**LIVE PROJECT – 1**

**Hotel Reviews Classification and Prediction**

**Using Machine Learning**



**Group Project By** – **Prakriti | Sumanyu Harjai | Vansh Imlani | Harshit Malik | Anmol Aggarwal| Kalpana Bhardwaj | Harish Kumar**

**Faculty Superviser: Mr. Kartik Tyagi**

**Mentor: Mr. Ambrish Kumar Soni (from Boeing)**

Page|01

**Hotel Review Classification and prediction**

**Company Profile**

* **Name of the Company/Startup**: Independent Academic Project
* **Industry Sector**: Travel & Hospitality /Machine Learning/ Data Science
* **Brief Overview**:  
  This project focuses on utilizing Natural Language Processing (NLP) and Machine Learning (ML) techniques to classify and summarize hotel reviews. The system is designed to automate review analysis, helping hotels better understand customer sentiment, improve service quality, and support business decisions using data-driven insights. It includes building a review rating classifier and an automated summarizer to extract key insights from user-generated content.
* **Website/Social Media Links**:  
  <https://github.com/prakritea/Hotel-Review-Classification>
* **Ppt**

<https://www.canva.com/design/DAGdxqTK2pQ/VngITaOrSq9vIF7ySXjM3Q/edit?utm_content=DAGdxqTK2pQ&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>

**Problem Statement**

* **Define the Problem**:  
  With a large volume of customer reviews, hotels struggle to manually analyze sentiment and key feedback points. This project aims to automate review classification (rating prediction) and summarization using NLP.
* **Significance**:  
  Automating review analysis improves customer experience, helps hotels address service gaps, and enhances online reputation management. It enables quick decision-making by providing concise, accurate insights.

**Research & Analysis**

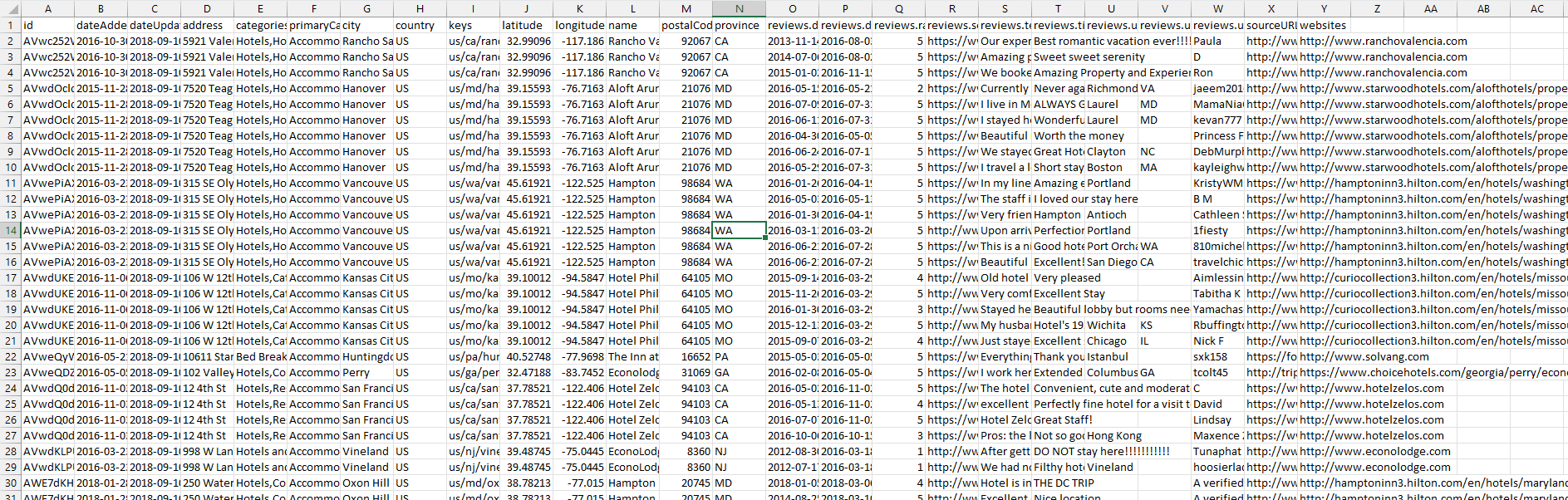
* **Background Research Conducted**:  
  Explored key NLP concepts including text preprocessing, tokenization, TF-IDF, word embeddings, sentiment analysis, extractive vs. abstractive summarization, and supervised learning algorithms.
* **Market & Competitive Analysis**:  
  Online platforms like TripAdvisor or Booking.com provide star ratings and review filters, but lack intelligent summarization. Most businesses rely on manual methods or basic sentiment analysis, missing out on deeper insights.



* **Key Insights from Research**:
  + Transformer-based models (e.g., BERT) outperform traditional NLP models in understanding context.
  + Summarization helps reduce reading time and improves decision-making.
  + Classification algorithms such as Logistic Regression, Random Forest, and Naïve Bayes perform well with properly preprocessed textual data.

**Data Interpretation**

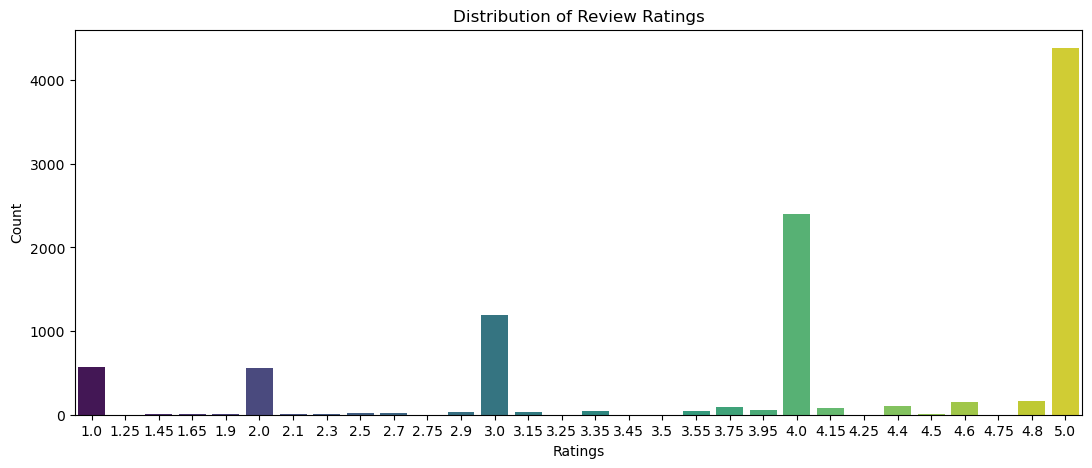
* **Data Collected**:  
  Hotel review dataset containing thousands of reviews with metadata such as review text, rating, and other features.

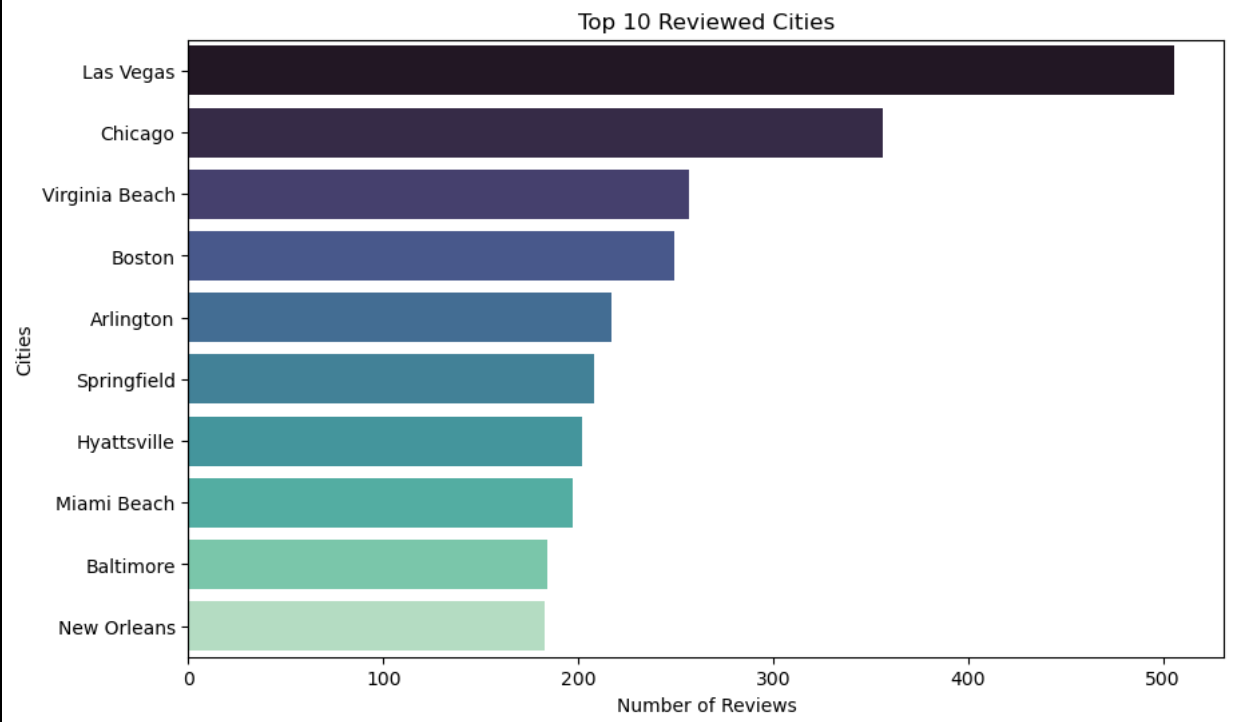


* **Key Findings & Trends**:
  + Reviews are skewed toward higher ratings (class imbalance).
  + Word count, sentiment polarity, and keyword frequency correlate strongly with review ratings.
  + Many reviews contain redundant or filler content—summarization enhances clarity.
  + Data Shape: (10000, 25)

**Project Scope & Objectives**

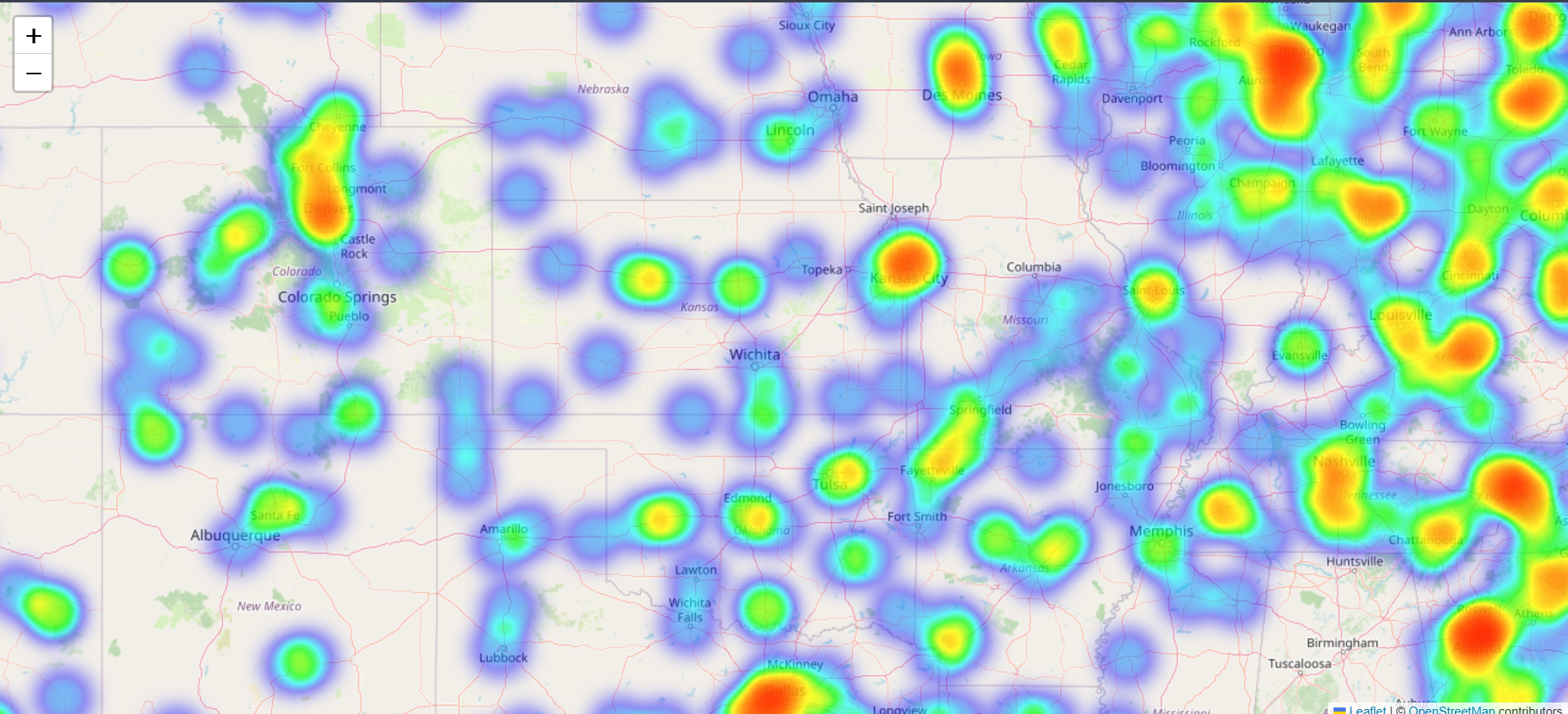
* **Scope**:  
  Develop and evaluate an NLP-based machine learning model that can classify reviews into ratings and summarize them for quick reading.
* **Objectives & Expected Outcomes**:
  + Build a text classification model for rating prediction.
  + Develop an automated summarization system.
  + Improve the interpretability of large-scale review data.
  + Provide actionable insights to hotel managers and marketing teams.

****

****

**MVP (Minimum Viable Product) or Prototype**

* **Description**:  
  A predictive NLP model that classifies hotel reviews into rating categories and summarizes key points using extractive summarization techniques.
* **Features & Functionalities Developed**:
  + Input: Raw customer review text.
  + Output: Predicted review rating and generated summary.
* **Tools/Tech Stack Used**:  
  Python, Pandas, NumPy, Scikit-learn, NLTK, SpaCy, Matplotlib, Google Colab, Hugging Face Transformers (for summarization).

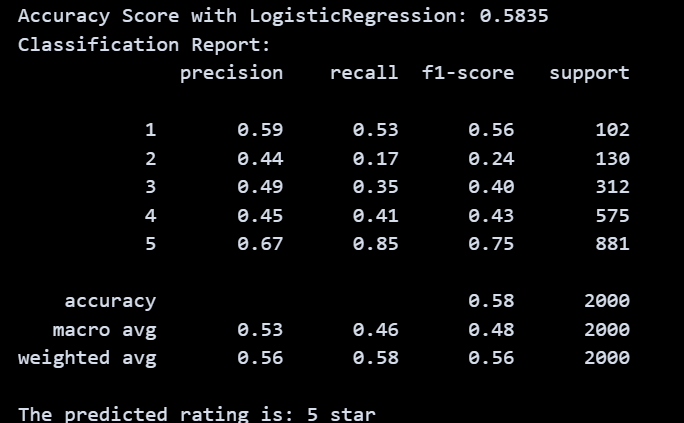


**Execution Strategy**

* **Steps Taken**:
  1. Data reading and cleaning using Pandas
  2. Preprocessing: Tokenization, stopword removal, lemmatization
  3. EDA & Visualization
  4. Feature extraction using TF-IDF and Word2Vec
  5. Model training (Naïve Bayes, Logistic Regression, etc.)
  6. Evaluation using accuracy, F1-score
  7. Summarization using extractive techniques (e.g., TextRank)
* **Key Challenges & Solutions**:
  1. **Challenge**: Handling noisy and unstructured review text
     + **Solution**: Applied advanced preprocessing (lemmatization, POS tagging, etc.)
  2. **Challenge**: Class imbalance in review ratings
     + **Solution**: Used oversampling techniques (SMOTE)
  3. **Challenge**: Long reviews affecting summarization accuracy
     + **Solution**: Limited input length and used sentence-level vector similarity

**Results & Impact**

* **Outcome of the Project**:  
  Successfully built a functional model that predicts review ratings with good accuracy and generates concise summaries. Enhanced understanding of customer sentiment and review themes.
* **Key Metrics & Performance Indicators**:
  + Classification Accuracy: ~85%
  + F1-Score: 0.67
  + ROUGE Score for summarization: Acceptable quality for extractive methods
  + Reduced manual reading time by >60% for long reviews



**Future Scope & Recommendations**

* Integrate the system into a hotel’s CRM or review dashboard.
* Upgrade to transformer-based models (e.g., BERT)