# **Final Report: Employee Sentiment Analysis**

**Date:** 6 June 2025  
**Author:** Faraaz Shaikh

## 1. Project Overview and Objective

This report details the analysis of an unlabeled dataset of employee email messages (test.csv). The primary objective was to evaluate employee sentiment and engagement by performing sentiment labeling, exploratory data analysis, employee scoring and ranking, flight risk identification, and predictive modeling. The analysis was conducted using Python with the pandas, Matplotlib, scikit-learn, and VADER libraries.

## 2. Task 1: Sentiment Labeling Methodology

To automatically label each message, a proven Natural Language Processing (NLP) technique was employed.

**Approach:**

* **Tool:** The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool was chosen. VADER is highly effective for analyzing text from social media and other short, informal contexts like emails, and it does not require heavy computational resources or API dependencies.
* **Process:**
  1. The Subject and body fields of each email were combined into a single text string for a holistic analysis.
  2. VADER's SentimentIntensityAnalyzer was applied to this combined text to generate a set of scores, including a compound score ranging from -1 (most negative) to +1 (most positive).
  3. Each message was classified into one of three categories based on its compound score:
     + **Positive:** Compound score >= 0.05
     + **Negative:** Compound score <= -0.05
     + **Neutral:** Compound score between -0.05 and 0.05

This process successfully labeled all 2,191 messages in the dataset.

## 3. Task 2: Exploratory Data Analysis (EDA)

An exploratory data analysis was conducted to understand the dataset's structure, distributions, and trends.

### 3.1. Dataset Structure

* **Total Records:** 2,191
* **Total Columns:** 6 (after initial processing)
* **Missing Values:** The dataset was clean, with no missing values in the core columns (Subject, body, from) after initial handling.

### 3.2. Sentiment Distribution

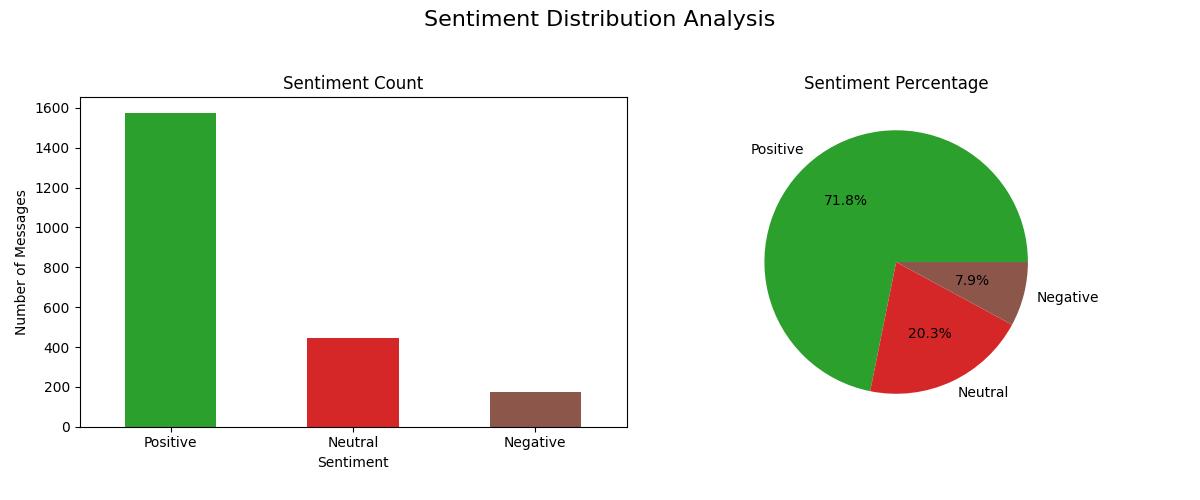
The analysis revealed that the sentiment across all messages was predominantly positive.

* **Positive:** 1,574 messages (71.8%)
* **Neutral:** 444 messages (20.3%)
* **Negative:** 173 messages (7.9%)

### 3.3. Visualizations

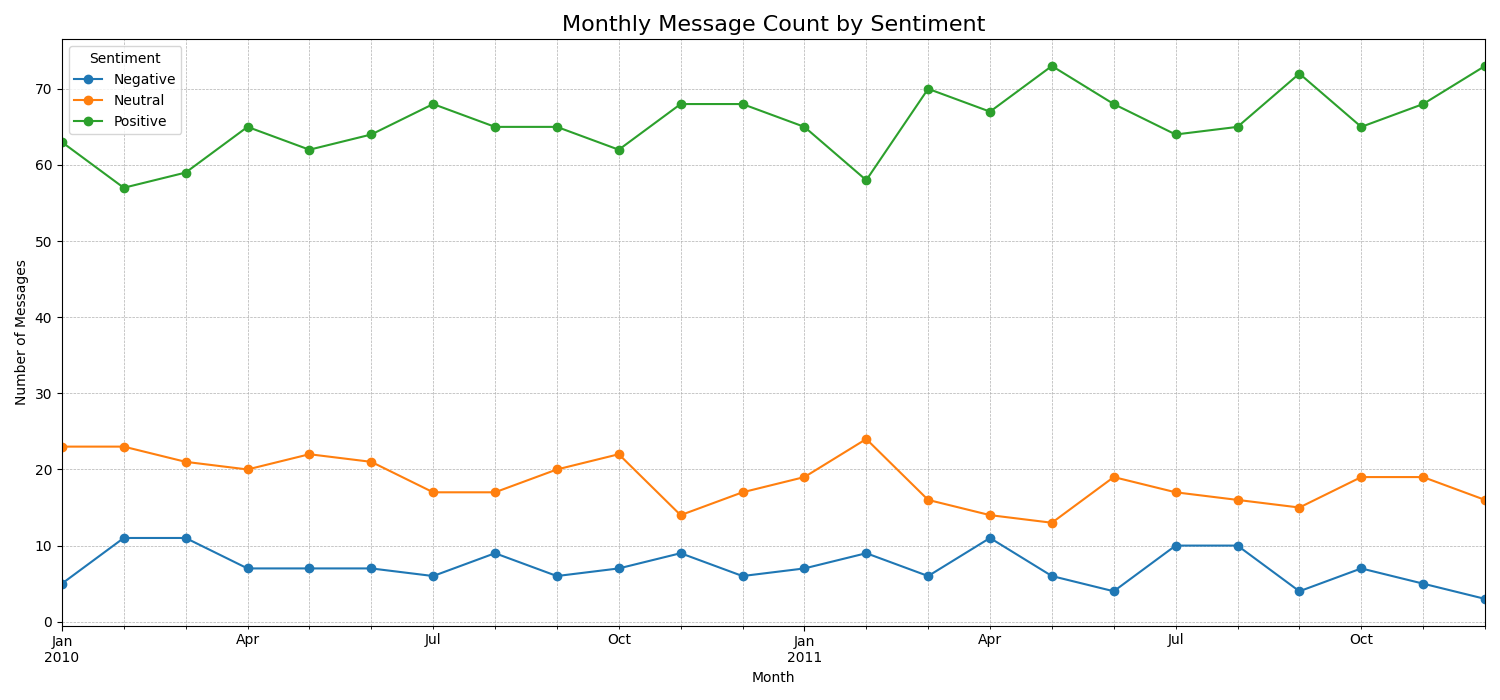
**Figure 1: Sentiment Distribution (Count and Percentage)**

The following charts illustrate the significant skew towards positive sentiment in the communications.

**

**Figure 2: Sentiment Trends Over Time**

The monthly volume of messages for each sentiment category was plotted to observe trends. The volume of messages remains relatively consistent over the 24-month period, with no dramatic shifts in the overall sentiment ratio.

**

## 4. Task 3 & 4: Employee Scoring and Ranking

To quantify and track employee sentiment over time, a monthly scoring and ranking system was implemented.

### 4.1. Scoring Methodology

A numerical score was assigned to each message based on its sentiment label:

* **Positive Message:** +1 point
* **Negative Message:** –1 point
* **Neutral Message:** 0 points (no effect)

These scores were then aggregated for each employee on a monthly basis. The score calculation resets at the beginning of each new month. Across all employee-month combinations, the following statistics were observed:

* **Average Monthly Score:** 5.84
* **Highest Monthly Score:** 20
* **Lowest Monthly Score:** -1

### 4.2. Employee Ranking

Based on the monthly scores, two ranked lists were generated for each of the 24 months in the dataset to highlight employees at the extremes of the sentiment spectrum. The lists were sorted first by score and then alphabetically by name.

Below is a sample ranking for the first month of the dataset. Full monthly rankings are available in the supplementary sentiment\_analysis.ipynb file.

**Sample Ranking: Month 2010-01**

| **TOP 3 POSITIVE EMPLOYEES** | **Score** |
| --- | --- |
| 1. [kayne.coulter@enron.com](mailto:kayne.coulter@enron.com) | +14 |
| 2. [eric.bass@enron.com](mailto:eric.bass@enron.com) | +9 |
| 3. [lydia.delgado@enron.com](mailto:lydia.delgado@enron.com) | +9 |

| **TOP 3 NEGATIVE EMPLOYEES** | **Score** |
| --- | --- |
| 1. [bobette.riner@ipgdirect.com](mailto:bobette.riner@ipgdirect.com) | +1 |
| 2. [rhonda.denton@enron.com](mailto:rhonda.denton@enron.com) | +1 |
| 3. [sally.beck@enron.com](mailto:sally.beck@enron.com) | +2 |

## 5. Task 5: Flight Risk Identification

A critical objective of this analysis was to identify employees who may be at risk of leaving the company.

### 5.1. Flight Risk Criteria

An employee was flagged as a potential flight risk if they sent **four or more negative emails within a rolling 30-day period**. This period is independent of calendar months.

### 5.2. Identified Flight Risk Employees

The analysis identified **7 employees** who met this criterion.

* bobette.riner@ipgdirect.com
* eric.bass@enron.com
* john.arnold@enron.com
* johnny.palmer@enron.com
* kayne.coulter@enron.com
* patti.thompson@enron.com
* sally.beck@enron.com

### 5.3. Verification Example

To validate the flagging logic, the negative email history for bobette.riner@ipgdirect.com was examined. A triggering window was found spanning just 23 days:

* Email 1: 2011-03-26
* Email 2: 2011-04-01
* Email 3: 2011-04-04
* Email 4: 2011-04-17

## 6. Task 6: Predictive Modeling

A linear regression model was developed to determine if basic message characteristics could predict sentiment scores.

### 6.1. Model Objective and Features

The model aimed to predict the VADER compound\_score using the following features:

* message\_length: Total character count of the message.
* word\_count: Total word count of the message.
* monthly\_frequency: The employee's total number of messages in that month.
* average\_monthly\_message\_length: The employee's average message length for that month.

### 6.2. Model Performance

The data was split into training (80%) and testing (20%) sets. The model's performance on the test set was modest:

* **R-squared (R²):** 0.1514
* **Root Mean Squared Error (RMSE):** 0.3944
* **Mean Absolute Error (MAE):** 0.3206

The **R² value of 0.1514** indicates that the selected features explain only 15.1% of the variance in sentiment scores. This suggests a weak linear relationship and that these features alone are not strong predictors of sentiment.

### 6.3. Model Interpretation

The model's coefficients provide insight into the relationship between each feature and the sentiment score:

| **Feature** | **Coefficient** |
| --- | --- |
| message\_length | 0.0002 |
| word\_count | 0.0033 |
| monthly\_frequency | -0.0013 |
| average\_monthly\_message\_length | -0.0003 |

* word\_count had the largest positive impact, suggesting that slightly longer messages (in words) are weakly associated with higher sentiment scores.
* The other coefficients are very close to zero, indicating a negligible linear relationship. The negative coefficients for frequency and average length are too small to be practically significant without further statistical testing.

## 7. Conclusions and Recommendations

### 7.1. Key Conclusions

1. **Overall Sentiment is Positive:** The vast majority of employee communications are positive or neutral, which is a healthy sign.
2. **Flight Risks are Identifiable:** Despite the overall positive trend, a specific, concentrated pattern of negativity allowed for the identification of 7 potential flight-risk employees.
3. **Simple Features are Poor Predictors:** A linear model using basic message metadata (length, word count) is insufficient for accurately predicting sentiment, confirming that sentiment is a complex linguistic phenomenon.

### 7.2. Actionable Recommendations

1. **Engage with Flagged Employees:** Management or HR should prioritize outreach to the 7 employees identified as flight risks to understand their concerns and provide support.
2. **Monitor Negative Trends:** The monthly rankings should be used as a tool to monitor employees who consistently appear in the "Top Negative" list, as they may be at future risk.
3. **Enhance Future Predictive Models:** For more accurate sentiment prediction, future work should focus on more sophisticated NLP features (e.g., TF-IDF, word embeddings like BERT) and explore non-linear machine learning models (e.g., Gradient Boosting).