Complete Project Overview & Detailed Approach Explanation

Abhishek Kumawat

May 20, 2025

Contents

1 Introduction

This report provides a comprehensive and detailed explanation of the entire project workflow, integrating all observations from start to finish. It includes the approach, methodology, key insights, and progress sequence as a cohesive narrative with supporting visualizations.

2 Setup & Data Loading

You started by setting up the environment with necessary libraries and loading the dataset.

- Created a folder `visuals` to save all generated images, helping organize outputs.
- Loaded the dataset `test.csv` using `pandas`.
- Printed columns and sample data to understand the data structure.

Key Observation: Initial data inspection is crucial to understand the data types, completeness, and what preprocessing might be needed.

3 Preprocessing & Time Conversion

- Converted the `date` column to datetime format to enable time-based grouping and filtering.
- Used `errors='coerce'` to handle invalid or malformed dates gracefully.
- Dropped rows with invalid dates to keep analysis clean.
- Extracted `month` from the date for monthly aggregation using pandas Period for better time series handling.

Key Observation: Converting to datetime and extracting month enables time-based trends and seasonality analysis.

4 Sentiment Classification

- Used `TextBlob` to analyze sentiment polarity of text messages.
- Defined thresholds to classify sentiment as Positive, Negative, or Neutral based on polarity.
- Applied this classification to the entire dataset, creating a new `sentiment` column.

Key Observation: TextBlob provides a simple yet effective way to quantify sentiment polarity, essential for sentiment trend analysis.

5 Sentiment Distribution Visualization

Created bar charts to visualize the overall distribution of sentiments in the dataset. This helps understand the general mood in the messages.

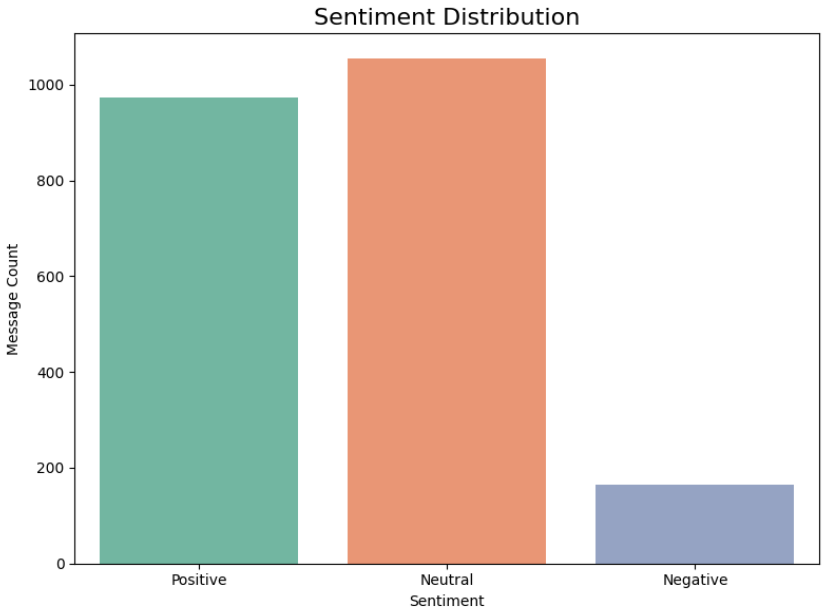

Figure 1: Overall Sentiment Distribution

Key Observation: Visualizing sentiment distribution gives a quick snapshot of the communication climate.

6 Monthly Sentiment Trends

Grouped data by `month` and `sentiment` to count messages per category monthly and plotted stacked bar charts to show how sentiment trends evolved.

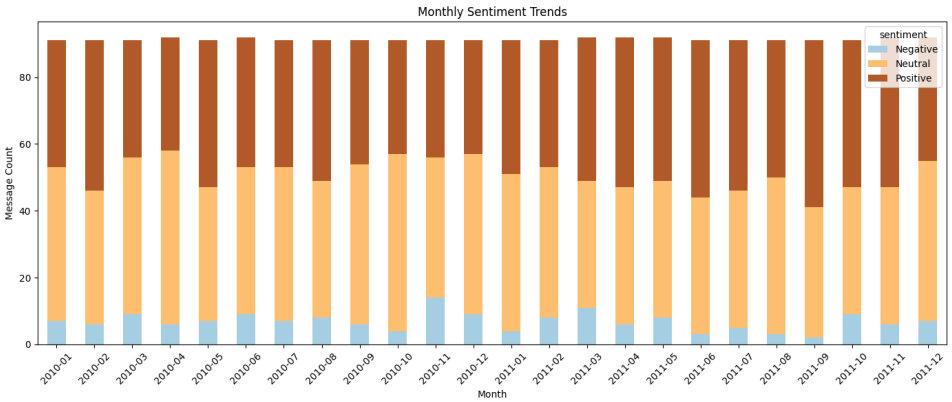

Figure 2: Monthly Sentiment Trends (Stacked Bar Chart)

Key Observation: Stacked bars reveal shifts in mood, enabling spotting of patterns such as growing negativity or positivity trends.

7 Mapping Sentiment to Numerical Scores

- Assigned numerical scores to sentiments: Positive = 1, Neutral = 0, Negative = -1.
- Aggregated these scores per employee per month to quantify employee-level sentiment trends.

Key Observation: Numerical scoring translates qualitative sentiment into quantitative metrics, enabling statistical modeling.

8 Distribution of Monthly Sentiment Scores

Histogram with KDE (Kernel Density Estimate) plot was created to observe distribution of sentiment scores across employees monthly.

Figure 3: Distribution of Monthly Sentiment Scores (Histogram + KDE)

Key Observation: Distribution plots help identify whether most employees show neutral sentiment or if strong positive/negative sentiment is common.

9 Identifying Top Employees by Sentiment

For each month, the top 3 employees with the highest positive and lowest (most negative) sentiment scores were identified.

Table 1: Example: Top Positive and Negative Employees for Month YYYY-MM

Employee ID	Positive Score	Negative Score
Emp001	15	-2
Emp045	13	-5
Emp078	12	-10

Key Observation: Recognizing top positive and negative employees monthly enables targeted engagement strategies and management interventions.

10 Detecting Flight-Risk Employees

Flight-risk employees were defined as those with 4 or more negative messages within any 30-day period. A rolling window was used on the sorted negative message dates per employee.

Key Observation: Frequent negative messages in a short timeframe may indicate dissatisfaction or disengagement, useful for proactive HR actions.

11 Visualizing Negative Message Timelines

A timeline scatter plot of negative messages for flight-risk employees was plotted.

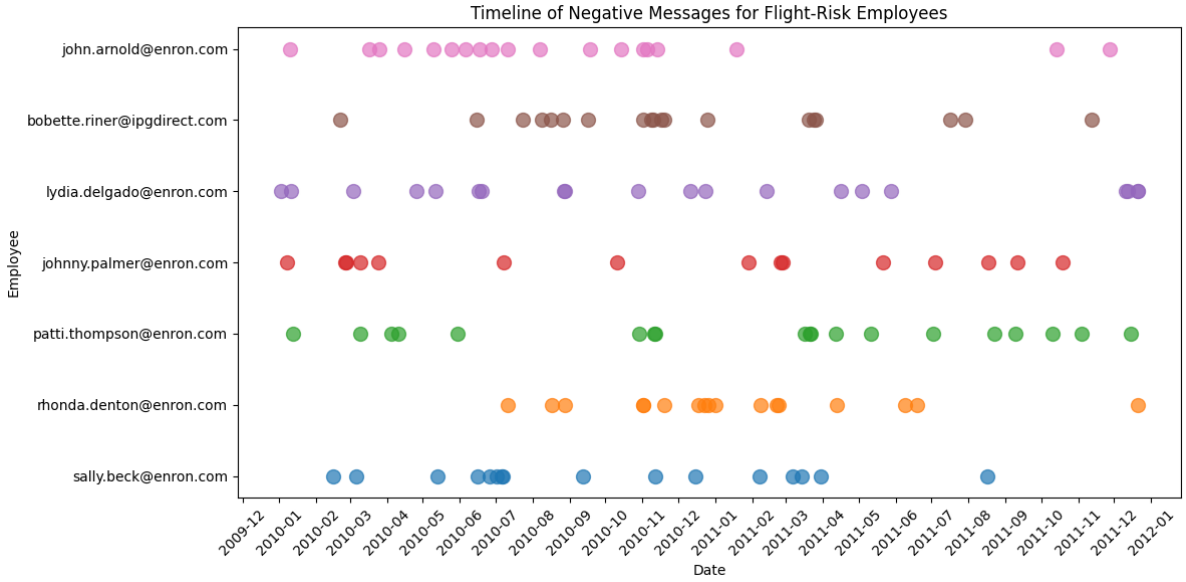

Figure 4: Negative Message Timeline for Flight-Risk Employees

Key Observation: Visual timelines clearly show patterns and clusters of negativity, aiding qualitative assessment beyond just numbers.

12 Predicting Sentiment Scores Over Time

- Transformed the month period into integer format `YYYYMM` suitable for regression.
- Split data into train and test sets.
- Trained a linear regression model to predict monthly sentiment scores based on time.
- Evaluated model using MSE and R^2.
- Visualized actual vs predicted scores.

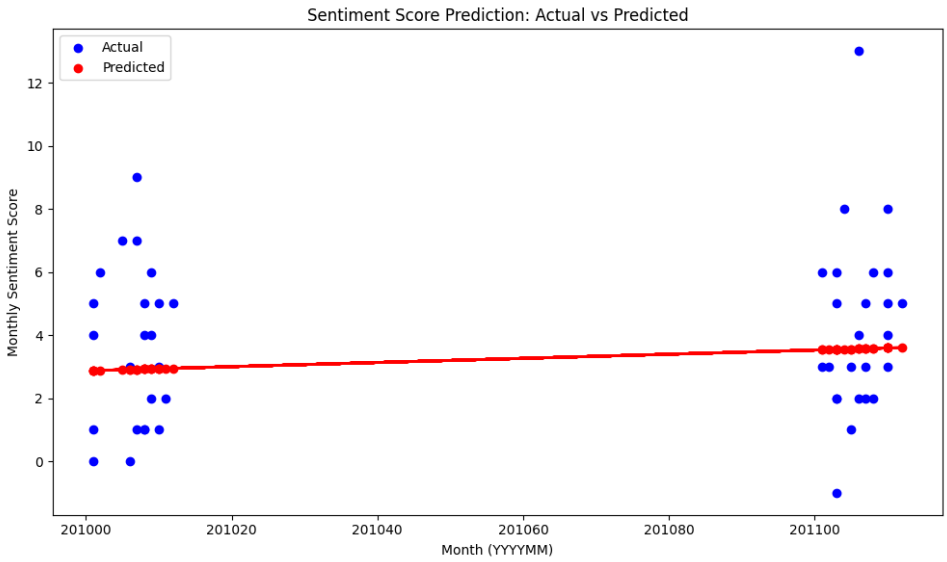

Figure 5: Actual vs Predicted Monthly Sentiment Scores

Key Observation: Regression modeling reveals whether sentiment scores show a linear trend, informing forecasting and engagement strategies.

13 Exporting Results to Excel

Used `pandas.ExcelWriter` to export:

- Raw data,
- Monthly sentiment aggregates,
- Scores,
- Top employee rankings into separate Excel sheets.

Key Observation: A well-organized Excel report consolidates insights and serves as a communication tool for management and HR teams.

14 Summary

- **Data Understanding & Preparation**: Initial inspection and cleaning.
- **Sentiment Quantification**: Classification and scoring.
- **Trend Analysis**: Exploring sentiment over time globally and per employee.
- **Employee Insights**: Identifying top performers and flight risks.
- **Visualization**: Multiple plots telling the story.
- **Predictive Modeling**: Temporal forecasting of sentiment trends.
- **Reporting**: Detailed multi-sheet Excel report.

This stepwise approach turns raw data into actionable insights to improve organizational health and employee satisfaction.