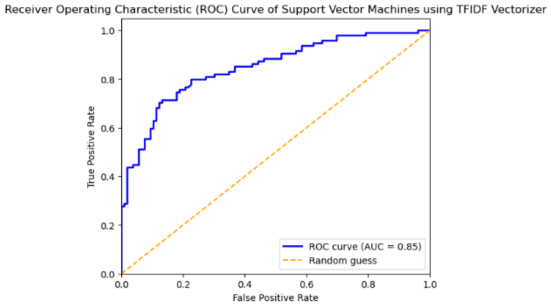

Figure 13: Receiver operating characteristics (ROC) curve of support vector machines with the term frequency-inverse document frequency (TF-IDF) vectorizer

Error analysis:

The following examples are the restaurant reviews that are misclassified as positive but are actually negative.

- ‘Furthermore, you can't even find hours of operation on the website!’

On the other hand, the following examples are the restaurant reviews that are misclassified as negative but are actually positive.

- ‘Not to mention the combination of pears, almonds and bacon is a big winner!.’

As the TF-IDF vectors represent text in a high-dimensional space, misclassification may occur due to overlapping classes or insufficient margins between classes. This will lead to difficulty in finding the optimal hyperplane.

Lessons learned and conclusions:

In conclusion, this report performed data visualization for the most frequent words by word cloud, the outliers of review length by boxplot, and the distribution of review length by histogram and compared the difference between the properties of data before and after text preprocessing. Also, text preprocessing is performed to remove irrelevant elements in the reviews. The experiment result shows that the multinomial Naïve Bayes model performs the best with the highest accuracy of 0.8 compared to the logistic regression and support vector machines. This does not support the first hypothesis that support vector machines perform the best which is the result of other papers mentioned in the previous section. In addition, it is clear that all models switching from count vectorizer to term frequency-inverse document frequency vectorizer enhanced the performance, especially the multinomial Naive Bayes and support vector machines having a significant increase in precision.

Reference:

Abbasi, A., Chen, H., & Salem, A. (2017). Sentiment analysis of online reviews for hotel services using ensemble learning. International Journal of Hospitality Management, 69, 77-88.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media, Inc.

Liu, B. (2012). Sentiment analysis and opinion mining. Morgan Kaufmann Publishers.

Kalchbrenner, J., Gρovdell, M., & Blomqvist, S. (2014). Deep convolutional neural networks for sentiment analysis of short texts. arXiv preprint arXiv:1406.2148.

Krishna, A., Dutt, V., & Singh, P. K. (2019). Sentiment analysis of restaurant reviews using machine learning techniques. 2019 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 130-134.

Morgado-Dias, J., Mateus, A., & Rocha, A. (2024, April). Machine learning for sentiment analysis and classification of restaurant reviews. In 2024 International Conference on Intelligent Systems and Networks (ISN) (pp. 123-128). IEEE.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Thelwall, M., Buckley, K., Vaughan, M., & Snowball, A. (2010). Comparative study of knowledge resources for sentiment analysis. Language Resources and Evaluation, 44(4), 169-186.

Vaithyanathan, S. (2002). Sentiment analysis of reviews. International Journal of Computational Linguistics and Speech Processing, 2(1), 127-137.

Yin, J., Shen, H., & Tang, Y. (2012). A hybrid sentiment analysis approach to customer reviews for online hotels. Knowledge-Based Systems, 25(2), 207-217.