# CSCE 5290: Natural Language Processing
# Project Proposal

**Title:** "Unlocking Business Insights: Sentiment Analysis Through User Review Text Summarization"

**GitHub Link:** https://github.com/Kishan-Kumar-Zalavadia/Natural_Language_Processing

## Team

| Group Number | 11 |
|---|---|
| **Team Member** | **Member ID** |
| Kishan Kumar Zalavadia | 11685261 |
| Sijo Rejigeorge | 11708233 |

## 1. Motivation

In the contemporary business environment, online reviews serve as a vital source of feedback and reputation management. However, the vast volume of textual data within these reviews poses a challenge for businesses seeking actionable insights. Manual analysis is time-intensive and inefficient, leading to missed opportunities for improvement and customer engagement. Hence, the motivation behind this project is to develop an automated system that streamlines review analysis through text summarization and sentiment analysis. By offering businesses concise summaries and sentiment assessments of their reviews, we aim to empower them to make data-driven decisions and enhance customer satisfaction.

## 2. Significance

This project's significance lies in its potential to transform how businesses utilize online reviews for strategic decision-making. By automating review analysis, businesses can efficiently identify trends, patterns, and sentiments across their customer base, thereby cutting expenses and improving business growth. This enables them to prioritize areas for enhancement, respond promptly to feedback, and tailor their offerings to meet customer needs better. Additionally, from a broader perspective, this project contributes to advancing natural language processing by exploring innovative techniques for summarizing large volumes of related textual data and extracting meaningful insights.

## 3. Objectives

### 3.1 Data Collection:

Gather data related to reviews about various businesses from different platforms, primarily utilizing the Yelp dataset, which includes information about businesses, users, and reviews.

### 3.2 Data analysis:

#### 3.2.1 Data Preprocessing:

Clean the data by removing noise, performing tokenization, lowercasing, and stemming/lemmatization, handling missing values, and filtering out irrelevant information such as stop words and special characters to ensure consistency and readiness for further processing.

#### 3.2.2 Data Visualization:

Create visualizations like histograms, scatter plots, and sentiment trend charts to aid in understanding the dataset's characteristics and provide insights into review distribution and sentiment.

### 3.3 Text Summarization:

Develop a module capable of generating concise summaries for each business based on their aggregated reviews, ensuring coherence and informative content.

### 3.4 Sentimental Analysis:

Perform sentiment analysis on the generated summaries to determine overall sentiment, assigning sentiment scores (e.g., positive, neutral, negative) and analyzing sentiment trends across businesses.

## 4. Features

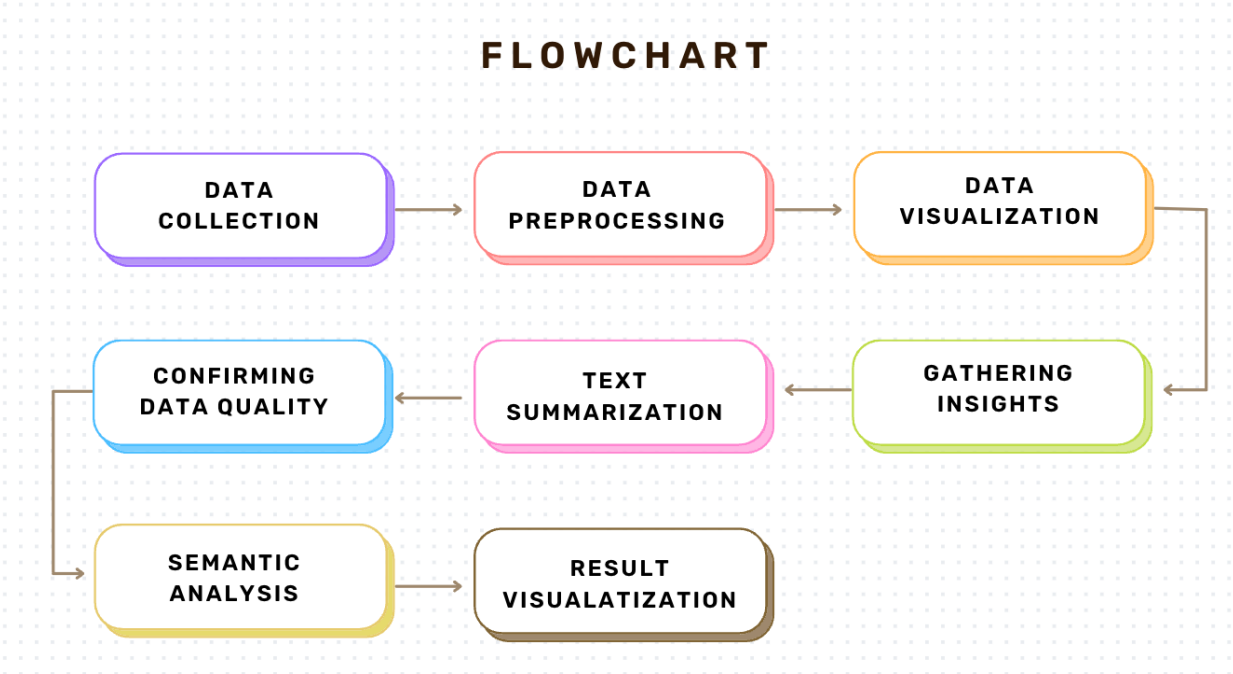
- Utilization of the Yelp dataset containing business, user, and user review data.
- Implementation of data preprocessing techniques including tokenization, stemming/lemmatization, and handling missing values while filtering out stop words.
- Utilization of data visualization tools and libraries such as Matplotlib and Seaborn.
- Text summarization by combining multiple reviews for each business to generate concise summaries.
- Sentiment analysis to determine the sentiment of the generated summaries.

## 5. Dataset

The project will primarily rely on the Yelp dataset, comprising information about businesses, users, and reviews. The dataset is in JSON format and totals approximately 8.65GB in size. It consists of multiple JSON files, including one containing business details such as names, locations, and related information, and another containing reviews submitted by various users.

## 6. Visualization

### 6.1 Workflow Diagram



## 6.2 Workflow Diagram Explanation

| Title | Explanation |
|---|---|
| **Data Collection** | Considering business review datasets from platforms like Yelp in JSON format.<br>Data includes business details, user reviews, etc. |
| **Data Preprocessing** | Clean raw data by removing noise, handling missing values, and normalizing text.<br>Tasks include tokenization, lowercasing, and removing stop words |
| **Data Visualization** | Create visualizations (histograms, word clouds, scatter plots) to understand the dataset.<br>Identify patterns, trends, and outliers to inform subsequent analysis. |
| **Gathering Insights** | Analyze visualizations to extract meaningful insights.<br>Identify popular businesses, common themes, and sentiment distribution. |
| **Text Summarization** | Combine multiple reviews for each business to generate concise summaries.<br>Capture key information and sentiment to provide an overview of customer feedback. |
| **Confirming Data Quality** | Validate the quality and accuracy of generated summaries.<br>Compare against ground truth or conduct manual review for coherence and relevance. |
| **Semantic Analysis** | Analyze summaries to determine overall sentiment expressed in reviews.<br>Assign sentiment scores (positive, neutral, negative) to each summary. |
| **Result Visualization** | Create visualizations to present sentiment analysis results.<br>Provide actionable insights for decision-making based on customer sentiment. |