# Yelp Review Classification Using NLP

**Milestone 3:** Project – Preprocessing and Transformation

**Student:** Urvashi Dube

+12368633298

dube.u@northeastern.edu

**Percentage of Effort Contributed by Student:** 100%

**Signature of Student :** Urvashi Dube

**Submission Date:** 22nd October, 2023

## Introduction

In today's fast-paced and highly competitive business landscape, understanding and responding to customer feedback is paramount. Customer reviews, particularly those shared on platforms like Yelp serve as valuable repositories of insights that can drive product enhancements, improve customer satisfaction, and shape strategic decision-making. However, the sheer volume of customer feedback, combined with its unstructured nature, makes manual analysis a time-consuming and error-prone endeavor.

This project endeavors to address the fundamental challenge of efficiently and effectively analyzing customer reviews by harnessing the power of Natural Language Processing (NLP) techniques. Yelp, as a leading platform for user-generated reviews, presents a treasure trove of opinions and sentiments, making it an ideal dataset for sentiment analysis. In this project, I will employ Natural Language Processing (NLP) techniques to classify Yelp reviews as either positive or negative based on user-assigned star ratings. Stars 1 and 2 are considered negative, while 3 and 4 indicate positive sentiment. By doing so, I aim to equip businesses with a powerful tool that not only expedites the review analysis process, showcasing the potential of NLP in deciphering human sentiment from text, but also provides actionable insights for enhancing their products, services, and overall customer experience.

## Problem Statement

The problem at hand is to develop an automated sentiment analysis solution for categorizing customer reviews on platforms like Yelp into positive and negative sentiments. This solution is essential to help businesses efficiently analyze customer feedback at a scale and make data-driven decisions based on customer sentiment.

## Data Used

### Dataset Overview

The "yelp_polarity" dataset is a comprehensive collection of Yelp reviews specifically curated for sentiment classification tasks by Xiang Zhang. It comprises a total of 560,000 Yelp reviews for training and an additional 38,000 reviews for testing. This dataset originates from the Yelp Dataset Challenge 2015. Its primary purpose is to facilitate sentiment analysis, a task where each review is categorized into one of two classes: negative or positive sentiment.

To establish these sentiment classes, reviews with ratings of 1 or 2 stars are labeled as negative (class 0), while those with 3 or 4 stars are classified as positive (class 1). The dataset is organized into two CSV files, namely "train.csv" and "test.csv," with each containing two columns: one denoting the class index (0 for negative and 1 for positive) and the other containing the actual review text. The review text is enclosed within double quotes, and any internal double quotes are escaped with double quotes as well. My goal is to develop and assess an NLP-based sentiment classification model using this dataset, providing businesses with a scalable solution for analyzing customer reviews' sentiments and responding to customer feedbacks proactively.

## Data Dependency

Effectively handling these dependencies will guarantee the reliable and precise classification of reviews.

1. **Class Imbalance Dependency:** A balanced distribution of negative and positive sentiment reviews is necessary for unbiased model training.

2. **Rating-Sentiment Mapping Dependency:** Accurate mapping of Yelp ratings to sentiment classes (1, 2 stars as negative; 3, 4 stars as positive).

3. **Textual Quality Dependency:** High-quality, clean review text is crucial for accurate sentiment analysis.

4. **Feature Engineering Dependency:** Choice of text features based on review text characteristics is crucial for better model performance.

5. **Evaluation Metric Selection Dependency:** Selection of appropriate evaluation metrics aligned with project objectives.

6. **Data Preprocessing Dependency:** Properly preprocess text data (tokenization, stop-word removal, normalization).

7. **Temporal Considerations:** Considering temporal trends or seasonality in sentiments

8. **Training and Test Set Split Dependency:** Ensuring a representative split of data for reliable model evaluation.

9. **Ethical Considerations:** Addressing privacy, bias, and fairness in sentiment analysis to respect rights and avoid biases.

## Analysis

### Data Preprocessing and transformation

The dataset consists of user reviews categorized into two labels: positive (1) and negative (0). Through meticulous data preprocessing and transformation, the textual data has been prepared for subsequent natural language processing tasks, such as sentiment analysis. The Yelp dataset contains a diverse set of user reviews, which are valuable resources for understanding public sentiments and opinions. Effective analysis of this data requires thorough preprocessing to ensure that the text data is clean, structured, and ready for analytical procedures. A critical aspect of data preprocessing is understanding the distribution of labels within the dataset. In the training dataset, it was observed that most of the reviews are labeled as either positive or negative.

Due to hardware complexity barriers only 100,000 entries in the training set and 10,000 entries in the test set were considered. Specifically, there are 53,197 negative reviews and 46,803 positive reviews in the training dataset, and 5,367 negative and 4,633 positive reviews in the test dataset, establishing a binary classification structure.

Data quality checks confirm that both the training and test datasets are complete and contain no missing values. The training dataset consists of two columns: 'text' containing preprocessed textual data and 'label' representing sentiment labels as integer values (0 or 1). The preprocessing techniques included converting text to lowercase, removing punctuation and digits, eliminating stopwords, and lemmatization. These steps ensure that the textual data is clean and structured, which is vital for subsequent analysis.

The next phase of preprocessing involves tokenization and data padding. A vocabulary size of 10,000 and an embedding size of 32 were established for tokenization using the Keras Tokenizer. This process converts words into unique integer tokens. Data padding ensures consistent input lengths for machine learning models. The resulting dataset comprises sequences of integers, padded to a consistent length, making it suitable for sentiment analysis and other natural language processing tasks.

The meticulous data preprocessing and transformation procedures have successfully converted raw textual data into a structured format, ready for use in machine learning models for sentiment analysis and related tasks. This dataset, consisting of sequences of integers and sentiment labels, is a valuable resource for analyzing user reviews and understanding the sentiments expressed. These processes have broad applications, including market research and customer feedback analysis, enabling the extraction of valuable insights and trends from textual data.

### Explanatory data analysis

The dataset under investigation comprises 100,000 Yelp reviews, with a clear distinction between positive and negative sentiments. The analysis revealed that positive reviews make up 46.80% of the dataset, while negative reviews account for the remaining 53.20% in both the train and test parts as shown in Fig.1 (a) and (b).

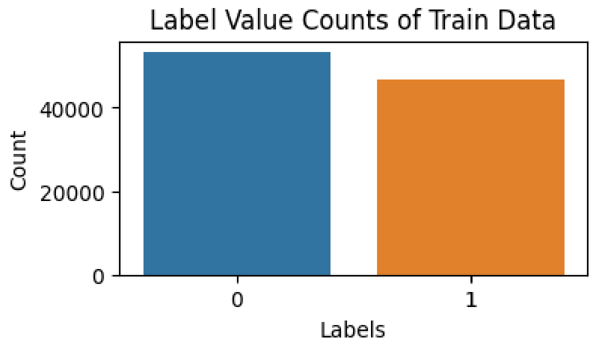


**Fig. 1 (a). Label Counts in Train data**

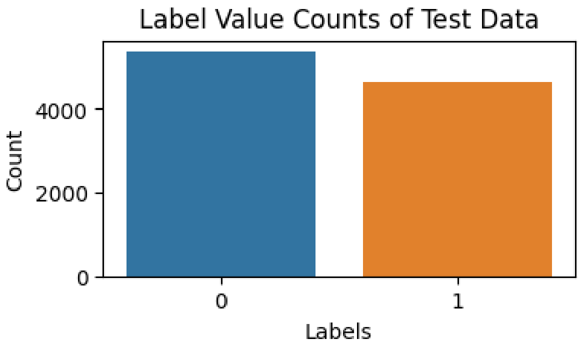


**Fig. 1 (b). Label Counts in Test data**

To gain insights into the most common words in the reviews, word clouds were generated separately for positive and negative sentiments. The word cloud for positive reviews (Shown in Fig. 2) showcased that frequently occurring terms included 'n', 'place', 'good', 'great', and 'food'. Conversely, the word cloud for negative reviews (Shown in Fig. 3) highlighted words like 'nothing', ‘never’, 'tried', 'time', and 'like', indicating common terms associated with unfavorable experiences.

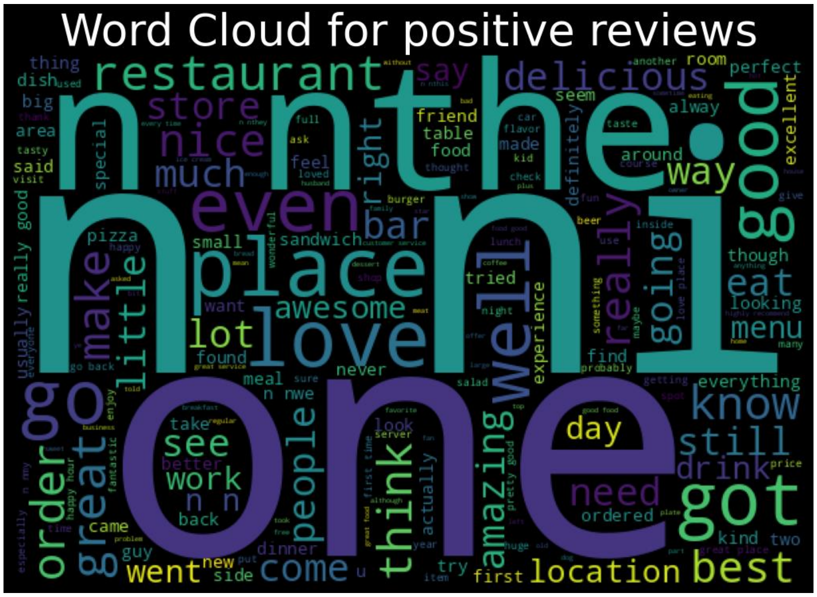


**Fig. 2. Word Cloud for Positive Reviews**

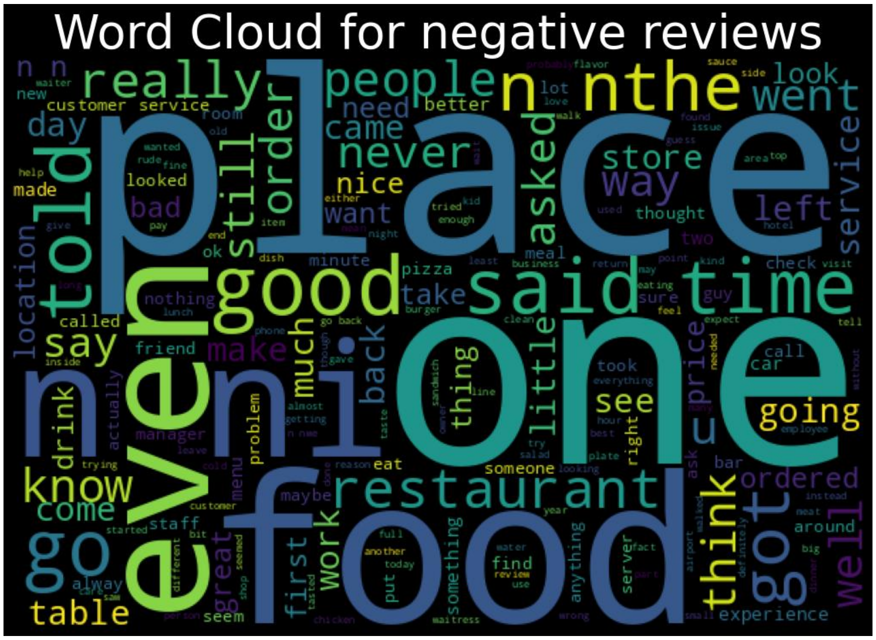


**Fig. 3. Word Cloud for Negative Reviews**

Furthermore, an examination of text length distribution in the reviews showed a right-skewed pattern as shown in Fig. 4. This suggests that most reviews are relatively short, which is essential to consider when performing Natural Language Processing (NLP) tasks, indicating that padding might be necessary to standardize text length for analysis.

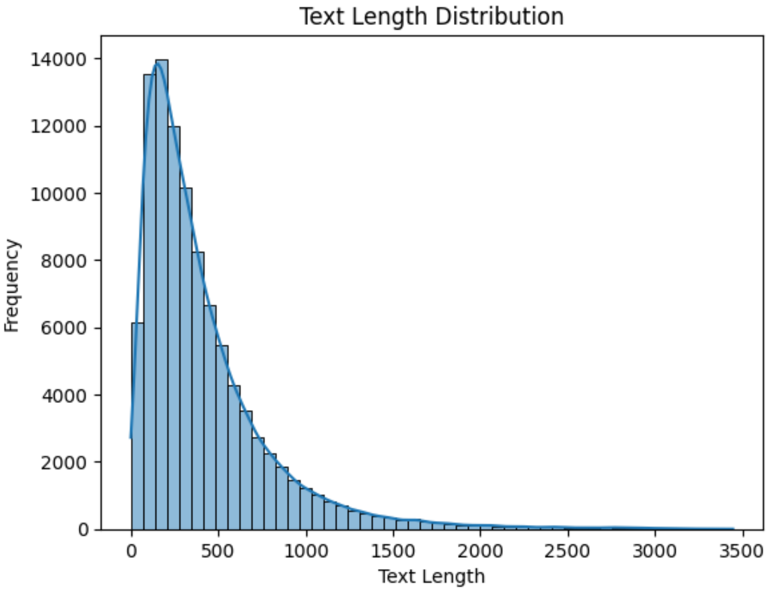
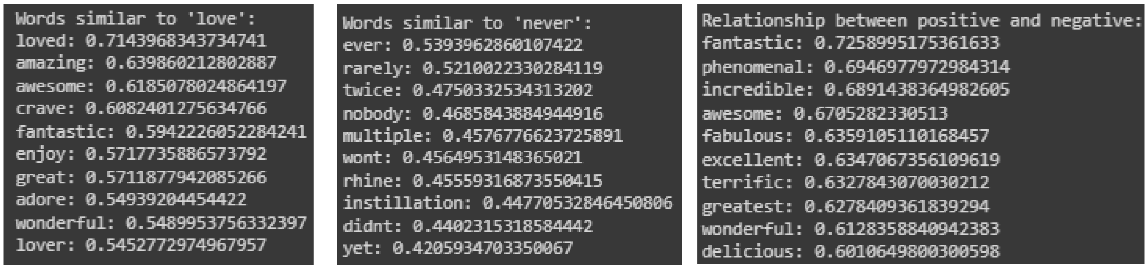


**Fig. 4. Text Length Distribution**

To understand word relationships and similarities, a Word2Vec model was trained on preprocessed text data. For example, words similar to 'love' included 'loved', 'amazing', 'awesome', and 'crave'. Conversely, words like 'never' encompassed 'ever', 'rarely', 'wont', and 'twice'. Additionally, exploring the relationship between 'amazing' and 'best' with 'never' removed yielded terms like 'fantastic', 'phenomenal', and 'incredible', indicating associations and contrasts in sentiment. Similar words with their scores are shown in Fig. 5.

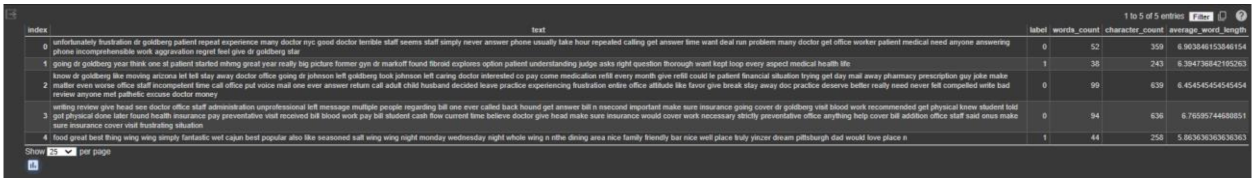


```
Words similar to 'love':
loved: 0.7143968343734741
amazing: 0.639860212802887
awesome: 0.6185078024864197
crave: 0.6082401275634766
fantastic: 0.5942226052284241
enjoy: 0.5717735886573792
great: 0.5711877942085266
adore: 0.54939204454422
wonderful: 0.5489953756332397
lover: 0.5452772974967957
```

```
Words similar to 'never':
ever: 0.5393962860107422
rarely: 0.5210022330284119
twice: 0.4750332534313202
nobody: 0.4685843884944916
multiple: 0.4576776623725891
wont: 0.4564953148365021
rhine: 0.45559316873550415
instillation: 0.44770532846450806
didnt: 0.4402315318584442
yet: 0.4205934703350067
```

```
Relationship between positive and negative:
fantastic: 0.7258995175361633
phenomenal: 0.6946977972984314
incredible: 0.6891438364982605
awesome: 0.6705282330513
fabulous: 0.6359105110168457
excellent: 0.6347067356109619
terrific: 0.6327843070030212
greatest: 0.6278409361839294
wonderful: 0.6128358840942383
delicious: 0.6010649800300598
```

**Fig. 5. Similarity of Positive and Negative words in the corpus and their relationship**



1 to 5 of 5 entries

| index | text | label | words_count | character_count | average_word_length |
|---|---|---|---|---|---|
| 0 | unfortunately frustration dr goldberg patient repeat experience many doctor nyc good doctor terrible staff seems staff simply never answer phone usually take hour repeated calling get answer time want deal run problem many doctor get office worker patient medical need anyone answering phone incomprehensible work aggravation regret feel give dr goldberg star | 0 | 52 | 359 | 6.903846153846154 |
| 1 | going dr goldberg year think one st patient started mhmg great year really big picture former gyn dr markoff found fibroid explores option patient understanding judge asks right question thorough want kept loop every aspect medical health life | 1 | 38 | 243 | 6.394736842105263 |
| 2 | know dr goldberg like moving arizona let tell stay away doctor office going dr johnson left goldberg took johnson left caring doctor interested co pay come medication refill every month give refill could le patient financial situation trying get day mail away pharmacy prescription guy joke make matter even worse office staff incompetent time call office put voice mail one ever answer return call adult child husband decided leave practice experiencing frustration entire office attitude like favor give break stay away doc practice deserve better really need never felt compelled write bad review anyone met pathetic excuse doctor money | 0 | 99 | 639 | 6.454545454545454 |
| 3 | writing review give head see doctor office staff administration unprofessional left message multiple people regarding bill one ever called back hound get answer bill n nsecond important make sure insurance going cover dr goldberg visit blood work recommended get physical knew student told got physical done later found health insurance pay preventative visit received bill blood work pay bill student cash flow current time believe doctor give head make sure insurance would cover work necessary strictly preventative office anything help cover bill addition office staff said onus make sure insurance cover visit frustrating situation | 0 | 94 | 636 | 6.76595744680851 |
| 4 | food great best thing wing wing simply fantastic wet cajun best popular also like seasoned salt wing wing night monday wednesday night whole wing n nthe dining area nice family friendly bar nice well place truly yinzer dream pittsburgh dad would love place n | 1 | 44 | 258 | 5.863636363636363 |

**Fig. 6. Five rows of the Dataset with word count, character count and average word length for each review**

Text length features, including average word length, character count, and word count, were computed and examined for their correlation with the sentiment label as shown in Fig. 6. However, the results showed minimal correlation between these features and the sentiment label, suggesting that other factors might influence the sentiment of the reviews as shown in Fig. 7.

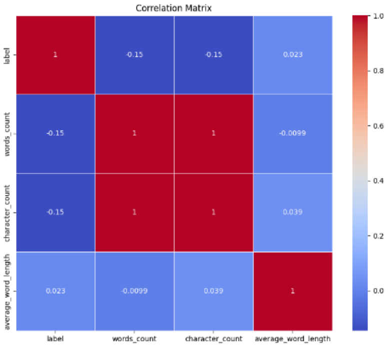


**Fig. 7. Correlation Matrix between newly formed columns**

Lastly, histograms were generated to compare the text length characteristics of positive and negative reviews. It was observed that both types of reviews mostly had word counts between 10-30. Character counts were in the range of 100-200 for positive reviews and 100-300 for negative reviews. Additionally, average word length was significantly higher in negative reviews compared to positive reviews. This pattern is shown in Fig. 8 (a), (b) and (c).

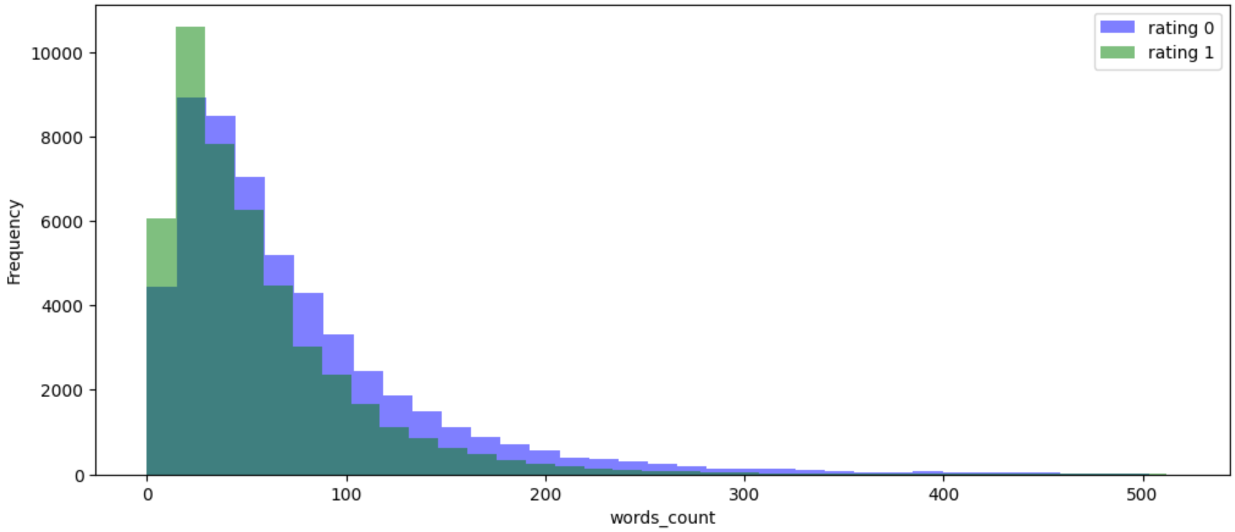


**Fig. 8 (a). Histogram plot of words count and its frequency of positive and negative reviews in training data**

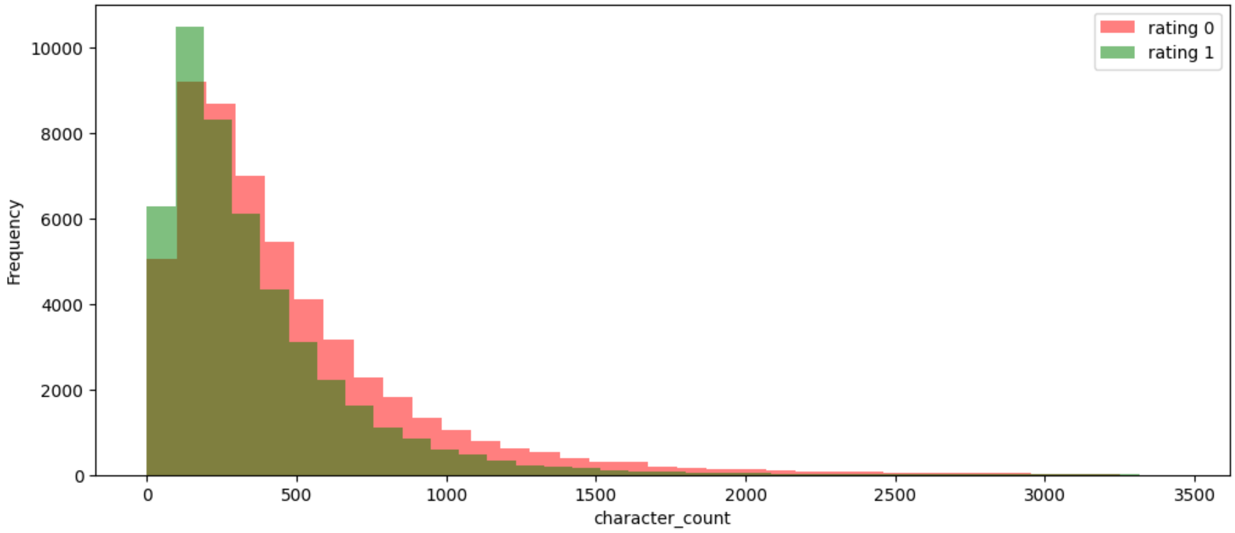


**Fig. 8 (b). Histogram plot of Character count and its frequency of positive and negative reviews in training data**

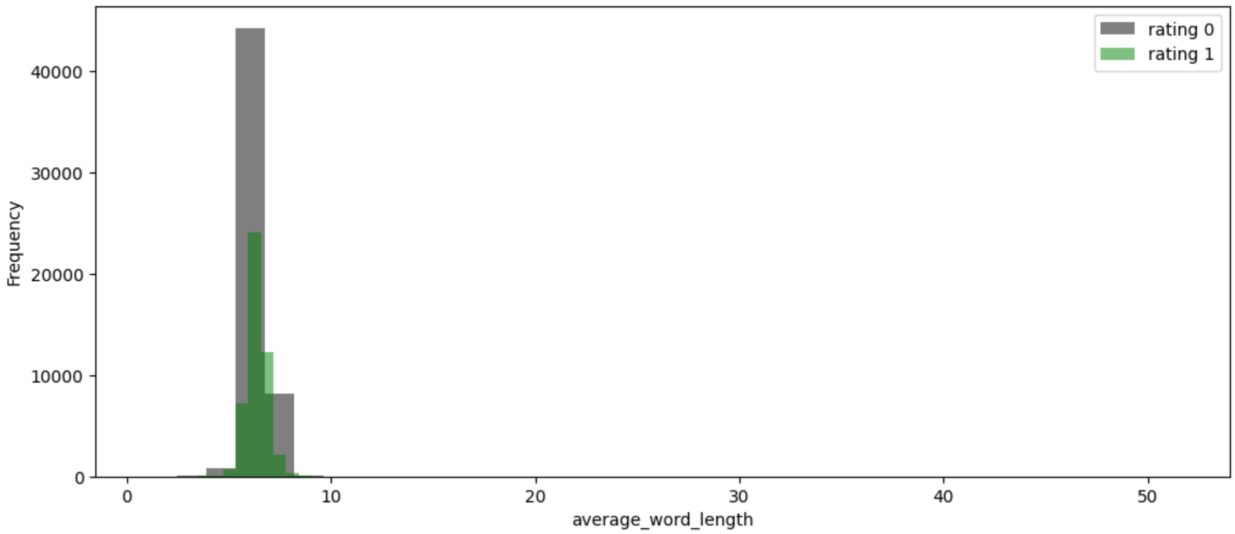


**Fig. 8 (c). Histogram plot of average word length and its frequency of positive and negative reviews in training data**

## NER Analysis:

I identified the following using a detailed NER analysis on the dataset:

Entities Identified:

Using the NER analysis, I identified various types of entities mentioned in the reviews. These entities fall into categories such as PERSON, TIME, DATE, CARDINAL, GPE (Geopolitical Entities), ORDINAL, ORG (Organizations), NORP (Nationalities, Religious, Political Groups), MONEY, PRODUCT, QUANTITY, LOC (Locations), WORK_OF_ART, FAC (Facilities), PERCENT, LANGUAGE, EVENT, and LAW. These entities can help gain insights into what aspects of the reviews are frequently discussed.

Frequent Mentions:

1. Locations (GPE): The analysis recognized many geographical locations, including cities, states, and countries. Popular mentions include Pittsburgh, Charlotte, Chicago, and various others. This suggests that reviewers often discuss the locations where they had their dining or service experiences.
2. Food and Cuisine (NORP): Various nationalities and types of cuisine were identified, such as Chinese, Italian, Mexican, and Thai. This highlights the significance of food and cuisine in customer reviews and provides insights into the preferences of reviewers.
3. Organizations (ORG): Names of restaurants and other businesses are frequently mentioned, such as Taco Bell, P.F. Chang, and McDonald's. This indicates that reviewers tend to name the establishments they visit.

4. Time and Dates (TIME, DATE): Time-related entities, including days of the week, months, and specific times of the day, are often discussed. This is crucial for understanding when reviewers visit the places they are reviewing.

Businesses can use this information to identify which entities have a significant impact on customer reviews. The NER analysis of Yelp reviews has provided valuable insights into the entities frequently mentioned in customer feedback. For example, positive mentions of specific food items or excellent service from an organization can be leveraged for marketing and improvement strategies. This information can be used to tailor strategies for improving customer satisfaction and enhancing the overall dining or service experience.

In conclusion, the exploratory data analysis provides valuable insights into the distribution of words, text length patterns, and word relationships within the Yelp review dataset. These findings can serve as a foundation for further analysis and modeling to enhance the accuracy of sentiment classification in NLP tasks.

**Feature engineering and feature selection**

The first step in feature engineering was handling the textual data present in the Yelp reviews. Since machine learning algorithms operate on numerical data, I need to convert the text into a suitable format. To achieve this, I employed a two-step approach:

1. **Count Vectorization**: I used the CountVectorizer from scikit-learn to convert the text into a matrix of word counts. This process not only transformed the text into numerical data but also represented the frequency of each word in the document. To limit the feature space, I set a maximum number of features to 5000 and removed common English stop words.

2. **TF-IDF Transformation**: After obtaining the word counts, I applied the TF-IDF (Term Frequency-Inverse Document Frequency) transformation. This technique considers the importance of a word in a specific document while also considering its significance across the entire corpus. TF-IDF helps in emphasizing words that are distinct to individual documents while de-emphasizing common words that appear in many documents.

The result of this process was two transformed datasets: one for training data and another for testing data. These datasets represented the Yelp reviews as numerical matrices suitable for machine learning.

To gain insights into the importance of individual words within the reviews, I calculated word frequencies within the training dataset. This analysis revealed the most frequently occurring words, shedding light on the language and terminology used in the Yelp reviews. The top 50 words by frequency included terms such as "food," "place," "good," "great," and "service," which are indicative of common themes in restaurant reviews.

The feature selection process was implicitly carried out during the TF-IDF transformation and word frequency analysis. TF-IDF inherently selects features that are relevant to the documents they appear in, as it assigns higher weights to words with unique information content. This means that common and uninformative words are automatically de-emphasized, reducing the dimensionality of the feature space.

Moreover, the word frequency analysis provided insights into the significance of individual words in the dataset. While all words are included as features, this analysis guides in understanding which words have a substantial impact on the sentiment expressed in the reviews. This knowledge can inform future feature selection efforts.

In summary, feature engineering and selection were essential steps in preparing the Yelp dataset for sentiment analysis. The transformation of textual data into numerical format through count vectorization and TF-IDF will allow to effectively use machine learning algorithms. Additionally, the analysis of word frequencies helped to identify the most important terms in the dataset.

## References

1. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. [Dataset]. Papers with Code. Retrieved from https://paperswithcode.com/dataset/yelp-review-polarity/

2. Zhang, X., Zhao, J., & LeCun, Y. (2016). Character-level Convolutional Networks for Text Classification. arXiv preprint arXiv:1509.01626.

3. Ghotra, Amandeep & Kudupudi, Chaitanya & Konda, Komali & Bizel, Gulhan & Gilkey, Joseph. (2020). Extraction of Aspects and Opinion Indicators from Yelp Reviews Using Different Methods of Sentiment Analysis. SSRN Electronic Journal. 10.2139/ssrn.3873335.

4. Monigatti, L. (Aug 31, 2022). "Fundamental EDA Techniques for NLP." Towards Data Science. Retrieved from https://towardsdatascience.com/fundamental-eda-techniques-for-nlp-f81a93696a75.

5. Wei, Jason, and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." EMNLP-IJCNLP 2019 short paper. arXiv:1901.11196 [cs.CL], 2019. DOI: 10.48550/arXiv.1901.11196.

6. Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

7. Barigou F (2018) Impact of instance selection on kNN-based text categorization. J Inf Process Syst 14(2):418–434