Final Report: Employee Communication Sentiment Analysis

# 1. Introduction

Employee communication reflects both individual attitudes and organizational culture. Analyzing these communications provides an opportunity to uncover insights about engagement, satisfaction, and retention risks.

In this project, I applied Natural Language Processing (NLP), statistical analysis, and predictive modeling to employee email data to:

* Perform sentiment analysis on individual messages.
* Score and rank employees based on their communication tone.
* Detect employees at flight risk through patterns of negativity.
* Build and evaluate a predictive model to estimate monthly sentiment scores.

The insights derived from this project can support HR in employee engagement initiatives, early warning systems for attrition, and organizational decision-making.

# 2. Methodology

2.1 Data Preprocessing

* Loaded raw email dataset (`test.csv`).
* Converted `date` column to datetime format.
* Cleaned missing values and standardized text fields.
* Extracted `Month` field from dates for temporal analysis.

2.2 Sentiment Analysis

* Used VADER Sentiment Analyzer from NLTK.
* Extracted compound sentiment score (range: -1 to +1).
* Classified messages into:
* Positive (≥ 0.05)
* Negative (≤ -0.05)
* Neutral (otherwise)

2.3 Feature Engineering

Created features at both message and employee level:

* Message-level features: word count, message length.
* Aggregated features: monthly message counts, monthly average sentiment, cumulative sentiment score.

# 3. Exploratory Data Analysis (EDA)

EDA provided initial insights into employee communication:

* + Sentiment distribution: Most messages were neutral, followed by positive, with a smaller fraction being negative.
  + Message volume: A few employees contributed the majority of communications, showing possible central roles.
  + Monthly trends: Average sentiment scores fluctuated, with certain months showing spikes in negativity.
  + Top employees by volume: High-volume communicators were not always the most positive, emphasizing the importance of sentiment over quantity.

Key Findings from EDA:

1. Communication is skewed toward neutral/positive, but negativity is concentrated in a small group of employees.

2. Message volume is unevenly distributed — a handful of employees dominate communication.

3. Temporal patterns exist — periods of increased negativity likely correspond to stressful organizational events.

4. Negative communicators overlap significantly with identified flight-risk employees, reinforcing the risk criteria.

A graph of a distribution of positive and negative

AI-generated content may be incorrect.

Figure 1: Sentiment Distribution

A graph of different colored lines

AI-generated content may be incorrect.

Figure 2: Monthly Sentiment Trends

A graph with blue and white stripes

AI-generated content may be incorrect.

Figure 3: Top 10 Employees by Message Volume

# 4. Employee Scoring and Ranking

4.1 Scoring Process

To measure employee sentiment, I assigned numeric scores to each message:

* + Positive = +1
  + Neutral = 0
  + Negative = -1

Then, I aggregated scores per employee per month. This showed how employee tone evolved over time.

4.2 Ranking Process

* + Cumulative sentiment scores were computed for each employee.
  + Employees were ranked from most positive to most negative.

This produced two leaderboards:

* + Top positive employees → strong contributors to positive culture.
  + Top negative employees → potential concerns for HR.

A graph with green and white stripes

AI-generated content may be incorrect.

Figure 4: Top 3 Positive Employees

A graph with red and white stripes

AI-generated content may be incorrect.

Figure 5: Top 5 Negative Employees

# 5. Flight Risk Analysis

5.1 Identification Criteria

I flagged employees as “at risk” if they:

* + Sent ≥4 negative messages within a 30-day rolling window.

This rule captures sustained negativity, avoiding false alarms from isolated negative messages.

5.2 Outcomes

* + Identified a subset of employees with persistent negativity.
  + Many matched with “bottom-ranked” employees, strengthening confidence in the method.
  + These employees represent high-priority cases for HR follow-up.

A graph with orange bars

AI-generated content may be incorrect.

Figure 6: Flight Risk Employees Bar Chart

# 6. Predictive Modeling

6.1 Overview

I built a Linear Regression model as a baseline to predict employee monthly sentiment scores.

Features used:

* + Average message length.
  + Average word count.
  + Monthly message count.
  + Average sentiment compound score.

Target variable: Monthly employee sentiment score.

6.2 Evaluation

* + Metrics: Mean Squared Error (MSE), R² score.

Findings:

* + Model captured some variance but showed modest predictive power (low R²).
  + Predictions aligned with actual values for employees with stable patterns but struggled with volatile cases.

6.3 Performance Visuals

* + Actual vs Predicted plot → showed general alignment but with noise.
  + Residual distribution → centered near zero but spread indicated underfitting.

A graph with blue dots and red line

AI-generated content may be incorrect.

Figure 7: Actual vs Predicted Sentiment Scores

A graph with purple lines and a purple line

AI-generated content may be incorrect.

Figure 8: Residuals Distribution

6.4 Limitations & Future Work

* + Linear Regression is too simple for complex communication data.
  + Improvements:
  + Use non-linear models (Random Forest, Gradient Boosting).
  + Add richer NLP features (TF-IDF, embeddings).
  + Incorporate context (time-of-day, communication network patterns).

# 7. Key Findings

1. Most communication is neutral/positive, but persistent negativity is localized to specific employees.

2. Flight-risk employees were successfully flagged by combining sentiment scoring with frequency thresholds.

3. Top negative employees overlap strongly with those flagged as at risk.

4. Predictive modeling shows promise, but requires more advanced techniques for accuracy.

# 8. Conclusion & Recommendations

This project demonstrates how NLP, statistical scoring, and modeling can be applied to employee communications to uncover insights about organizational health.

Recommendations:

1. Continuous Monitoring: Implement dashboards to track sentiment and employee rankings monthly.

2. Early Interventions: Use flight-risk alerts to trigger HR outreach before attrition occurs.

3. Enhanced Models: Adopt advanced NLP methods (TF-IDF, transformers like BERT) for better prediction.

4. Holistic Analysis: Combine communication sentiment with performance and survey data for well-rounded HR insights.

By following this framework, organizations can proactively manage engagement, improve culture, and reduce attrition risk.