

Yelp Review Classification Using NLP

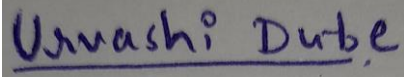
Milestone 2: Dataset Collection

Student: Urvashi Dube

+12368633298

dube.u@northeastern.edu

Percentage of Effort Contributed by Student: 100%

Signature of Student : 

Submission Date: 24th September, 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 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 1), while those with 3 or 4 stars are classified as positive (class 2). The dataset is organized into two CSV files, namely "train.csv" and "test.csv," with each containing two columns: one denoting the class index (1 for negative and 2 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 respond 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.

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.