

# Yelp Review Classification Using NLP

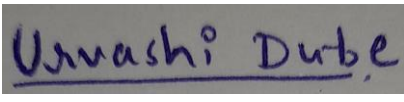
**Milestone 1:** Dataset Selection and Proposal

**Student:** Urvashi Dube

**Student Telephone Number:** +12368633298

[dube.u@northeastern.edu](mailto:dube.u@northeastern.edu)

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

**Signature of Student :** 

**Submission Date:** 17<sup>th</sup> September, 2023

### **What is the business problem? Explain a few sentences about the business problem**

I am focused on using Natural Language Processing (NLP) techniques to solve the business problem of sentiment analysis and customer feedback classification. My goal is to create a robust model that can automatically categorize customer reviews as positive or negative based on their text content. This solution is essential for businesses looking to extract insights from customer feedback, enhance customer satisfaction, and make data-driven decisions to improve their products or services.

The core issue I want to address involves efficiently analyzing and categorizing customer reviews, particularly those found on platforms like Yelp. Online reviews greatly influence consumer perceptions and buying choices, but manually processing a large number of reviews is time-consuming and prone to errors. By employing NLP and machine learning, I aim to automate this process, helping businesses track sentiment trends, pinpoint areas for improvement, and respond to customer feedback more effectively. This project aligns with the growing need for businesses to be customer-centric and data-driven in today's competitive market.

### **Explain one or two sentences on the dataset**

I'll be using the Yelp Reviews Polarity dataset from the 2015 Yelp Dataset Challenge, which contains 1,569,264 review samples, each with its text. This dataset has been pre-processed to label reviews as negative (for stars 1 and 2) or positive (for stars 3 and 4). It includes 560,000 training samples and 38,000 test samples for each sentiment category. 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.

*Link to the dataset:*

[https://paperswithcode.com/dataset/yelp-review-polarity\](https://paperswithcode.com/dataset/yelp-review-polarity)

[https://huggingface.co/datasets/yelp\\_polarity](https://huggingface.co/datasets/yelp_polarity)

### **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.

### **Hypothesis**

I hypothesize that by employing advanced NLP techniques and machine learning algorithms, I can accurately classify customer reviews into positive and negative sentiments. I believe that automated sentiment analysis will significantly reduce the manual effort required for review categorization, resulting in improved operational efficiency and timely insights for businesses.

## Objective and Solving the Problem

I plan to solve this problem by leveraging state-of-the-art NLP techniques and machine learning models. Through extensive data preprocessing, feature extraction, and model selection, I aim to build a robust sentiment classification model. This model will serve as an automation tool for businesses, allowing them to process and categorize a large volume of customer reviews accurately and efficiently. The combination of automation and data-driven insights will empower businesses to make informed decisions, ultimately enhancing customer satisfaction and business performance.

## How is your machine approach or automation going to solve the business use case?

My approach involves the following key steps:

1. **Data Preprocessing:** NLP techniques are used during data preprocessing. This includes tasks like text tokenization, removing stop words, and handling special characters. These are standard NLP preprocessing steps to prepare the textual data for further analysis.
2. **Feature Extraction:** I will use techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec, GloVe) to convert the review text into numerical features that can be fed into machine learning models.
3. **Model Selection:** I will experiment with various machine learning algorithms, such as logistic regression, Naïve Bayes, and deep learning models like LSTM or Transformer-based models (e.g., BERT), to build a sentiment classification model.
4. **Model Training:** The selected model will be trained on the labeled training data (560,000 samples) to learn the underlying patterns of positive and negative sentiments.
5. **Model Evaluation:** I will assess the model's performance using the test data (38,000 samples), measuring metrics like accuracy, precision, recall, and F1-score.

## Business Benefits:

The business stands to gain several benefits from this solution:

- Automation of sentiment analysis reduces manual effort and time spent on reviewing customer feedback.
- Businesses can quickly identify emerging sentiment trends and address customer concerns promptly.
- Access to sentiment analysis enables data-driven decision-making, leading to improved products or services.
- The ability to respond to feedback proactively can result in increased customer satisfaction and loyalty.