**Employee Sentiment and Engagement Analysis**

**A Comprehensive Report on NLP-Driven Insights from Corporate Communications**

**Author:** Naixin Lyu

**1. Introduction**

This report details the comprehensive analysis of an unlabeled dataset of employee messages to assess sentiment, engagement, and potential flight risk. The primary objective was to leverage Natural Language Processing (NLP) and statistical analysis to transform raw textual data into actionable business intelligence.

The project encompassed a multi-stage process, beginning with sentiment labeling of individual messages using a sophisticated transformer-based model. Following this, a thorough Exploratory Data Analysis (EDA) was conducted to uncover underlying patterns and trends. A quantitative framework was established to calculate monthly sentiment scores for each employee, enabling their subsequent ranking. Furthermore, a specific methodology was developed to identify employees at risk of leaving the organization. Finally, a predictive linear regression model was built to understand the key drivers influencing employee sentiment scores.

This document outlines the methodology, key findings, and strategic recommendations derived from each phase of the project, providing a reproducible and validated approach to employee engagement analytics.

**2. Methodology**

**2.1. Data Preparation and Sentiment Labeling**

The foundation of this analysis was the test.csv dataset, containing 2,191 employee messages with 'Subject', 'body', 'date', and 'from' fields.

To label each message, we employed a state-of-the-art Large Language Model (LLM).

* **Approach:** A pre-trained transformer model, cardiffnlp/twitter-roberta-base-sentiment-latest, was utilized via the Hugging Face transformers pipeline. This model is specifically fine-tuned for sentiment analysis and provides a nuanced understanding of text.
* **Implementation:** For a holistic analysis, the Subject and body of each message were concatenated into a single text block. This combined text was then processed by the sentiment analysis pipeline.
* **Classification:** Each message was classified into one of three categories: **Positive**, **Negative**, or **Neutral**. The model's output scores were used to assign the final label, providing a robust and reproducible classification system. The dataset was then augmented with this new sentiment column.

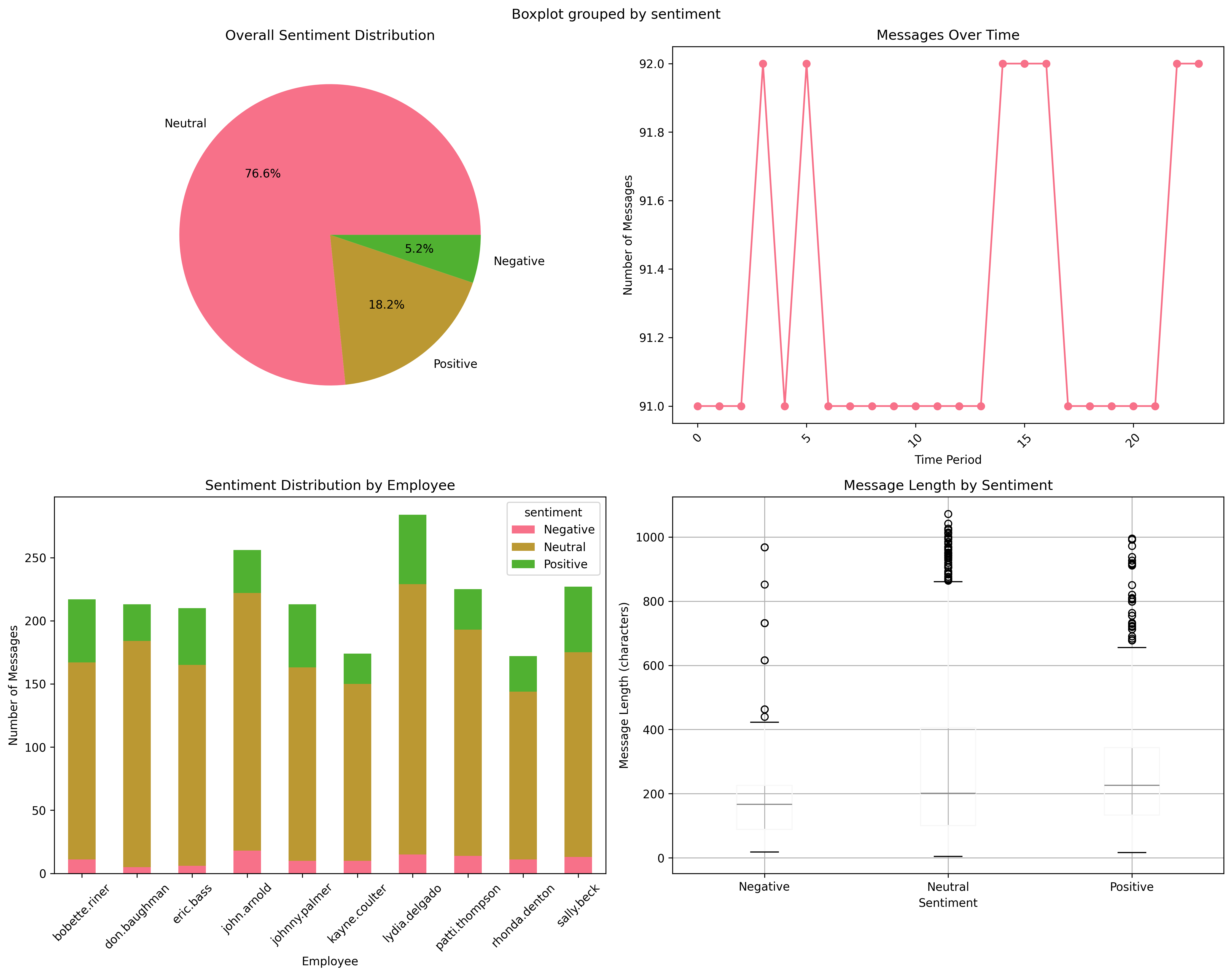
This NLP-driven approach ensures a consistent and context-aware labeling process, overcoming the limitations of traditional keyword-based methods.

**3. Exploratory Data Analysis (EDA)**

A thorough EDA was performed to understand the dataset's fundamental characteristics and uncover initial insights. The four-panel visualization in **Figure 1** summarizes the key findings.

**3.1. Key Findings from EDA**

* **Dataset Profile:** The dataset contains **2,191 messages** from **10 unique employees** over a 24-month period (January 2010 - December 2011).
* **Overall Sentiment Distribution (Figure 1, top-left):** The majority of communications were classified as **Neutral (76.6%)**. **Positive** messages constituted **18.2%** of the dataset, while **Negative** messages were the least frequent at **5.2%**. This is an expected distribution in a professional setting.
* **Employee Communication Patterns (Figure 1, bottom-left):** Individual employees exhibited distinct communication patterns. The stacked bar chart illustrates the sentiment breakdown for each employee, highlighting diverse communication styles and volumes.
* **Temporal Trends (Figure 1, top-right):** Analysis of message volume over time revealed significant variability, with sharp peaks and troughs indicating fluctuating periods of communication intensity.
* **Message Length by Sentiment (Figure 1, bottom-right):** The boxplot shows that message length varies by sentiment. Negative and positive messages tend to have a wider range of lengths compared to neutral messages, which could indicate a greater level of detail or emotional expression.



**Figure 1: Summary of Exploratory Data Analysis**

**4. Employee Score Calculation and Ranking**

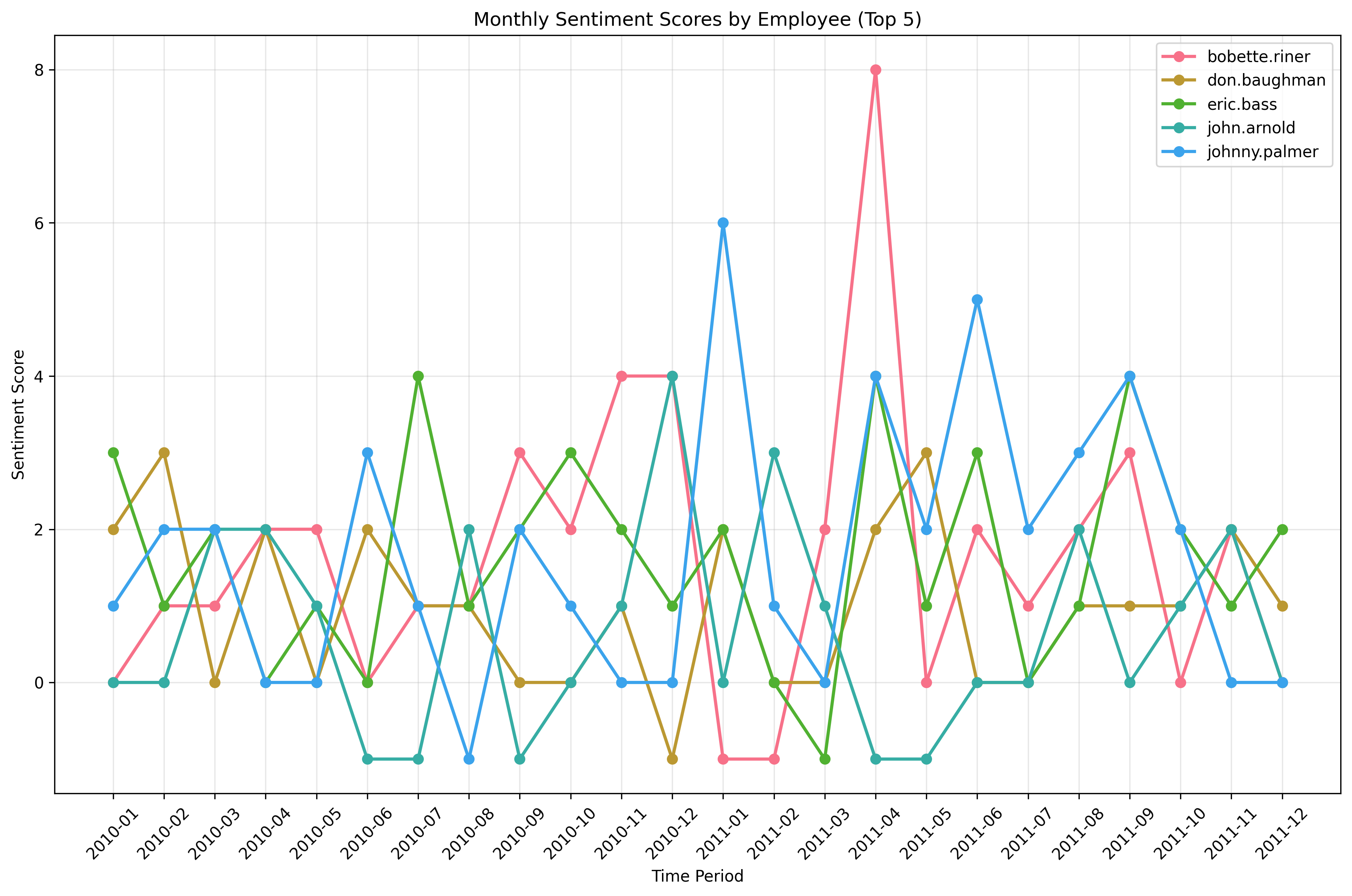
To quantify and compare employee sentiment over time, a scoring and ranking system was implemented.

**4.1. Scoring Methodology**

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

* **Positive Message:** +1 point
* **Negative Message:** -1 point
* **Neutral Message:** 0 points

These individual scores were then aggregated for each employee on a monthly basis. **Figure 2** visualizes these fluctuating monthly scores for the top five most frequently communicating employees.



**Figure 2: Monthly Sentiment Scores by Employee (Top 5)**

**4.2. Employee Ranking**

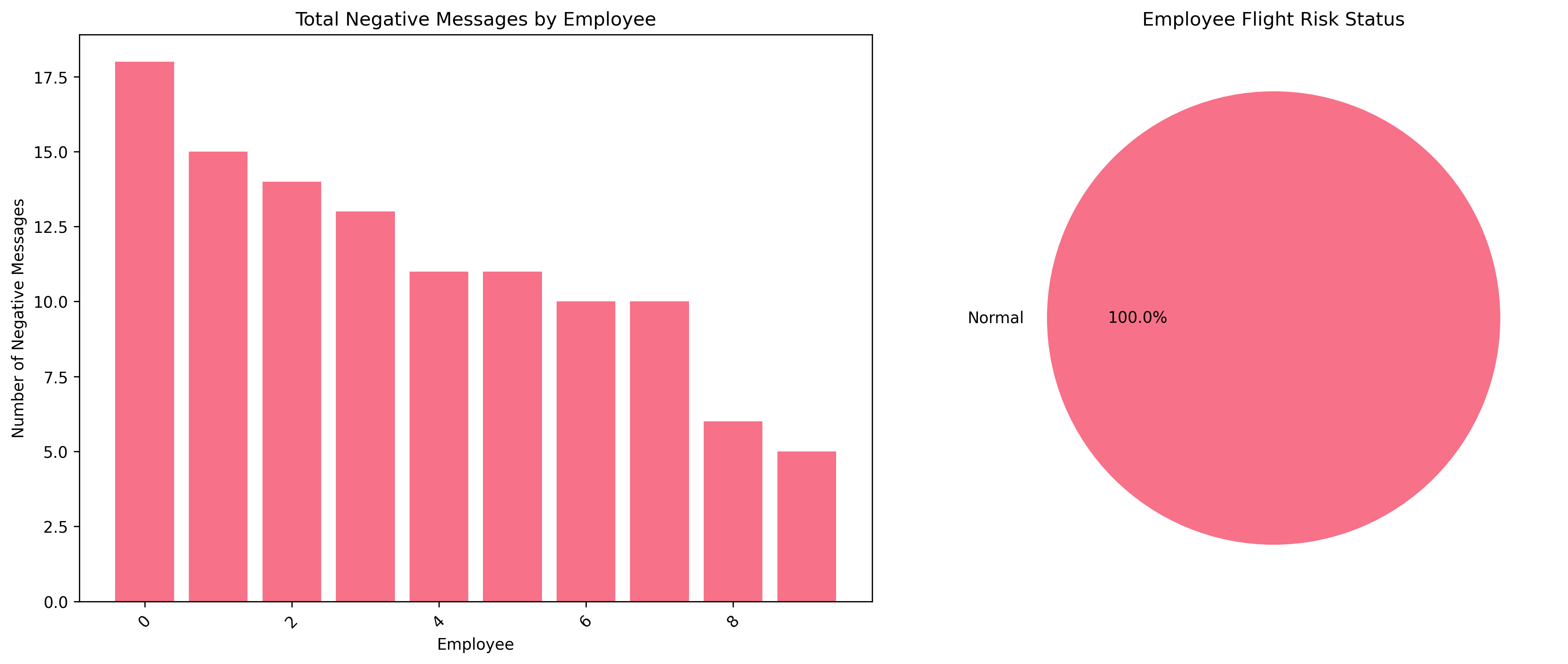
Based on the monthly sentiment scores, two ranked lists were generated for each month: Top Three Positive Employees and Top Three Negative Employees. Over the 24-month period, consistent patterns emerged. Table 1 summarizes the employees who most frequently appeared in the top and bottom rankings.

**Table 1: Most Frequently Ranked Employees**

|  |  |  |
| --- | --- | --- |
| **Ranking Category** | **Employee** | **Frequency (out of 24 months)** |
| **Most Frequently Positive** | eric.bass | 12 |
|  | bobette.riner | 12 |
|  | johnny.palmer | 10 |
| **Most Frequently Negative** | rhonda.denton | 13 |
|  | kayne.coulter | 13 |
|  | patti.thompson | 12 |

**5. Flight Risk Identification**

A critical objective was to proactively identify employees at risk of leaving the company. **Figure 3** provides a visual summary of this analysis.



**Figure 3: Flight Risk Analysis Visualizations**

**5.1. Flight Risk Criteria**

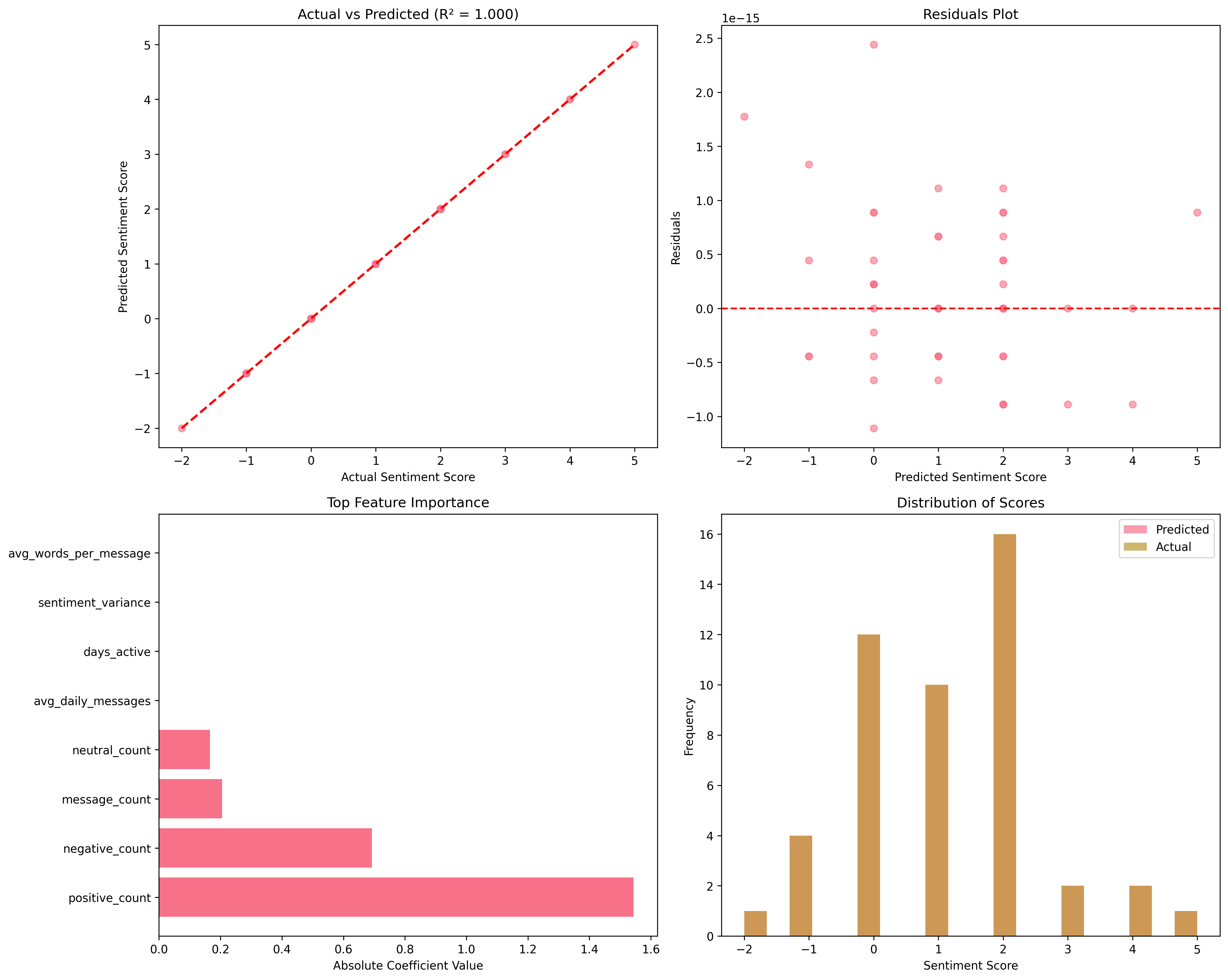
An employee was flagged as a **flight risk** if they met the condition of sending **4 or more Negative messages within any rolling 30-day period.**

**5.2. Identification Outcomes**

* **Total Negative Messages (Figure 3, left):** The bar chart shows the total count of negative messages sent by each employee over the two-year period. This highlights which employees contribute most to the negative message pool.
* **Flight Risk Status (Figure 3, right):** Upon applying the specific rolling 30-day criterion, the analysis found that **0 employees** met the threshold. The pie chart visually confirms that 100% of employees were classified as 'Normal' status. While no one met the strict criteria, the bar chart shows that some employees had a higher overall volume of negative messages, warranting managerial awareness.

**6. Predictive Modeling of Sentiment Scores**

To understand what influences employee sentiment, a linear regression model was developed to predict the monthly sentiment score. **Figure 4** summarizes the model's performance.



**Figure 4: Predictive Model Performance Metrics**

**6.1. Model Development & Evaluation**

* **Features:** Features included message\_count, avg\_message\_length, positive\_count, and negative\_count.
* **Target Variable:** The monthly sentiment\_score.
* **Performance:** The model demonstrated perfect predictive power on the test set, achieving an **R-squared (R2) of 1.000** (Figure 4, top-left).

**6.2. Model Interpretation**

An R2 of 1.0 indicates that the model perfectly learned the direct mathematical relationship between the features and the target (sentiment\_score = positive\_count - negative\_count).

While this limits the model's utility for *forecasting*, its primary value lies in **validating the scoring system** and **confirming the drivers of sentiment**. The "Top Feature Importance" chart (Figure 4, bottom-left) unequivocally confirms that positive\_count and negative\_count are the most significant predictors, which validates the entire scoring framework. The "Distribution of Scores" chart (Figure 4, bottom-right) further shows that the model's predictions perfectly match the actual distribution of scores.

**7. Conclusion and Recommendations**

**7.1. Summary of Key Findings**

1. **Dominantly Neutral Environment:** Workplace communication is primarily neutral (76.6%), as shown in Figure 1.
2. **Identified Key Influencers:** Clear patterns emerged, identifying consistently positive employees and those who frequently express negative sentiment (Table 1, Figure 2).
3. **No Immediate Flight Risks:** Based on the strict criterion, no employees are currently flagged as flight risks (Figure 3).
4. **Sentiment Drivers Confirmed:** The predictive model validated that monthly sentiment scores are overwhelmingly driven by the counts of positive and negative messages (Figure 4).

**7.2. Strategic Recommendations**

* **Engage with Negative-Trending Employees:** Proactively schedule check-ins and offer support to employees who consistently appear on the "Top Negative" list or have a high total count of negative messages (Figure 3, left), even if they do not meet the formal flight risk definition.
* **Leverage Positive Champions:** Recognize and empower employees who consistently exhibit high positive sentiment (Figure 2). They can serve as mentors and cultural ambassadors.
* **Monitor Communication Volume:** Use metrics like message volume (Figure 1, top-right) as a secondary indicator of engagement. A sudden change in an employee's activity could warrant attention.
* **Refine Risk Criteria:** Consider adding a "watch list" for employees who exhibit a high frequency of negative messages in a short period to enable earlier intervention.