INDIAN BANKS CUSTOMERS REVIEWS ANALYSIS

Contents

Indian Banks Customers Reviews Analysis Dashboard

1. Project Overview

This project analyzes customer reviews for major Indian banks to understand satisfaction levels, identify key service issues, and track trends over time. Using a structured data model and AI-based text analysis, the insights are presented in an interactive Power BI dashboard designed to support executive decision-making and improve customer experience.

2. Project Planning & Management

2.1 Objective

To transform raw bank review data into actionable business intelligence using a comprehensive data model and visualization.

2.2 Scope

- **Overall Overview:** Display key metrics such as total reviews, average ratings, and customer opinions summary.
- **Bank Comparison:** Compare major banks based on performance indicators like service quality and customer satisfaction.
- **Review Insights:** Analyze customer feedback to identify common themes, recurring issues, and positive highlights.
- **Trends Over Time:** Monitor how ratings, opinions, and review activity evolve over time to detect performance patterns.

2.3 Project Plan

Phase	Task	Milestone
1 – Data Selection	Select an appropriate dataset to be used for analysis and visualization	Dataset selected
2 – Data Cleaning & Transformation	Clean and preprocess the data by removing duplicates, handling missing values, and creating additional analytical columns	Cleaned and enhanced dataset ready

3 – Data Modeling	Develop a star schema by organizing data into fact and dimension tables	Data model structured
4 – Data Analysis & Metrics Creation	Generate analytical measures and identify key insights such as satisfaction levels and service patterns	Key insights identified
5 – Dashboard Development	Design and build an interactive Power BI dashboard with relevant visuals and KPIs	First version of dashboard
6 – Review & Refinement	Validate data accuracy, optimize visuals, and finalize dashboard presentation	Final dashboard ready
7 – Presentation & Documentation	Document the process, summarize findings, and present the results	Project completed

2.4 Risk Assessment & Mitigation Plan

1- Data Inconsistency	Conduct thorough validation and data cleaning to maintain accuracy and reliability.
2- Performance Issues	Implement a well-structured star schema and efficient DAX measures to enhance performance.
3- Visualization Overload	Design clear, focused visuals to emphasize key insights and maintain readability.
4- Misinterpretation of Results	Include descriptive labels, legends, and tooltips to ensure clarity and accurate interpretation.
5- Unbalanced Data Distribution	Recognize variations in review counts among banks when analyzing and presenting comparisons.

2.5 Dataset Overview

The foundational data for this project originates from a publicly available dataset on Kaggle, which initially contained approximately 1,000 customer reviews. This original dataset was expanded by AI to 3,000 entries to create a more robust dataset for analysis.

a) Dataset Scope and Volume

The analysis spans a large dataset of customer feedback, providing wide coverage across the financial sector:

- **Number of Columns:** 10 Columns (7 text & 3 numerical)

- **Total Records:** 3K customer reviews.
- **Entities Analyzed:** 12 distinct banks.
- **Geographic Coverage:** Feedback is collected from 325 different cities.

b) Key Features

Column Name	***Description***
author	The user who authored the review providing valuable insights into the reviewer's identity and perspective.
date	The date when the review was submitted, offering a temporal dimension to the dataset and enabling time-based analysis.
address	The geographical location from which the review was written, contributing to understanding regional trends and variations in banking experiences.
bank	The name of the reviewed bank, serving as a key identifier for the financial institution being assessed.
rating	The user's numerical assessment of the bank's service, indicating user satisfaction on a numerical scale.
review title by user	The user-assigned title to their review, summarizing the essence of their feedback in a concise manner.
review	The detailed content of the user's review about the bank, providing the primary textual data for analysis and insights.

Column Name	*Description*
bank image	The URL pointing to the bank's logo or image relevant to the review, facilitating visual associations with the bank.
rating title by user	The user-assigned title to their rating, potentially offering additional context to the rating value.
useful count	The count of users who found the review helpful, reflecting the impact and usefulness of the review among other users.

3. Data Preparation and Transformation

The initial raw dataset required cleaning and transformation to ensure data quality and suitability for analytical modeling.

3.1 Cleaning Steps

a) Date Conversion

The Date column, initially in a text-based format (e.g., *21-Mar-20*), was reformatted into a consistent numeric date format (*e.g., 21/03/2020*).

This transformation ensures uniformity across the dataset, enabling accurate time-series analysis, proper sorting, and efficient filtering operations.

b) Missing Value Handling

Missing entries and null values were strategically imputed to maintain data completeness and analytical consistency

Author: Missing values were replaced with the placeholder "Unknown user".

Bank name: Missing values were replaced with the placeholder "Unknown bank".

Rating Title by User: Blank entries were replaced with "Untitled Review" to ensure every review is characterized by a descriptive title for visualization and analysis.

Useful Count: Null values were replaced with 0, based on the assumption that a missing usefulness score implies the review has not yet received any votes.

c) Data Type Validation & Standardization

- **Text Cleaning:** Extra white spaces were trimmed from the Author and Address columns.
- **Casing:** Each word in the Author and Address columns was capitalized for standardization.
- **Column Header Standardization:** All dataset headers were standardized by capitalizing the column names and replacing underscores (_) with spaces for improved readability.
- **Column Renaming:**
  - The Author column was renamed to Reviewer to more accurately reflect the person's role in providing feedback.
  - The Date column was relabeled Review Date to clearly indicate the submission date.

3.2 Data Enrichment and Transformation

a) Rating Categorization

Two features were created to categorize the numerical rating:

Service Quality: The raw ratings were grouped into three levels of service performance:

Ratings (0 – 2.5): Bad Service

Ratings (2.5 – 3.5): Good Service

Ratings (3.5 – 5): Excellent Service

Rating Category: Customer feedback was grouped into predefined categories for sentiment analysis:

Ratings (4–5): Positive

Rating (3): Neutral

Ratings (1–2): Negative

b) Location Standardization

A new City column was created based on the Address column to unify all entries as valid cities in India. This process ensures reliability for location-based analysis. An AI-assisted tool was used for the following mapping logic:

- **Valid Cities:** Kept as valid city entries.
- **Districts / Localities / Suburbs:** Replaced with the nearest major city.
- **States:** Replaced with the capital city.
- **Variants** (old or duplicate city names): Unified by retaining only the official/modern city name.

c) Review Metrics and Scoring

- **Review Word Count:** This metric measures the total number of words in the review text. It is used as a descriptive variable to gain insights into the verbosity of customer feedback (i.e., whether the feedback is detailed or concise).
- **Influential Review:** This is a binary indicator designed to identify reviews that have achieved significantly higher engagement, thus holding greater analytical weight.

  Yes: The review's Useful Count is above the 75th percentile of all reviews.

  No: The review's Useful Count is at or below this threshold.

This indicator suggests the review resonates strongly with the customer base, as evidenced by its high vote count.

- **Bank Composite Score:** An Overall Performance Score was calculated using a weighted average of key bank performance dimensions:

  Rating (60%)

  Engagement (20%)

  Share of Voice (20%)

d) Text Analysis

An AI tool was utilized to generate a Python script that performed analysis on the text columns. The script output was an Excel file containing three sheets for detailed consumption.

Sheet Name	Source Column	Reason for Column Choice
1- Top Words per Category	Rating Title by User	The Rating Title by User column was selected due to its concise and targeted nature, which facilitated the extraction of highly specific and indicative keywords correlating directly with each rating category
2- Words Summary by Rating		
3- Word Cloud Data	Review	The Review column, containing the full text, was utilized to create the Word Cloud visualization. This approach ensures a broad and comprehensive representation of the entire customer feedback, which is appropriate for a word cloud visual where a larger volume of source text is desirable.

I. Text Exclusion Logic

To ensure the words generated are relevant to the analysis, the AI script implemented specific exclusion lists beyond standard English stop words:

- **Stop Words:** Standard English stop words (e.g., 'the', 'is', 'a') were removed from all analyses.
- **Bank and Generic Noise:** Terms directly relating to the analysis subject, such as bank names (e.g., 'sbi', 'hdfc'), business entities ('bank', 'banking'), and generic review related terms ('customer', 'service', 'account', 'review'), were removed across all sheets.
- **Vague Adjectives:** For the Top Words per Category analysis (Sheet 3), a dedicated list of vague or non-sentiment-bearing adjectives (e.g., 'other', 'major', 'different') was excluded to isolate words that specifically reflect sentiment or key attributes.

II. Data Model Connectivity

The output tables are integrated with the data model as follows:

- **Top Words per Category:** This table is connected to the primary Rating Category dimension table (often referred to as DimRatingCategory) and subsequently links to the fact table (FactReviews) in a minor snowflake schema for filtering and dimensional analysis.
- **Words Summary By Rating:** The required columns from this table were merged directly into the Rating Category dimension table (DimRatingCategory) for immediate access to the most and least frequent terms associated with each sentiment group.

4. Data Modeling

The data structure implemented adheres to principles of dimensional modeling, utilizing a Hybrid Schema (Starflake) design. This approach leverages the performance benefits of a Star Schema while incorporating the analytical depth of normalization where required, specifically for integrating text analysis results.

a) Schema Identification

The model is classified as a Hybrid Schema due to its composition:

- **Star Schema Core:** The central **`FactReviews`** table connects directly to the majority of primary dimension tables (e.g., `DimBank`, `DimDate`), forming the high-performing core of the Star Schema.
- **Snowflake Branching:** Normalization is applied to specific dimensions to connect specialized lookup tables. For instance, **`DimRatingCategories`** branches out to link the **`Top Words per Category`** analysis, creating a controlled **Snowflake structure** for 33.3. Table Integration and Relationships

b) Model Components

Table Name	Type	Description
FactReviews	Fact	Contains one row per bank review. It includes foreign keys (IDs) linking to all dimension tables, along with raw measures and attributes like Review Word Count and Useful Count.
DimDate	Dimension	Provides Time Intelligence attributes (e.g., Year, Month, Qtr, Weekday) derived from the review submission date, enabling time-series analysis.
DimBank	Dimension	Contains attributes about the bank, including the unique BankID and the calculated Composite Score.
DimReviewer	Dimension	Contains attributes about the person who submitted the feedback (e.g., Reviewer Name, Address).
DimCity	Dimension	Contains Cities derived from the Address column, linking reviews to the standardized city.
DimRating	Dimension	Contains the 1 to 5 rating stars.
DimServiceQuality	Dimension	Contains the three service quality categories (Bad, Good, Excellent) derived from the raw rating.
DimRatingCategories	Dimension	Contains the derived rating categories (Positive, Neutral, Negative). It has been enriched by merging the summary data from the Words Summary By Rating analysis (e.g., Most Frequent Word).

Table Name	Type	Description
Top Words per Category	Analysis/Lookup	A Snowflake table linked to DimRatingCategories. It contains the top words and their counts, enabling dimensional analysis of the most indicative terminology per sentiment group.
Word Cloud Data	Analysis/Lookup	Contains the top 50 most frequent words/phrases across the entire Reviews and their Occurrence Count, used exclusively for visual word cloud generation.

c) Table Integration and Relationships

The FactReviews table is the centralized entity, connected to its primary dimensions using one-to-many (1: *) relationships (Fact-to-Dimension):

- FactReviews connects to:
  - DimBank (on BankID)
  - DimDate (on DateID)
  - DimReviewer (on ReviewerID)
  - DimCity (on CityID)
  - DimRating (on RatingID)
  - DimRatingCategories (on RatingCategoriesID)
  - DimServiceQuality (on ServiceCategoriesID)
- Dimensional Branching: DimRatingCategories connects to the Top Words per Category table, allowing granular word analysis to be filtered by the sentiment category.

5. Data Analysis and Key Performance Measures

a) Advanced Text Analysis Integration

A critical phase of the project involved transforming unstructured customer feedback into structured, high-value analytical dimensions. An AI-generated Python script was executed to perform comprehensive text mining on the review data, yielding three specialized lookup tables integrated into the dimensional model.

Analysis Table	Source Column	Integration and Granularity
Word Cloud Data	Review (Full Text)	A standalone lookup table used for visual word cloud generation.
Words Summary By Rating	Rating Title by User	Merged/Enriched into the DimRatingCategories table to provide immediate vocabulary context alongside sentiment (e.g., *Most Frequent Word*).
Top Words per Category	Rating Title by User	Joined to the DimRatingCategories dimension, allowing performance metrics to be filtered based on the specific words mentioned (e.g., isolating reviews that contain the term "disappointed").

b) Key Performance Indicators and Analytical Measures

Key business metrics and analytical measures were developed in the modeling layer (DAX) to enable dynamic, accurate, and high-performance calculations for reporting and analysis.

I. Descriptive and Foundational Measures

Metric	Analytical Value
Total Reviews	The primary measure of volume; used as the denominator for all proportional calculations.
Number of Banks	Quantifies the number of unique entities analyzed.
Number of Cities	Quantifies the geographic breadth of the customer feedback collected.
Average Rating	Provides a single, high-level metric for overall bank satisfaction.
Average Word Count	Indicates the typical verbosity or detail level of customer feedback.

II. Customer Engagement Measures

Metric	Analytical Value
Total Engagement	The cumulative count of all "useful" votes, representing total validation across the reviews.
Average Engagement	The average number of useful votes per review, indicating the typical influence of a single review.
Total Influential Reviews	Counts the number of reviews that exceeded the 75th percentile for usefulness, highlighting high-impact content.
Most Useful Review	Identifies the single most resonant piece of customer feedback for qualitative inspection.

III. Bank Performance and Comparative Metrics

Metric	Analytical Value
Composite Score Normalized	Provides a single, holistic performance score for each bank, incorporating rating, engagement, and share of voice.
Share of Voice %	Measures a bank's visibility and market presence by calculating the percentage of all reviews it accounts for.
Reviews & Engagement Distribution	It shows the breakdown of sentiment (e.g., Positive Share) and engagement across all Rating Categories, revealing which segments drive the most volume or interaction.

IV. Time Intelligence Measures

These measures rely on the DimDate table and DAX's time intelligence functions.

Metric	Purpose
Month-over-Month (MoM)	Identifies short-term trends by calculating the percentage change in key measures compared to the previous month.
Year-over-Year (YoY)	Provides insights into long-term performance stability and growth by calculating the change compared to the same period in the prior year.

6. Dashboard Analysis and Key Insights Detected

The dashboards are designed to enable multi-dimensional analysis across the customer feedback lifecycle, providing actionable insights across four key areas.

a) Overview and Descriptive Analysis

This initial dashboard provides a holistic, single-view snapshot of the entire customer feedback, focusing on high-level descriptive statistics and key comparative visuals.

Key Analysis Components

- **Key Performance Indicators (KPIs):** The top bar displays foundational metrics for context, including Total Reviews (3K), Average Rating (3.40), Average Word Count (33), and Number of Cities (325).
- **Engagement and Sentiment Summary:**
  - **Customer Feedback Breakdown:** Shows the proportional distribution of sentiment (e.g., Positive at 54.7%, Negative at 29.9%).
  - **Engagement Overview:** Tracks the Total Engagement (497K) and the volume of Influential Reviews (750), quantifying audience resonance.
- **Comparative Visualizations:**
  - **Top 5 Banks by Average Review Score:** A dedicated visualization ranks the top five banks based purely on their Average Review Score (e.g., Citibank at 4.64).
  - **Ratings vs. Reviews Scatter Plot:** This critical plot maps Total Reviews (Volume) against Average Rating (Quality), enabling initial segmentation of market leaders (high volume, high quality) from niche or underperforming entities.
- **Geographic and Quality Analysis:**
  - **Reviews Density Map:** Visualizes the spatial distribution and concentration of customer reviews, assisting in location-based market analysis.

- **Reviews Split by Service Quality per Bank:** Displays a stacked bar chart showing the percentage distribution of Bad Service, Good Service, and Excellent Service ratings for each bank.

Filters and Context Slicers

- **Date Range Slider:** Allows selection of a specific period using the DimDate table.
- **City Slicer:** Filters all data by the standardized location (City), leveraging the DimCity table.

b) Bank Performance and Comparative Analysis

This dedicated page focuses on ranking and detailed evaluation of individual banks, utilizing derived metrics like the Composite Score and offering deep drill-down capabilities.

Key Visuals and Analysis

- **Bank Composite Score Ranking:** The main visual ranks all banks based on the Overall Performance Score, a weighted metric combining Rating (60%), Engagement (20%), and Share of Voice (20%).
- **Bank Key Metrics Cards:** Displays a summary of the selected bank's performance: Total Reviews, Total Engagement, Average Rating, Share of Voice, and the final Composite Score.
- **Feedback Breakdown:** A pie chart showing the Reviews Distribution by Rating Category (Positive, Neutral, Negative) for the selected bank.
- **Detailed Review Feed:** A table displaying individual review records for the selected bank, dynamically filtered by the Service Quality selection.

+ **Bank Tooltip Analysis:** A dedicated tooltip provides a concise bank summary upon hovering over a bank in the ranking, showing its Total Reviews, Average Rating, Share of Voice, and Composite Score.

Filters and Context Slicers

- **Bank Name Slicer:** Allows users to select a single bank (e.g., "Axis Bank") for detailed, page-specific filtering.
- **Service Quality Radio Buttons:** Filters the detailed review feed by the derived categories: Excellent, Good, or Bad Service.

c) Review Insights and Thematic Analysis

This page is focused on Thematic and Sentiment Analysis, integrating the output from the AI-generated text mining tables to provide vocabulary and qualitative insights.

Key Visuals and Analysis

- **Overall Vocabulary Summary:** Displays the Most Frequent Word ("Blown Away") and Least Frequent Word ("Unacceptable") across the entire dataset, sourced from the Words Summary by Rating table.
- **Reviews Word Cloud:** Visualizes the WordCloudData table, showcasing the most prominent recurring themes in customer conversations.
- **Filtered Measures based on the selected Rating Category to show:**
  - **KPI cards** that update dynamically and shows the Total Reviews, Average Word Count, and Average Engagement specific to that sentiment group.
  - **Most Helpful Review:** Highlights the review with the highest "Likes," providing necessary qualitative context (Date, Bank Name, Rating, Review Title, and Likes Count).
  - **Most Frequent Words Chart:** A bar chart displaying the top words/phrases relevant to the currently selected sentiment category, sourced from the Top Words per Category table.

Filters and Context Slicers

  - **Rating Category Buttons:** These buttons act as a field parameter, applying the filter (Positive, Neutral, or Negative) only to the lower analysis panel to generate detailed, specific insights for the selected sentiment group.

d) Time-Based Trend Analysis

This page is dedicated to Time Intelligence analysis, allowing users to track performance and sentiment fluctuations dynamically over time.

Key Visuals and Analysis

- **Sentiment Trend Visuals:** Features four area charts, tracking the chosen metric (Reviews or Engagement) across the selected time granularity: All Reviews, Positive Reviews, Neutral Reviews, and Negative Reviews.

+ Time Intelligence Tooltip (Dynamic MoM/YoY): Upon hovering over any data point, the tooltip provides change analysis based on the "View By" selection:

- If "View By" = Month, it displays Month-over-Month (MoM) change.
- If "View By" = Year, it displays Year-over-Year (YoY) change.
- Metrics include the Current Period Total Reviews, Previous Period Total Reviews, the Difference, and the calculated % Change.

Filters and Context Slicers

- Bank and City Slicer
- **"View By" Buttons (Time Granularity):** Allows users to select the visualization grain: Month or Year.
- **Metric Buttons:** Allows users to select the metric for analysis: Reviews (Volume) or Engagement (Total Likes).
- **Year and Month Slicers:** Dedicated slicers for fine-tuning the time period beyond the global range.

7. Executive Summary & Key Strategic Takeaways

This analysis successfully leveraged a Hybrid Schema data model and AI-driven text mining to synthesize 3,000 customer reviews from 325 cities into clear, actionable business intelligence. The following strategic takeaways provide a high-level summary of the most critical performance insights, service gaps, and temporal dynamics detected across the dataset.

a) Financial Sector Performance and Quality Leadership

- **Benchmark Performance:** Citibank is identified as the clear market leader, achieving the highest Average Review Score (4.64) and the leading Composite Score (60.0%). A remarkable finding is that 100% of Citibank's categorized reviews fall under Excellent Service, setting the gold standard for customer experience.

- **Critical Service Gaps:** Bank of Baroda and PNB represent the most significant areas of risk based on the Composite Score ranking and review segmentation. Analysis shows low Average Ratings (2.90 and 3.10, respectively) and a high proportion of "Bad Service" reviews.

b) Voice of the Customer and Thematic Analysis

- **Positive Dominance with Depth:** The overall sentiment remains majority positive, accounting for 54.7% of all feedback. The most frequent word is the highly emotive "Blown Away" (665 occurrences), directly correlating with 5-star reviews. Furthermore, positive reviews are, on average, longer (41 words), indicating that satisfied customers are more descriptive.

- **Root Cause of Negative Sentiment:** The most frequent negative word is "Disappointed" (152 occurrences). Critically, the most helpful Negative review for PNB explicitly cites unhelpful staff and very slow service as the primary issues, suggesting that operational efficiency and personnel training are the key areas for service recovery.

- **Neutrality as Opportunity:** The most frequent word for the Neutral segment is "Okay". The most helpful Neutral review notes a "Decent Experience" but stresses the need for improvement, highlighting an opportunity to convert ambivalent customers into highly positive advocates through targeted service upgrades.

c) Temporal Dynamics and Engagement Volatility

- **Annual Seasonality:** Review volume exhibits a clear annual cycle, peaking in both January and December. This Q4 surge indicates a critical period for resource allocation and service readiness.

- **Outsized Impact of Negative Feedback:** Despite accounting for only 29.9% of total reviews, Negative Reviews drive a high degree of audience resonance. When viewed year-over-year, the engagement (likes/votes) for Negative Reviews has a high and volatile peak (up to 44K in 2023), nearly matching the engagement level of Positive Reviews. This volatility mandates the immediate implementation of protocols to detect and rapidly respond to high-impact negative feedback.