IRCTC Sentiment Analytics - Project Report

Industry: Travel and Transportation

Name: Teesa Sharma

College: School of Management Studies, University of Hyderabad

Course: Text, Web and Social Media Analytics

Date: November 2025

Problem Overview:

The Indian Railway Catering and Tourism Corporation (IRCTC) receives thousands of customer reviews and social media posts daily. Manually understanding this data is time-consuming and ineffective. The project aims to automate the process using Natural Language Processing (NLP) and Social Media Analytics to identify sentiment, extract key discussion themes, and provide insights for service improvement.

Keywords: IRCTC, NLP, Sentiment Analysis, Text Mining, Social Media Analytics

Objectives and Methodology

Objectives

1. To analyze IRCTC customer feedback and tweets using NLP techniques.

2. To classify sentiments into positive, negative, and neutral categories.

3. To identify key discussion topics using topic modeling.

4. To develop machine learning models for sentiment prediction.

5. To visualize and interpret results through 2D and 3D analytics.

Methodology

The project follows a stepwise analytical workflow that integrates NLP, text mining, and visualization:

Data Collection -> Text Preprocessing -> Sentiment Analysis -> Topic Modeling -> Text Classification -> Visualization & Database Storage

Each step uses Python-based tools:

Text Cleaning: NLTK, TextBlob

Sentiment Scoring: VADER & TextBlob

Topic Discovery: Latent Dirichlet Allocation (LDA)

Classification: Naive Bayes, Logistic Regression

Visualization: Plotly, Matplotlib, Seaborn

Database: SQLite

Use Cases and Implementation

Use Case 1: Text Preprocessing

All IRCTC reviews were cleaned using tokenization, stopword removal, and lemmatization to prepare data for further NLP tasks.

Use Case 2: Sentiment Analysis

VADER and TextBlob models were applied to categorize reviews into positive, negative, or neutral sentiments.

Use Case 3: Topic Modeling

Latent Dirichlet Allocation (LDA) was used to identify top discussion themes such as booking experience, food quality, and refund process.

Use Case 4: Text Classification

Machine learning models (Naive Bayes and Logistic Regression) were trained on pre-labeled sentiment data. Logistic Regression achieved 91 percent accuracy.

Use Case 5: Social Media Analytics

Simulated Twitter data was analyzed to track public opinion and sentiment trends, represented through 3D visualizations.

Results and Key Insights

Results Summary

Sentiment Distribution: 60 percent positive, 25 percent negative, 15 percent neutral

Logistic Regression Accuracy: 91 percent

Key Topics: Tatkal Booking, Food Service, App Performance, Refund Process

Visual Outputs:

- Word Cloud of frequent terms

- 3D Sentiment Trend chart

- Confusion Matrix for model evaluation

- Topic Strength visualization (LDA)

Insights

1. IRCTC customers express high satisfaction with refund and reliability.

2. Technical and booking challenges remain top areas of concern.

3. Sentiment patterns are consistent between customer reviews and Twitter data.

4. NLP methods provided reliable and interpretable insights from unstructured text.

Conclusion, Future Scope, and Repository Link

Conclusion

The project demonstrates that NLP and Social Media Analytics can effectively extract actionable insights from textual data. Analysis of IRCTC reviews and tweets revealed strong customer satisfaction with certain services and recurring challenges with booking and app usability. The framework can be adapted to other public service platforms to enhance decision-making.

Future Scope

- Development of a real-time sentiment analysis dashboard.

- Inclusion of multilingual analysis (Hindi, Tamil, etc.).

- Integration with AI chatbots for live feedback collection.

Repository Submission

All relevant code, datasets, and documentation have been uploaded to a public GitHub repository.

GitHub Repository Link:

https://github.com/TeesaSharma/IRCTC-Sentiment-Analytics

Deliverables Submitted:

- Project Presentation (PPT / PDF ? 5 Pages)

- Source Code and Datasets

- SQLite Database (irctc_reviews.db)

- requirements.txt file

- Complete Documentation in GitHub Repository