CAPSTONE PROJECT PART 2

Problem Statement & Dataset

Daniel Yap

Customer Reviews Analysis

Natural Language Processing
(NLP)

Data Source

Kaggle

"AirBnb Listings in California"

https://www.kaggle.com/datasets/setseries/airbnb-listings-in-california

Problem Statement

Airbnb thrives on providing exceptional guest experiences. However, with a vast amount of textual customer reviews, it's challenging to manually analyze sentiment and identify areas for improvement. This limits Airbnb's ability to:

- Understand guest satisfaction: Using manual analysis, it is time consuming to determine the guest satisfaction from a large review data.

- Pinpoint improvement areas: Analyzing the large review data manually can be tedious, yet it is crucial as we can identify specific aspects of the rental experience (cleanliness, communication, location) that guests frequently praise or criticize.

- Personalize guest communication: Lack of categorization in terms of customer reviews whether their reviews are positive, negative or neutral.

Objectives

1. Understand guest satisfaction: We need an efficient method to automatically analyze the sentiment expressed in guest reviews to gauge overall satisfaction with the Airbnb experience.

2. Pinpoint improvement areas: By leveraging automation methods, we can identify specific aspects of the rental experience (cleanliness, communication, location) that guests frequently praise or criticize.

3. Personalize guest communication: Automation methods can help us categorize reviews and tailor communication strategies based on guest sentiment (positive, negative, neutral).

Methods & Models

Method:
Sentiment Analysis

Model:
Hybrid Models (Rule-based or Machine Learning)

Data Risks

Risks	Descriptions
Bias	Sampling bias, selection bias can lead to biased models
Noise	Irrelevant/misleading information can degrade the ML performance
Imbalance	Skewness may happen, leading to poor performance of model's predictions
Missing Values	Miss handling of missing values leads to biased analysis
Outliers	Distortion of ML models/statistical analysis

Data Assumptions

Assumptions	Descriptions
Random Sampling	This means each data point has an equal chance of being selected, and the selection of one point doesn't influence the selection of others.
Independence	Data points shouldn't be related to each other. For instance, if you're analyzing customer purchases, one person's purchase shouldn't influence what another person buys.
Normality	This is the assumption that the data is distributed in a bell-shaped curve, like a normal distribution. It's a common assumption for many statistical tests.
Equal Variance	This means the spread of the data is similar across different groups you might be comparing. Imagine comparing heights of men and women. Ideally, the variation (how spread out the data points are) in heights would be similar for both groups.
Accuracy & Precision	Ideally, the data points themselves should be reliable and measured correctly. Inaccurate or imprecise data can skew your findings.