Final Report:
Employee Sentiment Analysis

Shaidatullisa Nadia Binti Saipudin

31st May 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 using features like

average word count and average sentiment score per message. This modeling helped me explore whether sentiment trends could be forecasted.

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

The EDA phase revealed several important insights:
- There were no missing values in the dataset, but 30 message bodies were empty.
- Negative messages dominated the dataset, followed by positive messages, with very few neutral messages.
- Employee message counts were generally uniform, but negative message shares hovered between 48% and 60%.
- Monthly sentiment trends showed a consistent pattern where negative sentiment outweighed positive sentiment.

See visual summaries below:

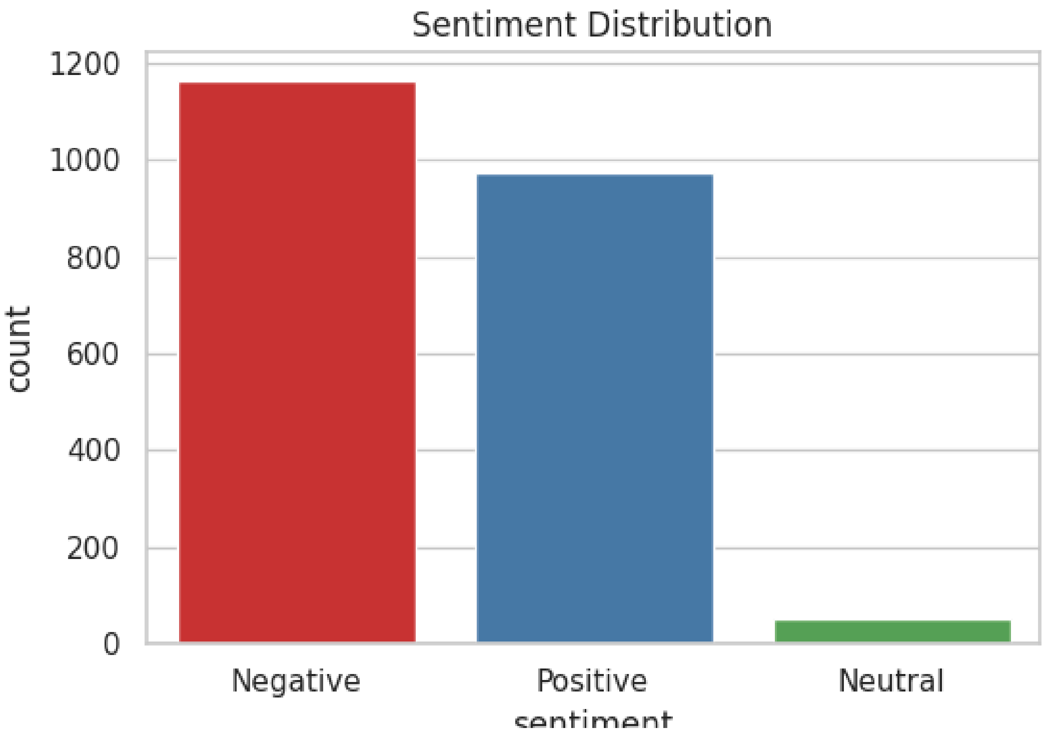

Figure 1: Sentiment Distribution

In Figure 1, the sentiment distribution chart shows that negative messages dominate the dataset, followed closely by positive messages, with neutral messages making up only a very small fraction. This imbalance suggests a workplace culture with consistently emotionally charged or dissatisfied communication.

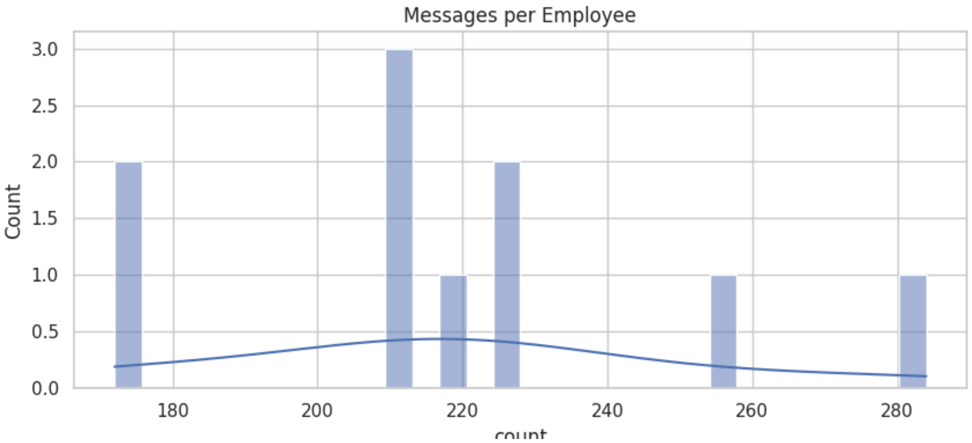

Figure 2: Messages per Employee by Count

In Figure 2, the histogram of messages per employee reveals that most employees sent a similar number of messages, with no extreme outliers. This indicates that message volume itself is relatively uniform and not a primary factor driving differences in sentiment.

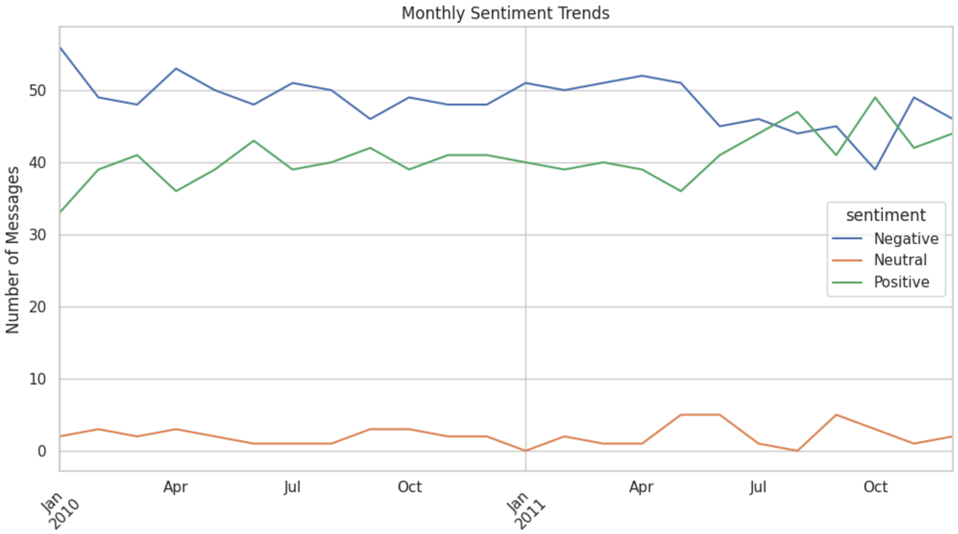

Figure 3: Monthly Sentiment Trends by Number of Messages

In Figure 3, the line chart of monthly sentiment trends demonstrates that negative sentiment consistently leads over time, while positive sentiment tracks slightly lower, and neutral sentiment remains almost flat. This stable but negative trend points to systemic issues rather than temporary fluctuations.

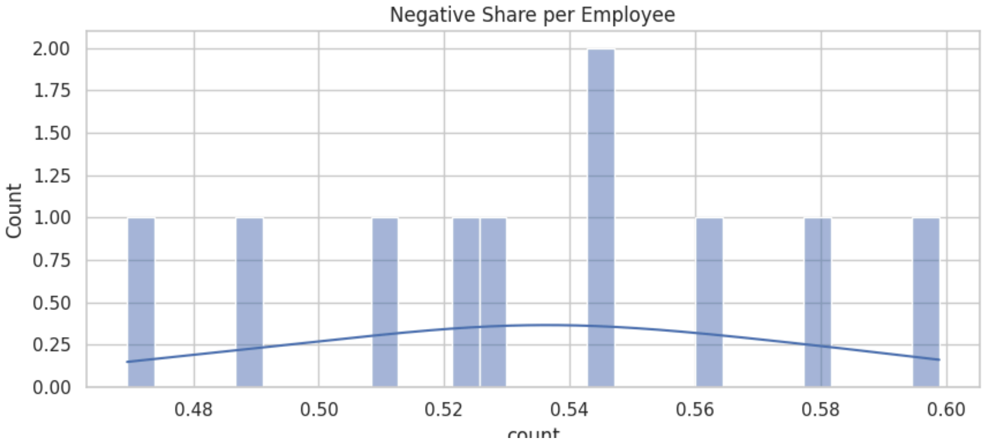

Figure 4: Negative Share per Employee by Count

In Figure 4, the negative share per employee chart shows that most employees have a negative message proportion clustered between 48% and 60%. This suggests that negative communication is not isolated but rather widespread across the employee base.

3. Employee Scoring and Ranking

Monthly sentiment scores were calculated by summing the sentiment values per employee (Positive = +1, Negative = -1, Neutral = 0). We identified the top three most positive and most negative employees each month as we can see in Figure 5 and Figure 6. This ranking helps pinpoint highly engaged employees and those who might require managerial attention.

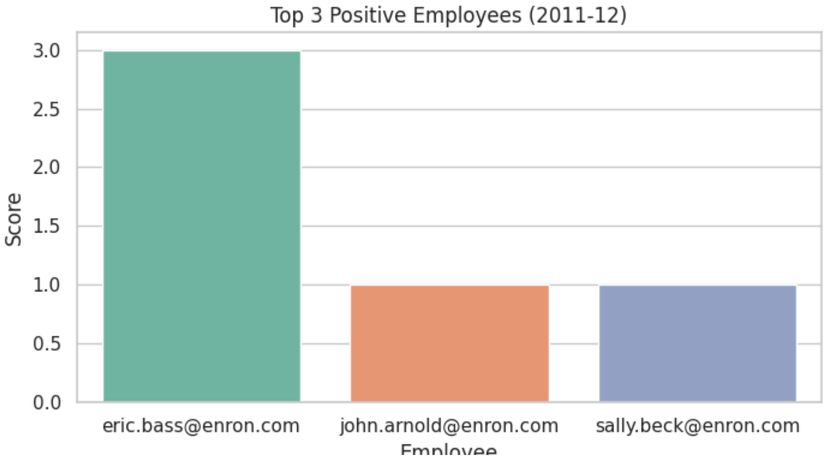

Figure 5: Top 3 Positive Employees of the month

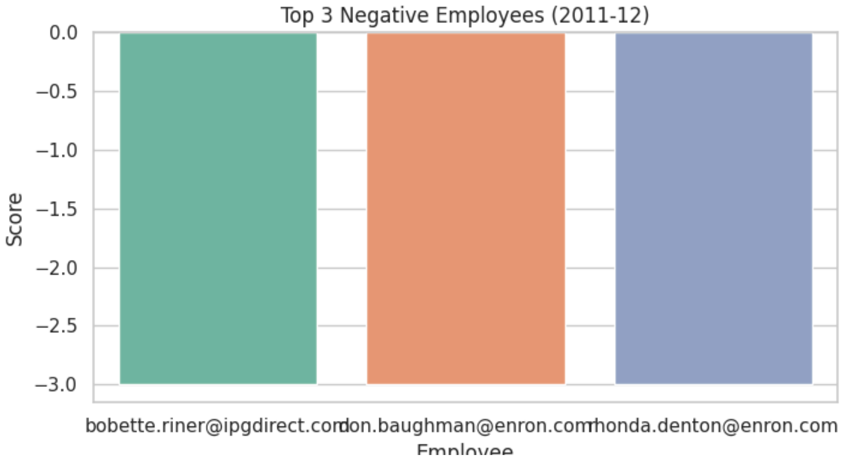

Figure 6: Top 3 Negative Employees of the month

4. Flight Risk Identification

An employee was flagged as a flight risk if they sent 4 or more negative messages in any rolling 30-day period. This process helps identify individuals who may be at risk of disengagement or resignation.

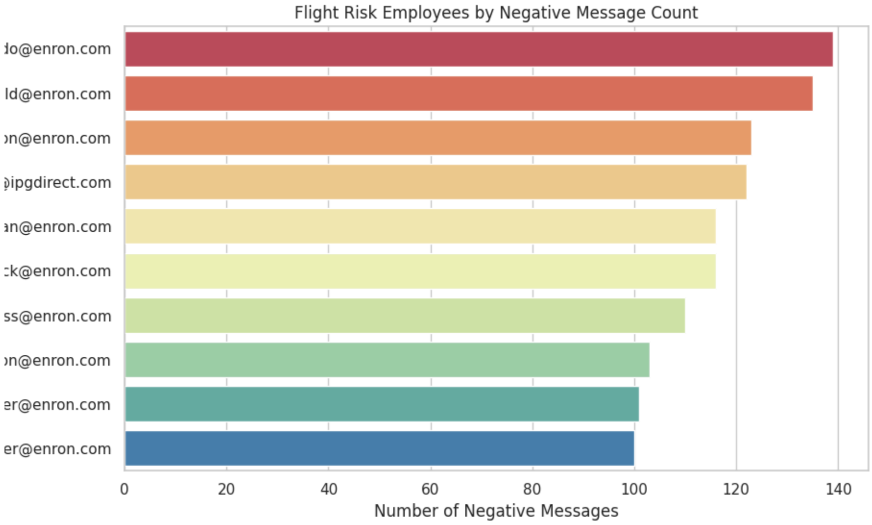

Figure 7: Flight Risk Employees by Negative Message Count

In Figure 7, the flight risk bar chart ranks employees by the number of negative messages sent, highlighting those most at risk of disengagement. The top employees have significantly higher negative counts, making them critical targets for intervention.

5. Predictive Model Overview and Evaluation

I developed a linear regression model using two sentiment-focused features:
- Average word count per message
- Average sentiment score per message

The model aimed to predict the monthly sentiment score sum. The evaluation metrics were:

Mean Squared Error (MSE): 3.61
R-squared (R^2): 0.52

This indicates the model explains approximately 52% of the variance in sentiment score outcomes, showing moderate predictive power.

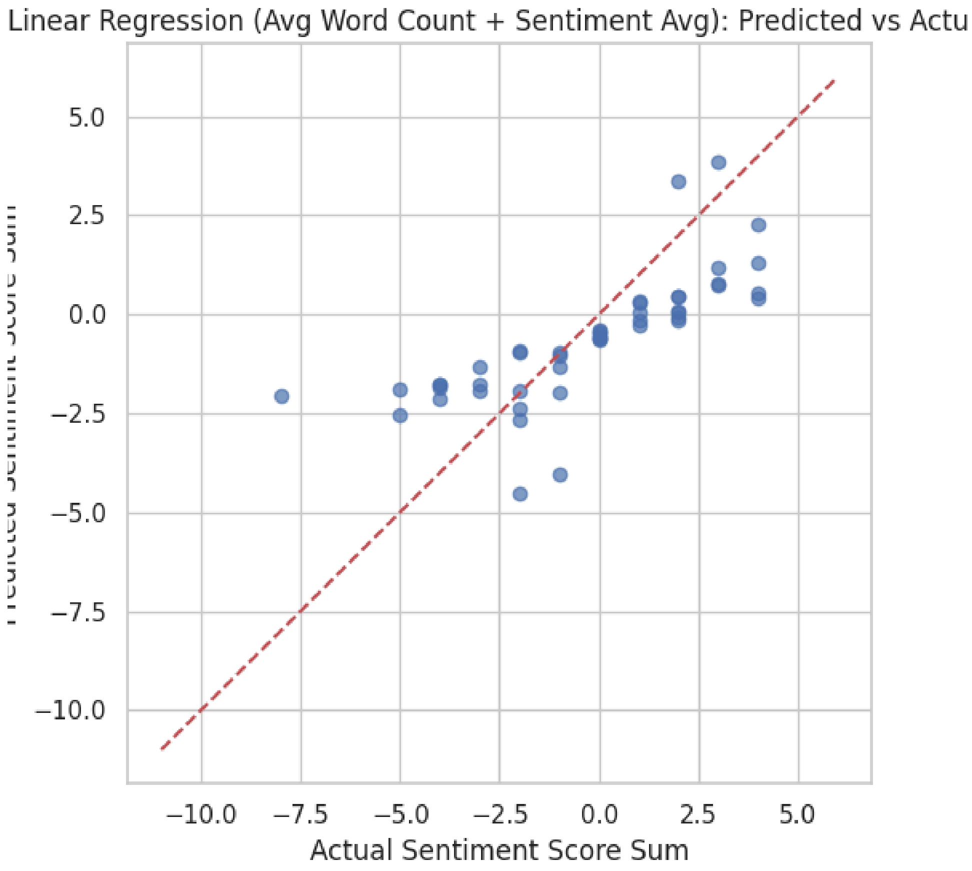

Figure 8: Predicted Vs Actual Sentiment Score Sum

Conclusion

This project successfully applied sentiment analysis and statistical modeling to assess employee sentiment trends, identify top performers and at-risk individuals, and explore the potential for predictive modeling. Key takeaways include the dominance of negative sentiment across communications, the importance of tracking monthly sentiment fluctuations, and the potential for using sentiment trends to inform proactive HR strategies. Future improvements could include more advanced machine learning models and incorporating additional contextual features.