DAV Project Phase 1: Project Proposal

Course: Data Analysis and Visualization (CS-352)
Phase Lead: Zaheer Abbas (NUM-BSCS-2022-50)
Group Members:
Usaira Shahbaz (NUM-BSCS-2022-27)
Ahmad Mustafa (NUM-BSCS-2022-37)
Department: Computer Science, Namal University, Mianwali

April 2025

Project Proposal: Sentiment Analysis of Customer Reviews in Brazilian E-Commerce Dataset

1 Dataset Source and Description

Source: The dataset is sourced from Kaggle, titled "Brazilian E-Commerce Public Dataset by Olist" (https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce).
Description: This publicly available dataset contains information on over 100,000 orders from 2016 to 2018 made through Olist, a Brazilian e-commerce platform. It includes multiple CSV files covering order details, customer information, product data, and customer reviews. For this project, we will primarily focus on the `olist_order_reviews_dataset.csv` file, which contains customer reviews along with ratings and timestamps. The review column (`review_comment_message`) provides textual feedback from customers, making it suitable for sentiment analysis. The dataset is substantial, with thousands of entries, and offers a real-world context for applying data analysis and visualization techniques.

2 Research Question(s)

The primary objective of this project is to perform sentiment analysis on the customer reviews to address the following research questions:

- What is the overall sentiment (positive, negative, or neutral) expressed in customer reviews of products sold on the Olist platform?
- How does sentiment correlate with review ratings (1 to 5 stars)?
- Are there temporal trends in sentiment (e.g., changes in customer satisfaction over time)?
- Can we identify key factors (e.g., product categories, delivery times) influencing positive or negative sentiments based on the reviews?

These questions aim to uncover actionable insights into customer satisfaction and experience on the Olist platform.

3 Preliminary Thoughts on Potential Challenges and Solutions

While the dataset is rich and well-suited for this project, several challenges may arise during the analysis:

- **Challenge 1: Merging the Files**
  The dataset consists of multiple CSV files (e.g., orders, reviews, products) that may need to be merged to analyze factors like product categories or delivery times alongside reviews, which could be complex due to differing keys and large data volumes.
  Solution: We will use Python libraries like `pandas` to perform efficient merges based on common keys (e.g., `order_id`, `product_id`) and optimize memory usage by selecting relevant columns. We will also validate the merged dataset to ensure data integrity before analysis.

- **Challenge 2: Missing Data**
  The `review_comment_message` column may contain missing values, as not all customers leave textual feedback.
  Solution: We will preprocess the data by filtering out rows with missing review comments for sentiment analysis and explore imputation techniques (e.g., labeling as neutral) if necessary. A preliminary analysis of missing data patterns will also be conducted.

- **Challenge 3: Language Barrier**
  The reviews are primarily in Portuguese, which may pose difficulties in sentiment analysis if tools are English-centric.
  Solution: We will use Python libraries like `googletrans` or `deep-translator` to translate reviews into English or leverage pre-trained sentiment analysis models that support Portuguese (e.g., using `transformers` from Hugging Face).

- **Challenge 4: Subjectivity and Noise in Text Data**
  Customer reviews may contain sarcasm, slang, or typos, which could affect sentiment classification accuracy.
  Solution: We will apply text preprocessing techniques (e.g., tokenization, stop-word removal, and lemmatization) and experiment with both rule-based (e.g., VADER) and machine learning-based (e.g., BERT) sentiment analysis approaches to improve robustness.

- **Challenge 5: Visualization Complexity**
  Representing sentiment trends across time, ratings, and product categories in a clear and insightful way may be challenging.
  Solution: We will use visualization tools like Matplotlib, Seaborn, and Plotly to create interactive plots (e.g., heatmaps, word clouds, and time-series graphs) to effectively communicate findings.

4 Conclusion

This project will involve data preprocessing, sentiment analysis, and visualization of the Brazilian e-commerce dataset to extract meaningful insights from customer reviews. The final deliverable will be a research report formatted per IEEE standards, detailing the methodology, results, and conclusions.