Final Report:
Employee Sentiment Analysis v2.0

Shaidatullisa Nadia Binti Saipudin

8th June 2025

Table of Contents

Introduction

This report presents a comprehensive analysis of an unlabeled dataset of employee messages. The goal was to assess employee sentiment and engagement using a combination of natural language processing (NLP), exploratory data analysis (EDA), scoring and ranking methods, flight risk identification, and predictive modeling. Each task was designed to systematically transform raw data into meaningful insights that can guide organizational understanding and proactive HR strategies.

1. Approach and Methodology

The project is divided into 6 tasks as follows:

Task 1: Sentiment Labeling
I used a sentiment analysis model to classify each employee message as Positive, Negative, or Neutral. This step was essential because it transformed raw, unstructured text into structured sentiment data, allowing me to quantify the emotional tone of communications.

Task 2: Exploratory Data Analysis (EDA)
I explored the dataset to check for missing or empty values, examine sentiment distributions, and analyze messaging trends over time. This gave me a critical understanding of the dataset's quality and revealed key patterns before moving to deeper analysis.

Task 3: Employee Score Calculation
I calculated monthly sentiment scores by assigning +1 to positive messages, –1 to negative messages, and 0 to neutral messages, summing these per employee each month. This aggregation revealed who was consistently positive or negative, setting the stage for ranking and risk detection.

Task 4: Employee Ranking
I ranked the top three most positive and top three most negative employees each month based on their sentiment scores. This helped spotlight highly engaged individuals as well as employees who might require attention or support.

Task 5: Flight Risk Identification
I flagged employees as flight risks if they sent four or more negative messages within any rolling 30-day window. This approach allowed me to surface individuals at risk of disengagement or resignation.

Task 6: Predictive Modeling
I developed a linear regression model to predict monthly sentiment scores for employees using

features derived from their messaging behavior, such as message count, word and character statistics, and sentiment ratios. The model achieved strong performance with an R^2 score of 0.71 and low prediction error (MSE = 2.45), showing that patterns in how and what employees communicate can reliably forecast sentiment trends.

2. Key Findings from the Exploratory Data Analysis (EDA)

A detailed EDA was conducted to uncover patterns in employee communication sentiment. Key insights from the analysis are summarized below:

Data Completeness: No missing values were found across the dataset, and datetime formatting was successfully applied. All 690 message records were structurally valid, with additional date-based features such as year_month, day, and message_length created for temporal analysis.

Sentiment Distribution: Contrary to expectations, positive messages were the most common, followed by neutral, with negative messages being the least frequent. This indicates a generally constructive tone in communication, though outliers still warranted deeper inspection.

Monthly Sentiment Trends: Over the 2-year period, positive sentiment consistently led each month, while negative sentiment remained relatively low but persistent. This pattern suggests stable overall engagement, with occasional dips that may relate to workload or project cycles.

Message Length by Sentiment: The longest messages were typically negative, suggesting that dissatisfied or concerned employees tend to provide more detailed responses. Neutral messages were shorter and more uniform, while positive messages showed more variability.

7-Day Rolling Sentiment Trend: Rolling averages revealed spikes and dips in sentiment at a finer resolution, particularly in neutral and positive tones. These micro-trends highlighted short-lived shifts in mood that could align with events, deadlines, or announcements.

Sentiment Ratios by Employee: Analysis of sentiment proportions per employee revealed that most individuals maintained a positive or balanced tone. A few employees, such as bobette.riner@ipgdirect.com, had a higher-than-average negative sentiment ratio (13.8%), which could flag a need for further attention.

Top Senders with High Negative Rates: Employees like lydia.delgado@enron.com and john.arnold@enron.com sent the highest volume of negative messages in absolute terms, though their overall message ratios remained within expected bounds. These employees may represent high-engagement roles under pressure or long-term friction points.

Negative Sentiment Heatmap: A heatmap of monthly negative message volume revealed clusters of negativity among specific employees. For example, Bobette and Lydia showed recurring negative messaging in several months, suggesting chronic dissatisfaction or communication issues.

See some visual summaries below:

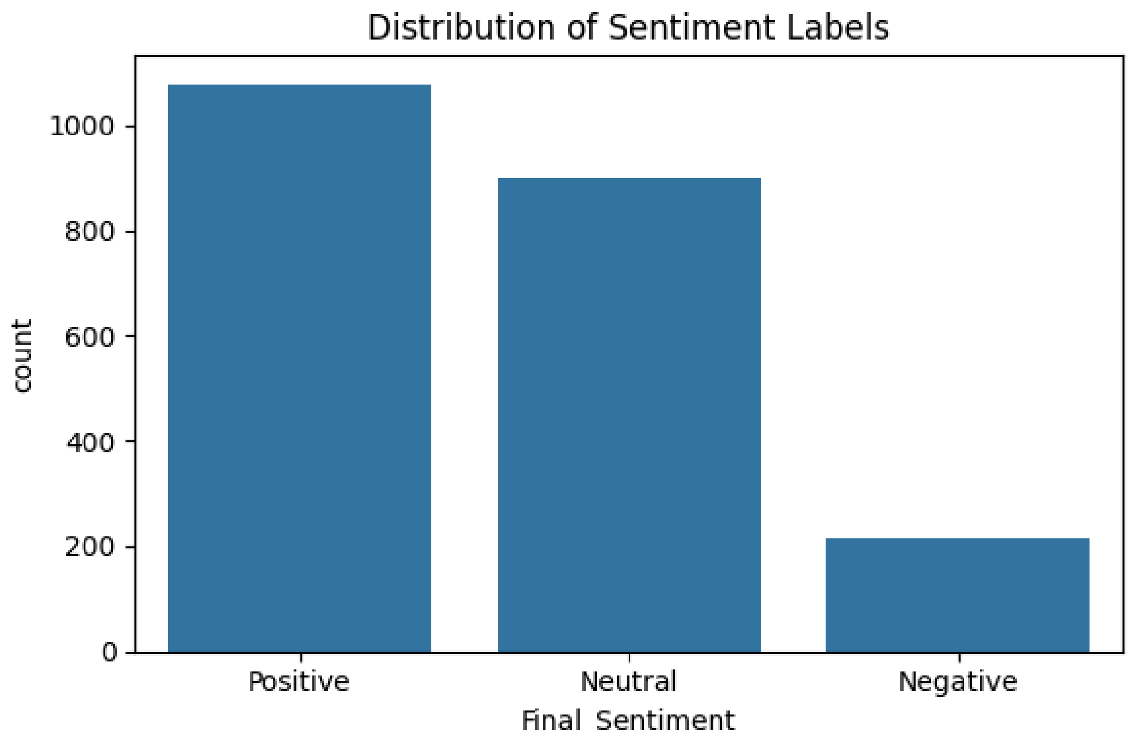

Figure 1: Sentiment Distribution

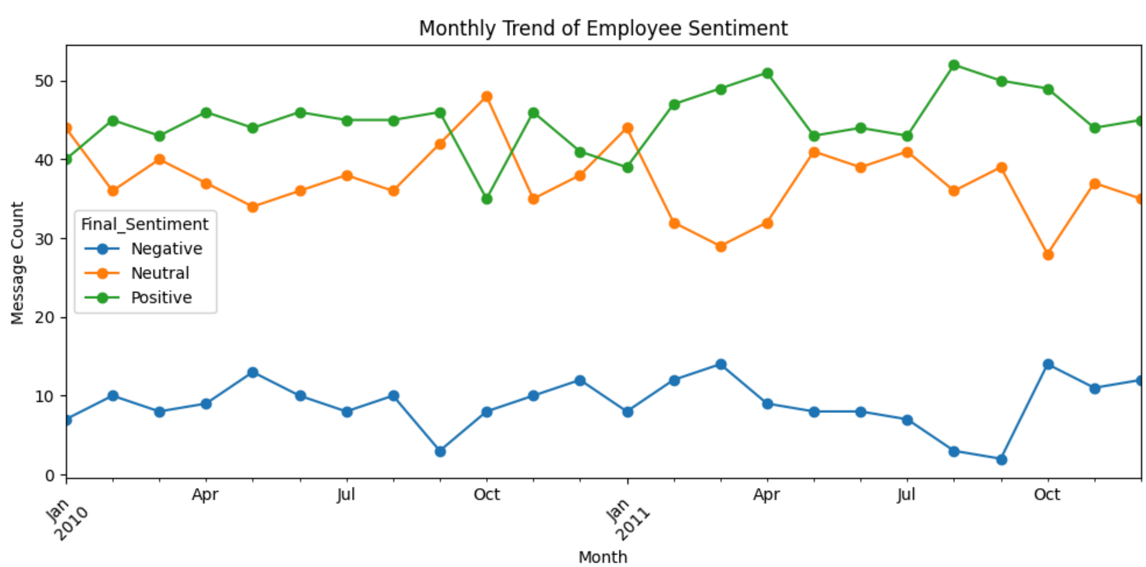

Figure 2: Monthly Trend of Employee Sentiment

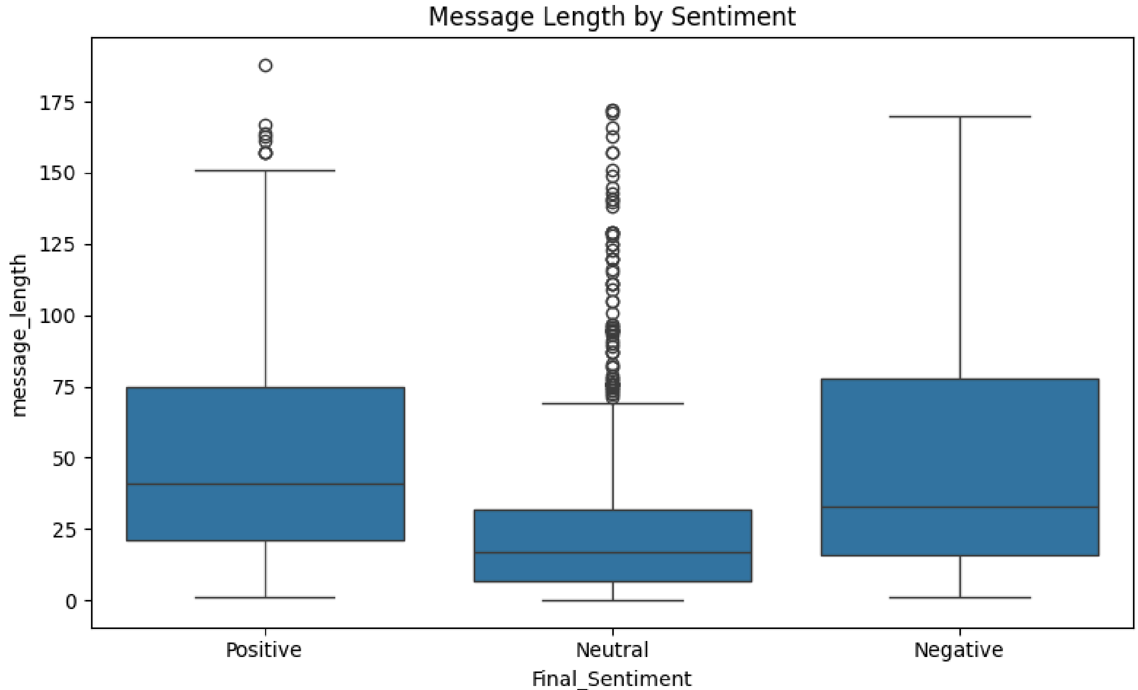

Figure 3: Message Length by Sentiment

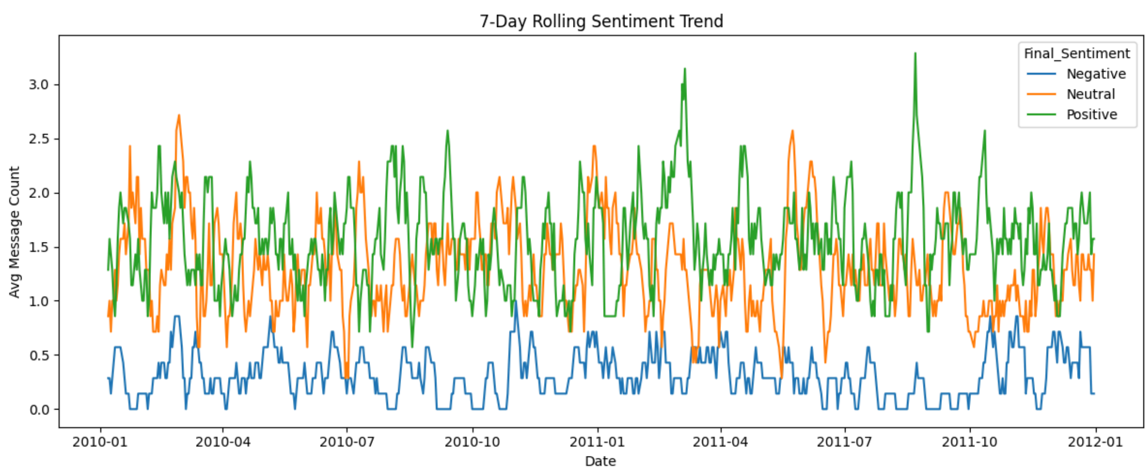

Figure 4: 7-Day Rolling Sentiment Trend

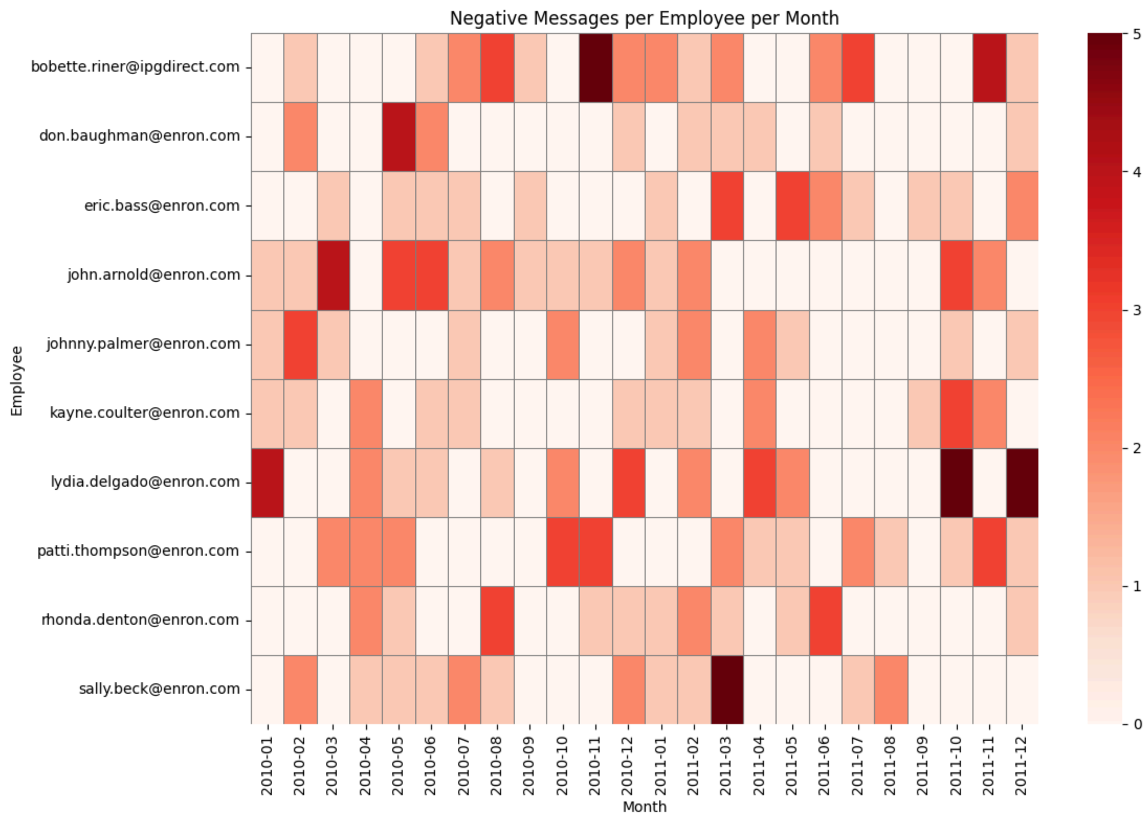

Figure 5: Negative Messages per Employee per Month

3. Employee Scoring and Ranking

In this task, I calculated a **monthly sentiment score** for each employee based on their labeled messages. This score reflects overall emotional tone and engagement in employee communication.

To compute the score, each message was assigned a value:
+1 for Positive, **–1** for Negative, and **0** for Neutral.
These values were then **summed per employee each calendar month**, creating a rolling sentiment profile that resets monthly for consistency.

This scoring method allowed me to identify which employees consistently expressed positive communication, and which ones tended to be more negative over time. The score provided a quantitative way to compare employee sentiment trends on a monthly basis.

The scores were saved into a new dataset and later used for **employee ranking (Task 4)** and **flight risk detection (Task 5)**. This step was critical in transforming individual message labels into actionable employee-level insights that could be tracked and compared month over month.

Below are the example visuals:

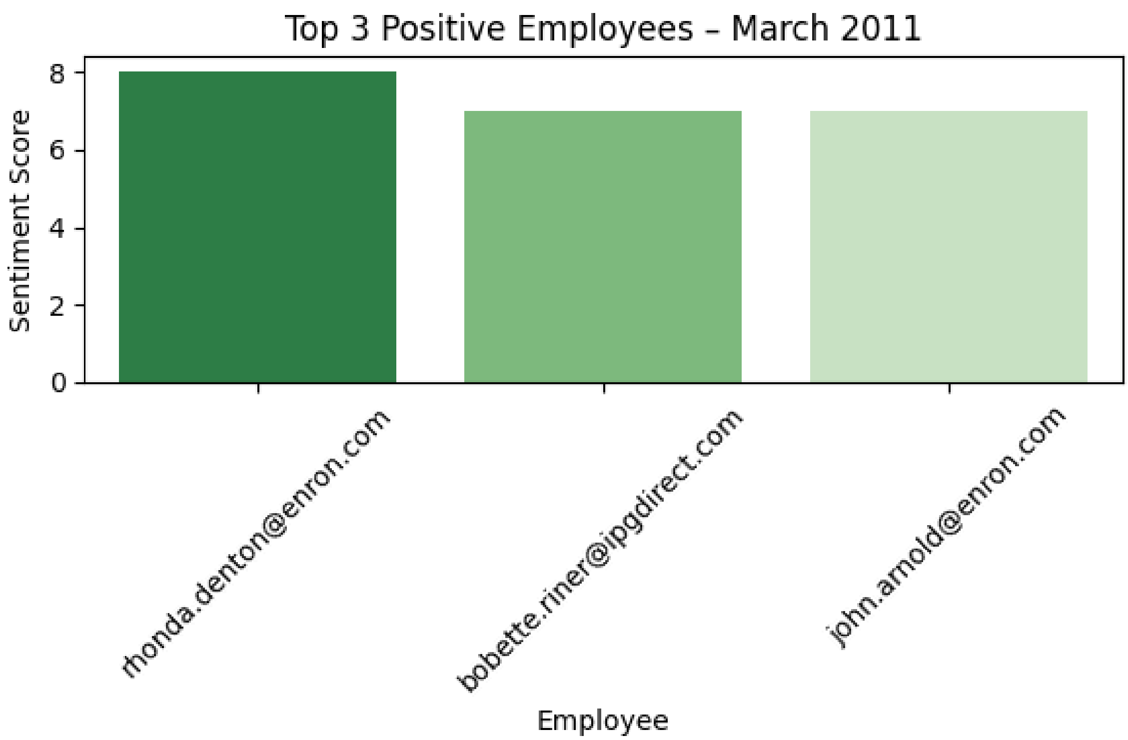

Figure 6: Top 3 Positive Employees of the month

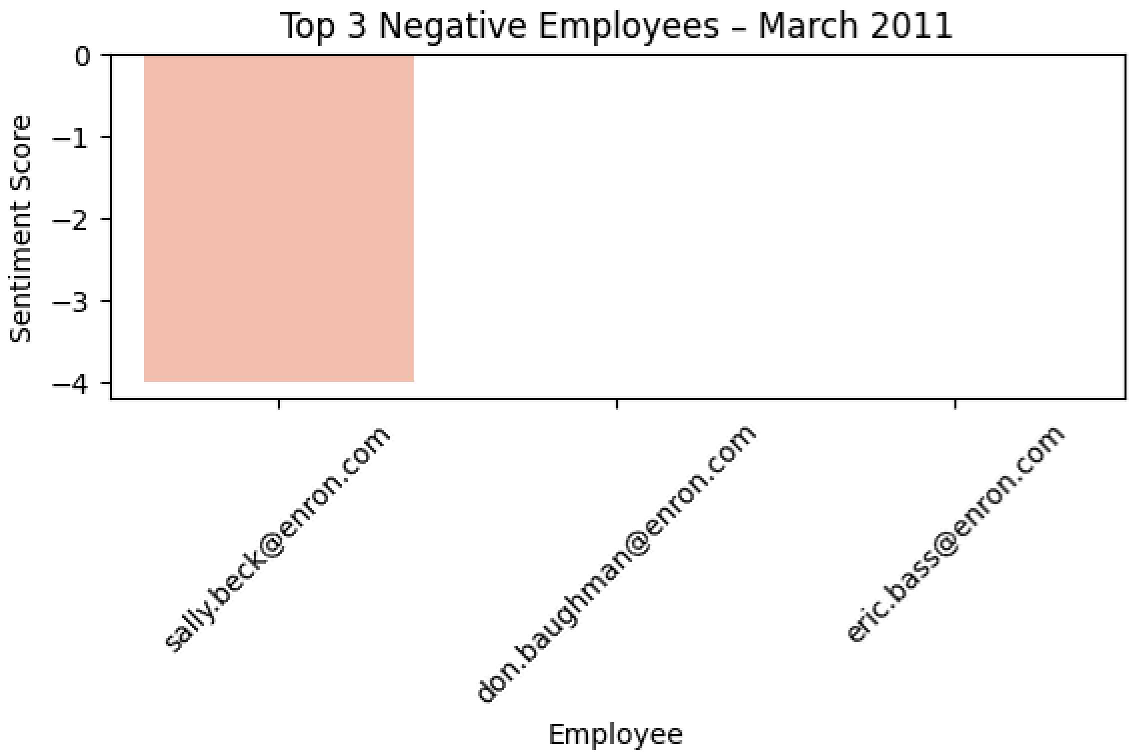

Figure 7: Top 3 Negative Employees of the month

4. Flight Risk Identification

- In this task, I analyzed employee messages to identify individuals at risk of disengagement or resignation. I focused specifically on those who exhibited a pattern of negative communication behavior.
- Using a **rolling 30-day window**, I flagged any employee who sent **4 or more negative messages within any 30-day span**. This method ensures that patterns are captured over time, rather than being restricted to calendar months.
- Employees who met the threshold were labeled as **flight risks**. I then counted the total number of negative messages each flagged employee sent, helping to prioritize who may need closer monitoring or support.

Visualizations and Interpretation:

1. Bar Chart – Flight Risk Employees by Negative Message Count

This chart shows the **total number of negative messages** sent by each employee flagged as a flight risk. Employees like **bobette.riner@ipgdirect.com**, **lydia.delgado@enron.com**, and **john.arnold@enron.com** were among the top contributors to negativity, with over **30 negative messages each**. This visualization makes it easy to spot individuals who not only triggered the flight risk threshold but also sustained negative sentiment over time.

2. Line Chart – Rolling 30-Day Negative Message Trend

This line chart tracks the **rolling 30-day count of negative messages** for a single flagged employee: bobette.riner@ipgdirect.com. Each point represents the number of negative messages sent in a 30-day period, sliding daily. Peaks reaching **4 or 5 messages** confirm repeated violation of the flight risk threshold. Notably, this employee shows **recurring spikes** across multiple months, indicating an ongoing pattern of disengagement, not a one-time event.

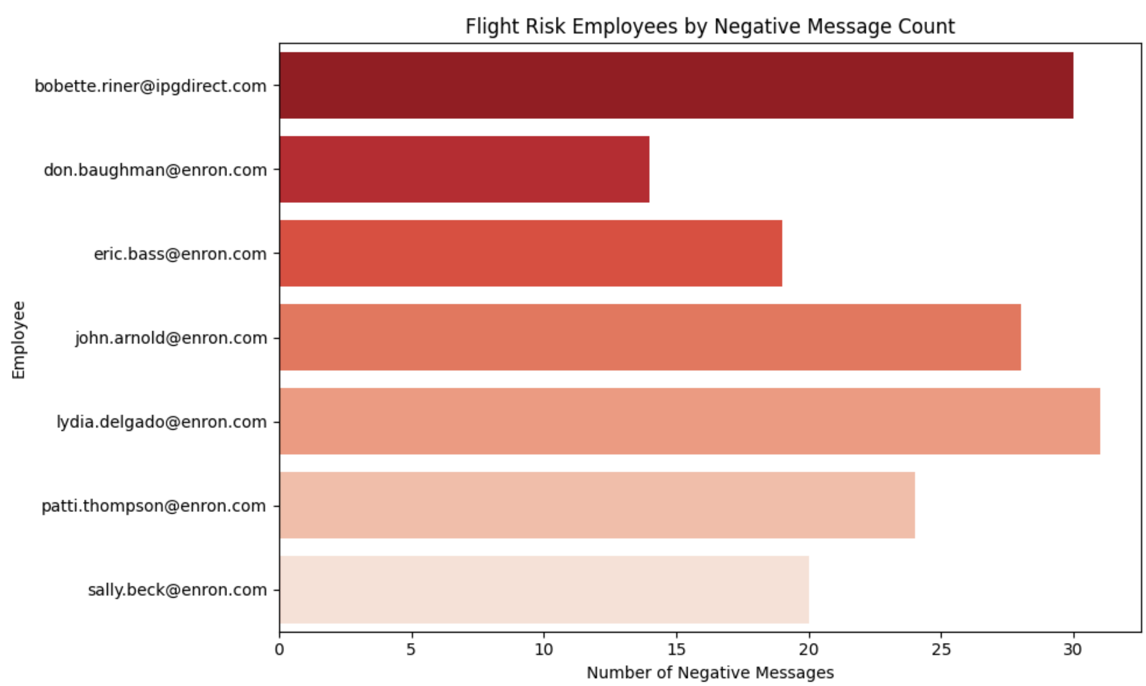

Figure 8: Flight Risk Employees by Negative Message Count

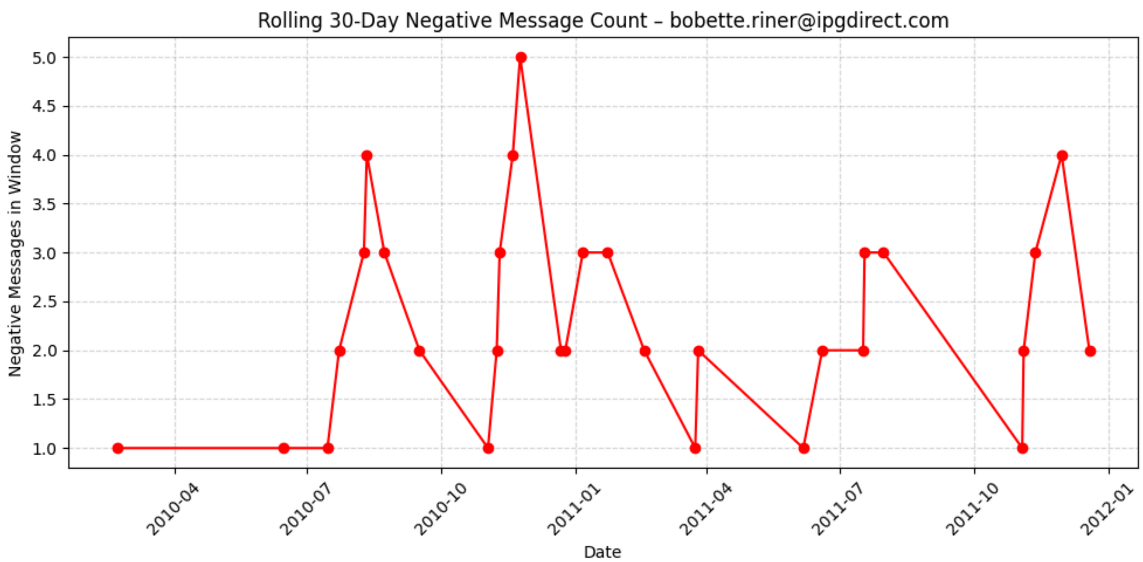

Figure 9: Rolling 30-Day Negative Message Count (one employee example)

5. Predictive Model Overview and Evaluation

The objective of Task 6 was to develop a predictive model capable of estimating an employee's monthly sentiment score based on their message behavior and sentiment distribution. A linear regression model was chosen for its interpretability and efficiency.

To train the model, I used features engineered from each employee's monthly communication patterns. These included:

- Message metrics: volume, total words, character counts, and density
- Sentiment ratios: proportions of positive, neutral, and negative messages

The model was trained and evaluated using an 80/20 train-test split. Evaluation metrics show that the model performed strongly, with:

- R^2 Score: 0.7075 – indicating that the model explains approximately 71% of the variance in sentiment scores
- Mean Squared Error (MSE): 2.45 – showing low average prediction error

The scatter plot of Actual vs Predicted Sentiment Scores confirms that the model fits the data well, with most predictions aligning closely with the diagonal reference line (perfect prediction).

Feature Importance (Model Coefficients):

- positive_ratio (+4.49) and negative_ratio (–4.69) were the most influential predictors, which aligns with the scoring logic (positive messages raise the score, negative messages reduce it).
- message_count also had a positive effect, suggesting that high message activity correlates with higher sentiment scores.
- Interestingly, avg_message_length and words_per_message had negative coefficients, indicating that longer or denser messages may slightly reduce sentiment, possibly due to being issue-driven or overly formal.

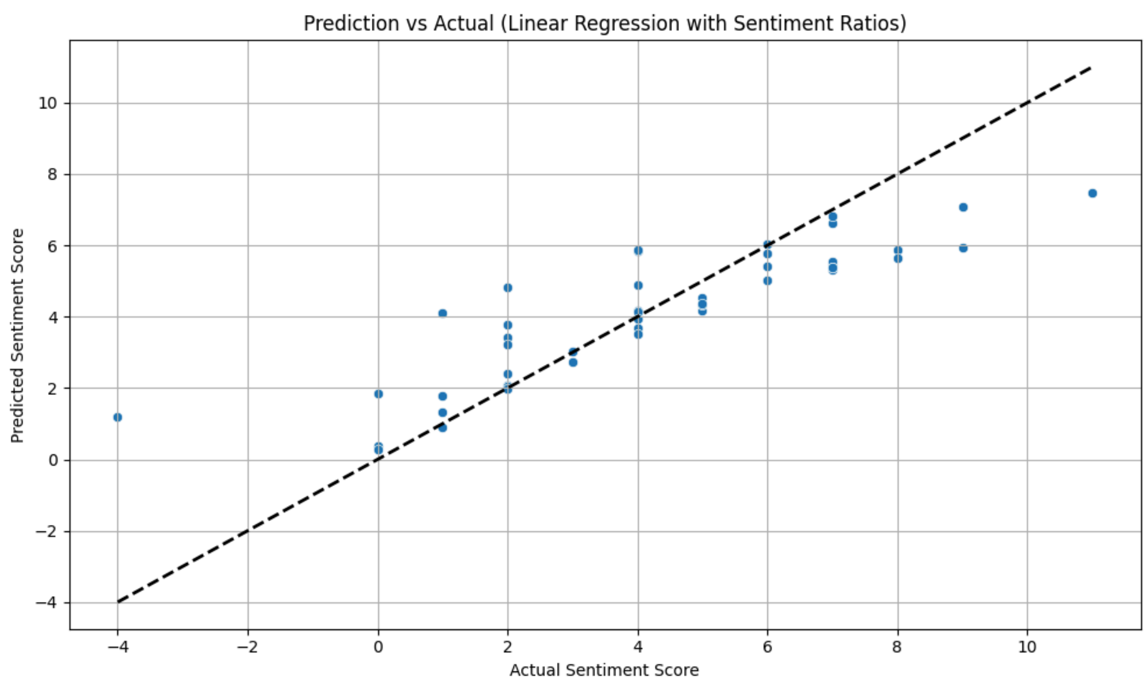

Figure 10: Predicted Vs Actual Sentiment Score Sum

Conclusion

This project turned a raw dataset of employee messages into meaningful insights about how people communicate at work. Using a mix of sentiment analysis, data exploration, scoring, risk identification, and predictive modeling, I built a complete workflow that helps understand employee engagement and sentiment over time.

In summary, this project proved that it's possible to understand and even forecast employee sentiment using internal communication data. The results can help HR teams monitor morale, identify potential concerns early, and make more informed decisions to support a healthy and engaged workplace.