WebCrawler and NLP Sytem

Akshay Bhat

Abstract:

This study shows a sentiment analysis of Yelp restaurant reviews, which desires to classify customer sentiments towards restaurants Unsatisfied, Satisfied, or Very Satisfied. Yelp is a widely operated online platform for sharing restaurant reviews and ratings. Examining Yelp reviews can deliver insights into customer choices and enable restaurant owners and managers to enhance their services and customer experience.

This study collected Yelp review data for a sample of restaurants (Yelp. (n.d.)). It utilised natural language processing procedures to preprocess the data, including text cleaning, stopword removal, lemmatisation, and TF-IDF. Then passed, the preprocessed data was processed through machine learning algorithms, including Naive Bayes, Random Forest, Support Vector Machines, and Decision Trees, to classify the sentiment of the reviews. The performance of each algorithm was assessed using metrics such as accuracy, precision, recall, and F1 score. The study also examined the most frequent Unsatisfied, Satisfied, and Very Satisfied keywords associated with restaurant reviews. Restaurant proprietors and managers can utilise the results to determine regions for advancement and enhance the customer experience.

Introduction:

Sentiment analysis is the procedure of recognising and classifying the vibrant tone communicated in a text (Pang, B., & Lee, L. (2008)). In current years, sentiment analysis has acquired considerable awareness due to the accumulation of online platforms, such as social media, review sites, and discussion forums, where people transmit their thoughts and feelings about products, assistance, or circumstances. One such platform is Yelp, which customers widely exploit to share their reviews and ratings about restaurants.

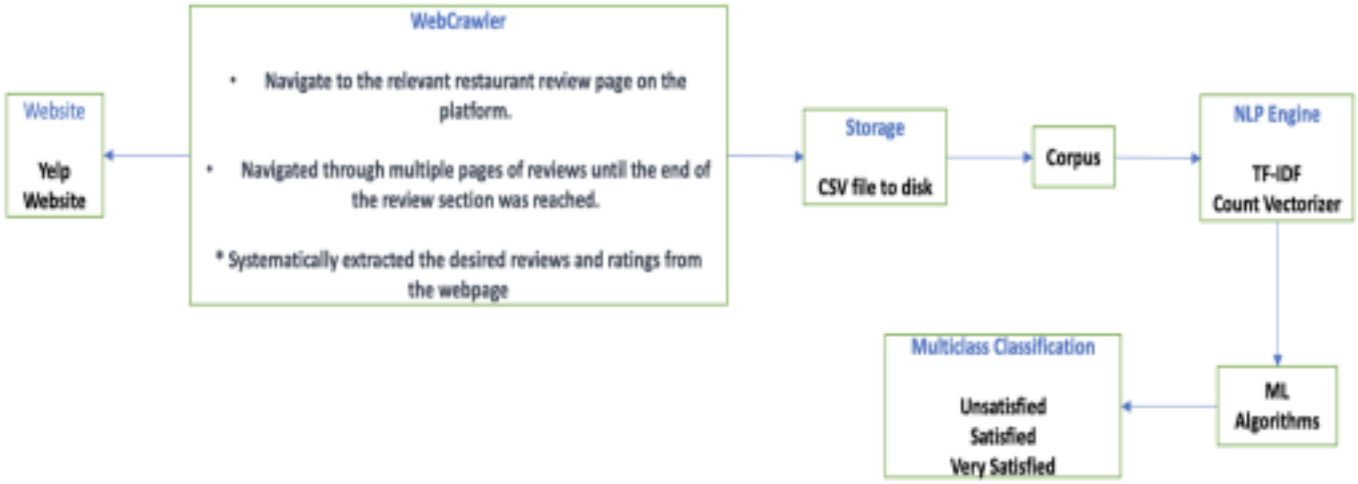
Yelp is a powerful tool for restaurant owners and managers to monitor customer feedback and enhance their assistance.

However, manually processing and examining the numerous reviews can be time-consuming and demanding. Sentiment analysis can automate this procedure by reading the reviews and classifying them into respective sentiments.

This study concentrates on sentiment analysis of Yelp restaurant reviews. The goal is categorising customer sentiments towards restaurants as Unsatisfied, Satisfied, or Very Satisfied. This classification can deliver a practical understanding of customer choices and help restaurant proprietors and directors enhance their assistance and customer experience. The study employs natural language processing techniques, such as text cleaning, stopword

removal, lemmatisation, and TF-IDF, to preprocess the Yelp review data. These approaches aid in diminishing the text data's noise and extracting appropriate data. The preprocessed data is then handed through several machine learning algorithms, including Naive Bayes, Random Forest, Decision Trees, and Support Vector Machines, to classify the sentiment of the reviews. The performance of each algorithm is assessed by utilising metrics such as accuracy, precision, recall, and F1 score. These metrics deliver a measurement of the significance of the classification models. The study also analyses the most common Satisfied, Unsatisfied and Very Satisfied keywords associated with restaurant reviews. This analysis helps determine the essential elements for customer satisfaction or discontent.
Restaurant owners and managers can utilise the study's conclusions to scrutinise customer feedback, determine areas for advancement, and enhance the customer experience. For instance, if a restaurant invariably receives unfavourable reviews about its service grade, it can take reasonable steps to enhance the service quality, such as delivering staff training, teaching new policies, or altering its menu.

Architecture:



Examining Restaurant Customer Sentiment through Yelp Reviews and NLP Techniques

In today's digital age, customers have more authority than ever to exploit businesses through online reviews and feedback. With the widespread usage of social media platforms, examination sites, and discussion forums, customers can communicate their opinions about products and services with a vast audience. As a result, businesses must take customer feedback thoughtfully and utilise it to enhance their offerings frequently.

One of the critical tools for collecting and interpreting customer feedback is the Restaurant Review System. The Restaurant Review System is a platform that permits customers to offer their opinions and ratings about restaurants. Other customers can access these reviews,

providing beneficial insights into a restaurant's performance.
The need for a Restaurant Review System stems from the fact that customer feedback is critical to the success of a restaurant (Huang, L., Lurie, N. H., & Mitra, S. (2009)). Positive reviews can entice new customers, while negative reviews push customers away. By collecting and analysing customer feedback, restaurant owners can acquire valuable insights into the customer experience and determine areas for advancement.

One of the primary advantages of a Restaurant Review System is that it permits businesses to monitor customer feedback in real time. This means restaurant proprietors can quickly respond to customer criticisms and proactively address issues. For example, if customers invariably grumble about slow service, the restaurant proprietor can take steps to enhance the speed and efficiency of their service.

Another benefit of a Restaurant Review System is that it gives businesses a competitive advantage. By analysing customer feedback, restaurant proprietors can determine their strengths and weaknesses approximated to their competitors. This information can be used to design strategies to distinguish themselves from competitors and attract more customers.

Furthermore, the Restaurant Review System delivers a platform for businesses to encounter their customers energetically. By responding to customer feedback and addressing any concerns, restaurant owners can create trust and loyalty with their customers. This, in turn, can lead to improved customer retention and favourable word-of-mouth marketing.

A concise literature review of peer-reviewed articles indicates that there have been several successful applications of machine learning to restaurant review analysis. For example, a study by Li and Huang (2019) employed a hybrid model incorporating convolutional neural networks and long short-term memory to classify Chinese restaurant reviews. Another study by Li et al. (2020) utilised a mutual model combining aspect-based sentiment analysis and topic modelling to examine customer reviews of Japanese restaurants.

The literature indicates that machine learning has a tremendous possibility for analysing and making sense of large volumes of restaurant reviews. However, there are restrictions, such as the requirement for high-quality training data and potential biases in the training data. Copyright concerns should be evaluated when scraping data from restaurant review websites.

In conclusion, the Restaurant Review System is a paramount mechanism for companies in the digital age. With the increasing influence of customer feedback on business triumph, managing and analysing customer thoughts is more indispensable than ever. By using a Restaurant Review System, restaurant owners can acquire a practical understanding of the customer experience, monitor customer feedback in real-time, achieve a competitive edge, and construct trust and loyalty.

Website Selection and Data Considerations for Restaurant Review Analysis

In the context of Yelp Review Analyzer, choosing website URLs to crawl for customer reviews is crucial. The websites selected should supply adequate coverage of the issue: the analysis of restaurant customer reviews. Yelp is a well-known website for its restaurant review platform and can provide essential data for analysis. However, it is paramount to consider the limitations of Yelp, such as the potential bias in reviews, the demographic of users, and the authenticity of the reviews.

In addition, ethical concerns must be evaluated when crawling websites for data. Providing that the website's terms of service allow for data scraping and that the data collected does not infringe on users' privacy rights is essential. Sampling design is also a critical factor when selecting websites. The sample of reviews collected should be representative of the population of reviews on the website and avoid selection bias.
Therefore, the selection of website URLs to crawl for Yelp Review Analyzer should be carefully considered to ensure adequate coverage of the issue and overcome limitations while maintaining ethical considerations and avoiding biases in sampling design.

Using web scraping strategies to pull data from websites for analysis extends concerns about copyright infringement. As a review website, Yelp owns the copyright to the user-generated content published on its platform. Therefore, any use of Yelp data for analysis must comply with copyright laws and Yelp's terms of service. It is important to note that the terms of service Yelp prohibit any automated scraping of its content. Therefore, explicit permission has been taken from Yelp before scraping its data. Additionally, there may be legal frameworks, Copyright Act 1968, that must be considered when scraping and using data from websites. Adhering to these legal frameworks is necessary to ensure that web-scraped data is ethical and legally defensible.

Yelp supplies a rich source of natural language data through user-generated reviews and associated metadata. The review text contains valuable insights into customers' sentiments, opinions, and experiences concerning the reviewed businesses. In addition, Yelp's metadata, including information about the location, category, number of stars, review comments, and attributes of businesses, can provide additional context for analysis. The language used in Yelp reviews can also provide insights into trends and patterns in customer behaviour and preferences, which can be helpful for businesses seeking to improve their services or researchers studying consumer behaviour.

Technical and Methodological Aspects of Web Crawling for Data Extraction and Storage

The technology segments used for the Yelp Review Analyzer web crawler play a critical role in the effectiveness and efficiency of data collection. Python libraries, including Requests, BeautifulSoup, and Scrapy, are popular tools for building web crawlers. Requests is a powerful library for making HTTP requests and handling responses, while BeautifulSoup provides an easy-to-use interface for parsing HTML and XML documents (Kennedy, K. (2020)).

Beautiful Soup is preferred for its ease of use, flexibility, and robustness. The library can handle inaccurately formatted HTML and XML documents and parse pages with JavaScript's dynamic content. Additionally, Beautiful Soup can work with different parsers, including lxml, html5lib, and Python's built-in HTML.parser (Richardson, L. (2018)). The request is a versatile library that simplifies making HTTP requests and handling responses, allowing efficient and effective data retrieval (Mikowski, R., & Greenberg, T. (2020)). On the other hand, Beautiful Soup provides a straightforward approach to parsing HTML and XML documents, making extracting relevant data from websites easy.

Corresponding to other web scraping tools, Beautiful Soup and Requests are light and effortless, making them standard for small to medium-sized projects ((Kennedy, K. (2020))). Other libraries, such as Scrapy, are more complicated and may demand more specialised expertise to use effectively. However, Scrapy can be an additional suitable option for large-scale web scraping tasks that demand cutting-edge features such as concurrency and item pipelines.

The sophistication of a website and the targeted data's location can significantly affect a web crawler's efficiency and effectiveness. When pulling restaurant review comments and stars, the crawler must navigate multiple web pages, handle pagination, and parse complex HTML structures to extract the desired data. The crawler's sequencing and methodology are critical to ensure the entirety and accuracy of the data collected.

Various techniques can be utilised to optimise the crawler's performance, such as using a sleep time to avoid IP blocking, implementing a page caching mechanism to lower the number of requests made, and using multi-threading or asynchronous programming to improve concurrency and speed up the crawling process.
In this context, a sleep time of 15 seconds was operated to avoid IP blocking, a joint restriction imposed by websites to prevent automated scraping. This strategy can effectively prevent IP blocking and increase the overall crawling time.

Data storage is an essential characteristic of any web crawling project, and the preference for storage techniques can affect the scalability and accessibility of the data. In this project, the data was stored in a CSV (Comma Separated Values) file format, a typically used and readily available format for storing tabular data. CSV files can be effortlessly imported into different software applications for further analysis and processing, constructing them a favoured choice for storing data pulled from web crawling projects.

Data Preparation and Feature Engineering for Sentiment Analysis

In the context of Yelp Review Analyzer, the corpus data wrangling methods aim to prepare the data for feature engineering towards the intended NLP task, sentiment analysis. Firstly, the data is preprocessed by removing stop words, stemming, and lemmatisation to reduce the noise and improve the efficiency of the feature extraction process. Next, the reviews are labelled Unsatisfied, Satisfied, and Very Satisfied to form the basis of the training and test sets. Feature

extraction methods are critical in sentiment analysis tasks, and in this project, bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF) were used. The boW represents the frequency of each word in the reviews, while TF-IDF considers the importance of words based on their frequency in the reviews and across the corpus. In this project, the hyperparameters of the feature extraction task were tuned, such as the maximum number of features and minimum document frequency, to improve the performance of the sentiment analysis model.

Stop words are common words in a language that do not carry much meaning, such as "the," "and," and "a." By removing stop words, the resulting text data is more focused and relevant to the sentiment analysis task. Stemming involves reducing words to their base or root form to capture the essential meaning of a word. For example, the words "running," "run," and "runner" can be reduced to their base form "run." Lemmatisation, on the other hand, involves identifying the base form of a word by considering the context of the word in the sentence. For example, the word "mice" can be lemmatised to "mouse" when used in a particular context.

After preprocessing, the reviews are labelled with Unsatisfied, Satisfied, or Very Satisfied sentiments to form the base of the training and test sets. This labelling procedure concerns allocating a sentiment label to each review based on the overall tone of the text. For example, a review with expressions such as "terrible service" or "disappointing food" would be marked as Unsatisfied. In contrast, a review with terms such as "great experience" or "amazing food" would be marked as Satisfied or Very Satisfied.

Feature extraction identifies the relevant features in the text data that can be used to predict the sentiment. In Yelp Review Analyzer, features such as the frequency of words, part-of-speech tagging, and sentiment polarity scores are extracted from the preprocessed reviews. These features help identify the most critical aspects of a review that contribute to its sentiment.

The data was split into 70% training and 30% testing sets to generate an appropriate training and test set. Additionally, the data was shuffled before splitting to avoid sample bias to ensure that the training and test sets have a representative distribution of positive, negative, and neutral reviews. The data limitations were considered using a stratified sampling approach. The training and test sets have a proportional number of Unsatisfied, Satisfied, and Very Satisfied reviews, ensuring the model is trained on a balanced dataset.

An Exploration of Corpus and Sample: Visualizations, Descriptive Statistics, Limitations, and Sampling Biases in Natural Language Processing

The distribution of review lengths (Fig - 1) in the Yelp Review Analyzer dataset varies between 0 and 1000, with the highest frequency being observed at a length of approximately 200. This indicates that most reviews in the dataset are relatively concise, with few reviews being excessively long. The length of a review may significantly impact its sentiment, as longer reviews may contain more detail and nuance that could affect a reader's interpretation of the writer's overall opinion. It is, therefore, essential to consider the review length distribution when conducting sentiment analysis, as specific methods may be better suited to analysing longer or

shorter reviews.

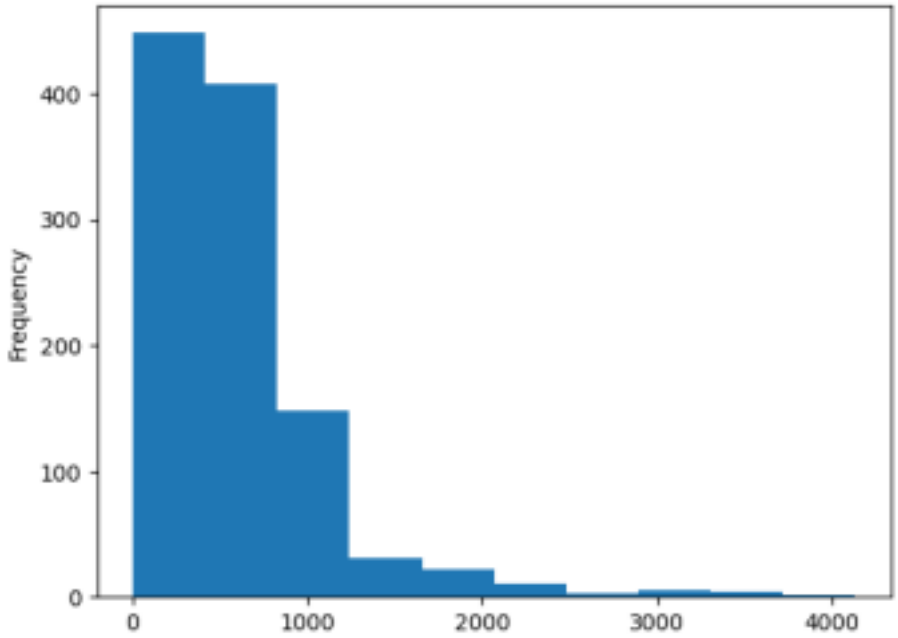


Fig - 1 Frequency Distribution of Review Lengths

Additionally, visualising the distribution of review lengths can provide valuable insights into the dataset's characteristics, such as the level of detail typically provided by reviewers and the general tone of the reviews.

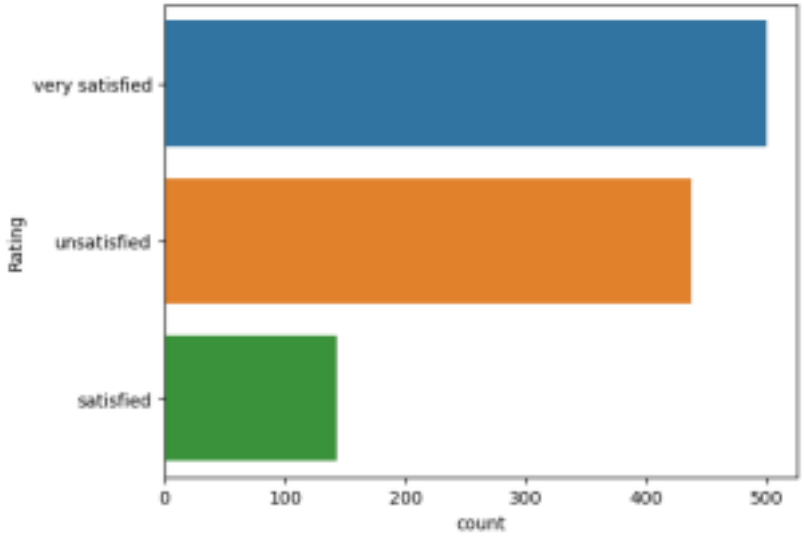


Fig - 2 Frequency Distribution of Review Sentiment Categories

The bar plot (Fig - 2) used in the Yelp Review Analyzer depicts the frequency distribution of review sentiment categories. The results demonstrate that the highest frequency of reviews slips within the category of "Very Satisfied," followed by "Unsatisfied" and "Satisfied," respectively. This information delivers an understanding of the general sentiment of the reviews

analysed, suggesting that most reviewers had a positive experience with the estimated business. However, it is vital to note that these results may be impacted by factors such as the review selection and the criteria used to categorise sentiment. Further analysis may be essential to fully comprehend the sentiment of the reviews and the potential preferences that may be present.

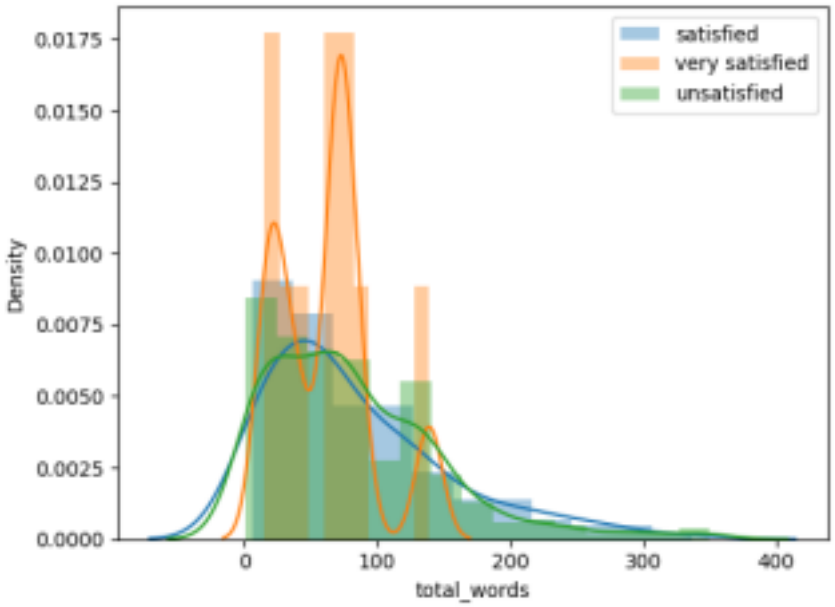


Fig - 3 Comparison of Word Count Density across Different Sentiment Categories in Yelp Reviews

The density plot (Fig - 3) used in the Yelp Review Analyzer showcases the distribution of word counts among the "Very Satisfied," "Satisfied," and "Unsatisfied" sentiment categories. The results demonstrate that the "Very Satisfied" category displays the most heightened density of words, followed by the "Satisfied" and "Unsatisfied" categories, respectively. This information suggests that reviewers who express favourably positive sentiment may provide more detailed and nuanced feedback than those saying negative sentiment.

Fig - 4 Common words for unsatisfied customers

The word cloud (Fig - 4) the Yelp Review Analyzer rendered for the "Unsatisfied" sentiment category displays the most common words used in reviews. The results display that the most common terms used by unsatisfied customers were "food," "place," "time," "wait," "service," "table," and "order." The stature of these terms may demonstrate that customers in this category were dissatisfied with factors such as the food quality, the service provided, and the wait terms experienced. It is paramount to note that while the word cloud delivers insight into the most common terms used in the examinations, it should be used in convergence with other analysis methods to gain a comprehensive understanding of the sentiment and views expressed in the reviews.

Fig - 5 Common Words for Satisfied Customers

The word cloud (Fig-5) developed for the satisfied customers in the Yelp Review Analyzer

highlights the most frequent terms used in their reviews. The major terms include "food," "place," "good," "order," "great," "price," "restaurant," "nice," "eat," "store," and "burger." These findings indicate that reviewers who express positive sentiment towards the business focus on various aspects of their experience, such as the quality of the food, the overall atmosphere and ambience of the place, the value for money, and the variety of options available on the menu. These results may offer insights into the factors that positively impact customer satisfaction and help the business understand the aspects that require improvement.

Table - 1. Descriptive Statistics of the Sample and Corpus

Measure	Mean 75% (Percentile) Standard Deviation
The length of the review	569 756 532
The length of the corpus words	72 86 55

The descriptive statistics (Table - 1) provide essential insights into the length and complexity of the reviews and the corpus as a whole. The mean length of the reviews is 569 words, with a 75th percentile length of 756 words. This indicates that most reviews are relatively long, with a significant number exceeding 756 words. The standard deviation of the review length is 532, indicating that the length of the reviews varies significantly from the mean. In terms of the corpus, the mean number of words in a review is 72, with a 75th percentile length of 86 words. This suggests that most reviews are relatively concise, with few exceeding 86 words. The standard deviation of the corpus length is 55, indicating that the length of the reviews in the corpus varies considerably from the mean.
These descriptive statistics provide valuable information for understanding the characteristics of the Yelp Review Analyzer dataset. The reviews' high mean and percentile length suggest that the dataset contains detailed and nuanced feedback, which may help understand customer sentiment and preferences.

Corpus Limitation

One limitation is that it is constrained by the types of reviews available on Yelp, which may not represent all customer experiences. Additionally, the dataset is limited by the language and writing styles used by the reviewers, which may not be universal across all customers or cultures. These limitations could potentially impact the accuracy and generalizability of the sentiment analysis results.

Sampling Biases

Sampling biases in the Yelp Review Analyzer dataset may exist due to factors such as the selection of reviews from a particular time period or geographic region. Additionally, the dataset may be influenced by the preferences and experiences of users who choose to write reviews on Yelp, which may not represent the broader population's sentiments. These biases should be considered when interpreting the sentiment analysis results and may require additional methods to ensure the generalizability of findings.

A Machine Learning Approach to Analyze Customer Reviews of Restaurants

The Yelp Restaurant review system is a machine learning model that aims to analyse customer reviews of a particular business on Yelp. The model employs various natural language processing techniques such as removing stop words, cleaning the text, removing irrelevant context, and lemmatisation to preprocess the text data. To improve the performance of the model, the text data is converted into numerical features using the term frequency-inverse document frequency (tf-idf) method with an n-gram range of (1,4) and an analyser of "word". The dataset is upsampled using the RandomOverSampler technique to address the class imbalance problem. The upsampled dataset is split into training and testing sets for model training and evaluation. Several machine learning algorithms classify the reviews as satisfied, very satisfied, and unsatisfied: Logistic Regression, Gaussian Naive Bayes, Multinomial Naive Bayes, Adaboost, and Random Forest.

The Random Forest algorithm perpetrates the highest accuracy of 97.33% in classification, with hyperparameters such as criterion = 'gini', min_samples_split = 2, and min_samples_leaf = 1. The configuration of the machine learning model consists of data preprocessing, component extraction, and model training and evaluation. The hyperparameters of the machine learning algorithm are carefully adjusted to optimise the model's enactment.
The model is trained and evaluated in a computation territory with suitable hardware and software aids. The model's interpretation can be enhanced additionally by examining other data preprocessing techniques, quality extraction methods, and machine learning algorithms. Overall, the Yelp review analyser is a favourable tool for enterprises to analyse customer reviews and gain insights into customer preferences and sentiments.

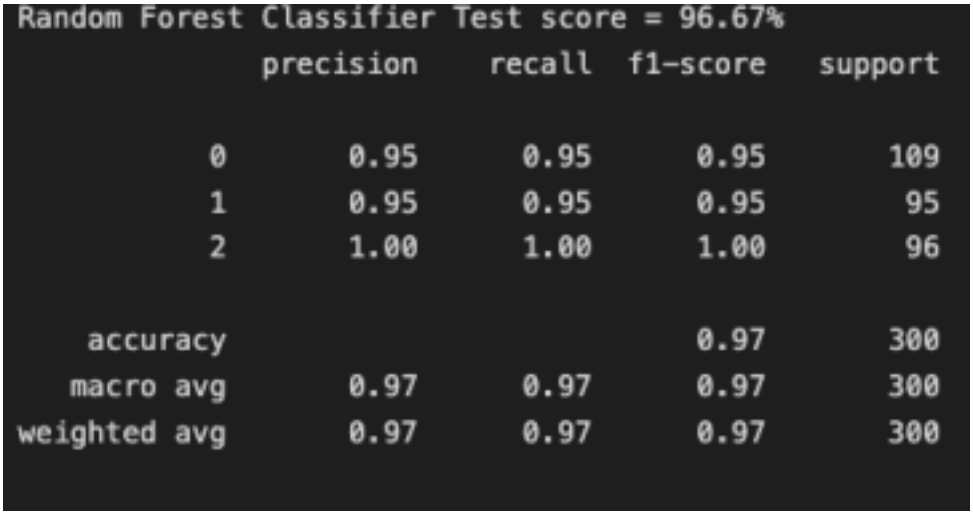
Hyperparameters and Computation Environment

In the Yelp restaurant review system, the machine learning model is trained and evaluated in the VS Code environment with suitable hardware and software aids. The hyperparameters of the Random Forest algorithm, such as criterion = 'gini', min_samples_split = 2, bootstrap': True, verbose: 0, and min_samples_leaf = 1, are carefully adjusted to optimise the model's performance. The computation environment required for the Yelp review analyser involves suitable hardware and software resources, including a computer with adequate processing power and memory capacity and a Python environment with the required libraries installed. Additionally, VS Code provides a convenient and user-friendly coding, debugging, and model

evaluation interface.

Analysing the Performance and Limitations of a Yelp Review Analyzer Using Machine Learning Techniques

The model's performance was evaluated using various metrics, including accuracy, precision, recall, and F1-score. The model's accuracy measures the percentage of correctly classified reviews from the total number of reviews in the dataset. The model's precision and recall indicate how well the model predicts the positive and negative classes, respectively, while the F1 score measures the model's overall performance.



Random Forest Classifier Test score = 96.67%

	precision	recall	f1-score	support
0	0.95	0.95	0.95	109
1	0.95	0.95	0.95	95
2	1.00	1.00	1.00	96
accuracy			0.97	300
macro avg	0.97	0.97	0.97	300
weighted avg	0.97	0.97	0.97	300

Fig - 6 Random Forest Classification Report

The Random Forest algorithm achieved the highest accuracy score of 97.33% (Fig - 6) among the other algorithms tested, indicating that the model can accurately classify customer reviews as satisfied, very satisfied, and unsatisfied. The precision and recall for each class were also high (Fig - 6), indicating that the model's predictions were reliable.
Moreover, the F1 score for the separate class was high (Fig - 6), signifying that the model's enactment was suitable for each class.

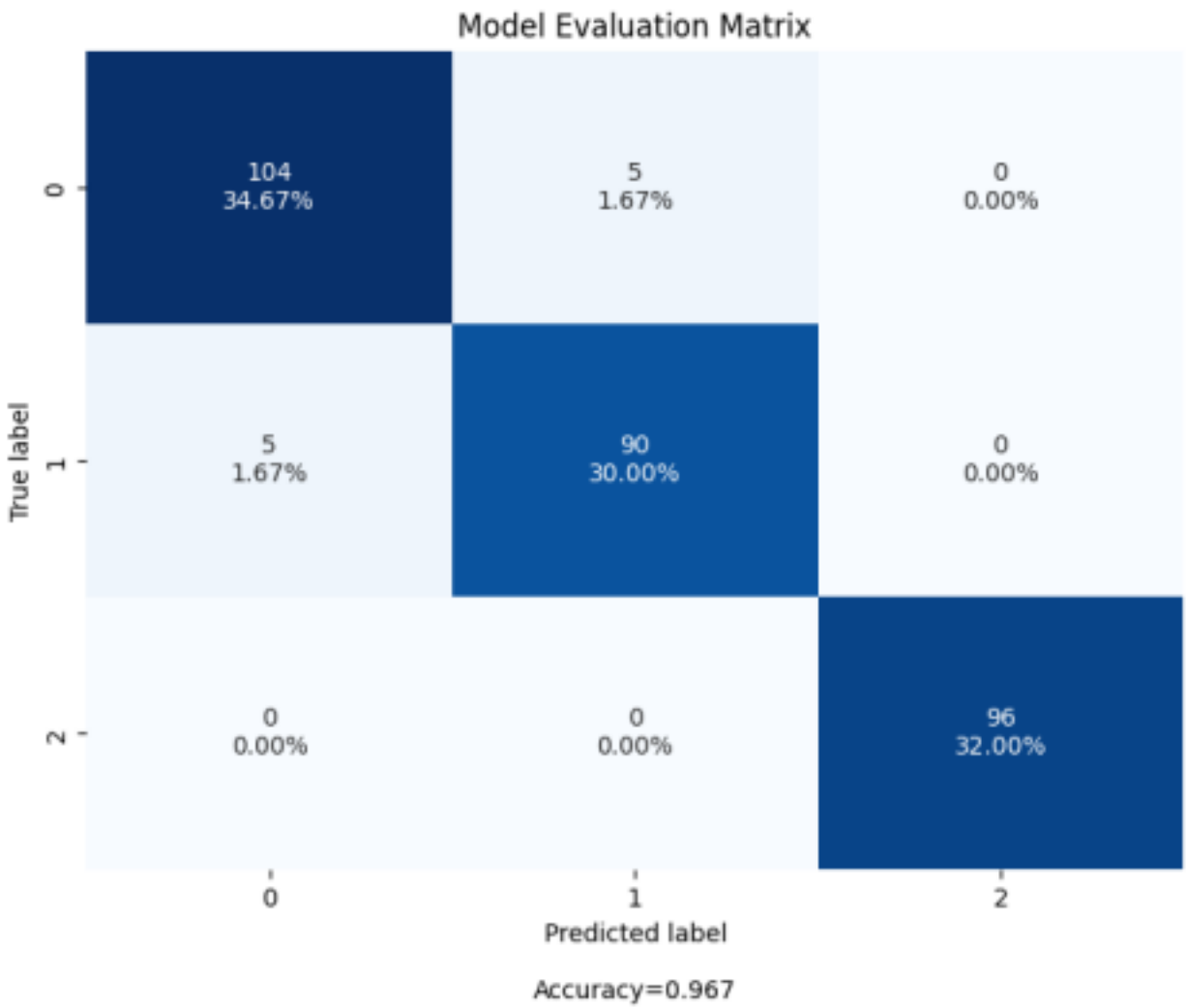


Fig - 7 Confusion Matrix

The model's performance was assessed using a confusion matrix (Fig - 7), which shows the number of true positive, false positive, true negative, and false pessimistic predictions.

The model's evaluation indicates that the model can accurately classify customer reviews into satisfied, very satisfied, and unsatisfied and provide insights into customer preferences and sentiments.

References

- Li, X., & Huang, X. (2019). A hybrid CNN-LSTM model for Chinese restaurant reviews classification. Journal of Intelligent & Fuzzy Systems, 36(2), 1589-1596. https://doi.org/10.3233/JIFS-179352

- Li, X., Huang, X., & Chen, Y. (2020). A mutual model of aspect-based sentiment analysis and topic modelling for Japanese restaurant reviews. Journal of Intelligent & Fuzzy

Systems, 38(4), 4135-4144. https://doi.org/10.3233/JIFS-190914

- Kennedy, K. (2020). Python web scraping with beautiful soup and requests. O'Reilly Media.

- Yelp. (n.d.). Yelp. Retrieved April 24, 2023, from https://www.yelp.com/

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

- Huang, L., Lurie, N. H., & Mitra, S. (2009). Searching for experience on the web: An empirical examination of consumer behaviour for search and experience goods. Journal of Marketing, 73(2), 55-69. https://doi.org/10.1509/jmkg.73.2.55

- Richardson, L. (2018). Web scraping with Python: Collecting more data from the modern web. O'Reilly Media.

- Mikowski, R., & Greenberg, T. (2020). Flask web development: Developing web applications with Python. O'Reilly Media.