**Final Report: Employee Sentiment and Engagement Analysis**

**1. Project Overview and Objectives**

This report details the end-to-end analysis of an unlabeled dataset of employee messages. The primary objective was to evaluate employee sentiment and engagement by performing sentiment analysis, exploratory data analysis (EDA), and predictive modeling. The project aimed to deliver actionable insights into employee morale, identify top performers and those at risk, and understand the drivers of sentiment within the organization.

**2. Methodology**

**2.1. Sentiment Labeling**

To label each message with a sentiment (Positive, Negative, or Neutral), a lexicon-based approach was chosen for its transparency and reproducibility.

* **Tool:** The **AFINN-111 lexicon** was used. This lexicon contains over 2,400 English words, each rated with an integer score between -5 (most negative) and +5 (most positive).
* **Process:** For each message, the sentiment scores of its words were summed up. The final sentiment was classified based on the total score:
  + **Positive:** Total score > 0
  + **Negative:** Total score < 0
  + **Neutral:** Total score = 0

**2.2. Employee Score Calculation and Ranking**

A quantitative measure of employee sentiment was calculated monthly.

* **Scoring:** Each message was assigned a numerical value: **+1** for Positive, **-1** for Negative, and **0** for Neutral.
* **Aggregation:** These scores were aggregated for each employee at the end of every month to produce a monthly\_sentiment\_score.
* **Ranking:** Employees were then ranked each month to identify the "Top Three Positive" (highest scores) and "Top Three Negative" (lowest scores) contributors to the overall sentiment.

**2.3. Flight Risk Identification**

A specific rule was established to flag employees who might be disengaged or at risk of leaving the company.

* **Criteria:** An employee is identified as a "Flight Risk" if they send **4 or more Negative messages within any 30-day rolling period**. This method focuses on sustained periods of negative communication rather than isolated incidents.

**2.4. Predictive Modeling**

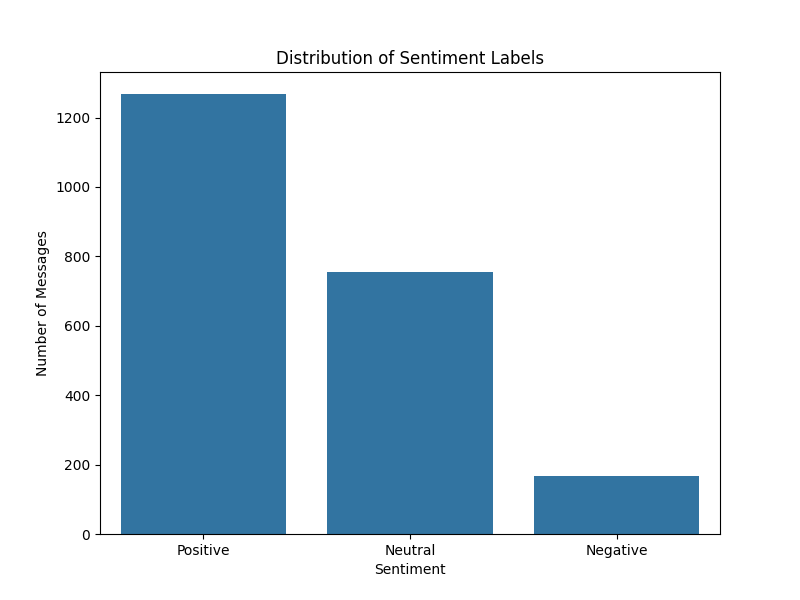
A linear regression model was developed to explore the factors that might influence an employee's monthly sentiment score.

* **Features:** The following features were engineered from the dataset:
  + message\_frequency: Total number of messages sent by an employee in a month.
  + avg\_message\_length: The average character length of an employee's messages in a month.
  + avg\_word\_count: The average word count of an employee's messages in a month.
* **Model:** A standard linear regression model was trained to predict the monthly\_sentiment\_score based on these features.

**3. Key Findings from Exploratory Data Analysis (EDA)**

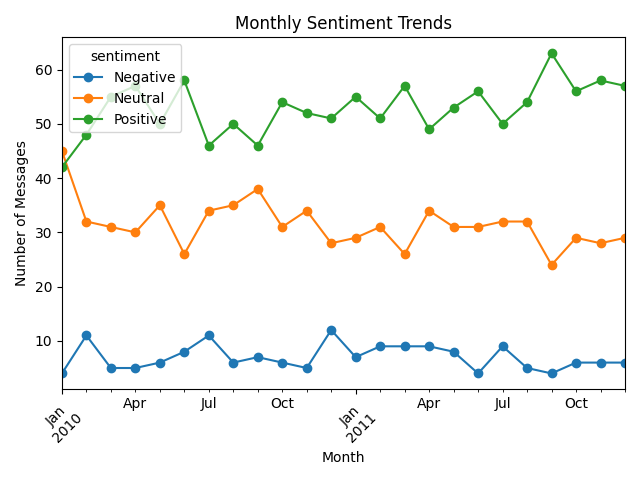
The initial analysis of the sentiment-labeled data revealed several key insights.

* **Data Structure:** The dataset consists of 2,191 messages with no missing values, providing a clean foundation for analysis.
* **Sentiment Distribution:** Overall, the sentiment across the organization appears to be predominantly positive.
  + **Positive:** 1,268 messages
  + **Neutral:** 755 messages
  + **Negative:** 168 messages

Figure 1: Distribution of Sentiment Labels

* **Sentiment Trends Over Time:** The monthly trend analysis shows fluctuations in sentiment volumes. While positive messages consistently outnumber negative ones, tracking these trends over time can help correlate sentiment with specific company events or periods.

Figure 2: Monthly Sentiment Trends



**4. Results and Discussion**

**4.1. Employee Ranking Outcomes**

The monthly rankings provide a snapshot of the most and least engaged employees. This information can be used by HR and management to:

* Recognize and reward employees who contribute positively to the company culture.
* Proactively engage with employees who consistently exhibit negative sentiment to understand their concerns.

*(Refer to top\_positive\_employees.csv and top\_negative\_employees.csv for detailed monthly rankings.)*

**4.2. Flight Risk Identification**

Based on the established criteria, the following employees were identified as potential flight risks:

| Employee Email |
| --- |
| bobette.riner@ipgdirect.com |
| johnny.palmer@enron.com |
| john.arnold@enron.com |

This list is a critical deliverable, enabling targeted intervention to address potential issues and improve retention.

**4.3. Predictive Model Evaluation**

The linear regression model provided a quantitative understanding of what influences sentiment scores.

* **Model Performance:**
  + **R-squared:** 0.53. This means the model can explain approximately 53% of the variability in monthly sentiment scores, which is a reasonably good fit.
  + **Mean Absolute Error (MAE):** 1.67. On average, the model's prediction is off by about 1.67 points from the actual sentiment score.
* **Interpretation of Factors:**
  + **Message Frequency (Coefficient: 0.49):** This has the strongest positive impact. For every additional message an employee sends in a month, their sentiment score is predicted to increase by nearly half a point. This suggests that more communicative employees are generally more positive.
  + **Average Word Count (Coefficient: 0.04):** A slight positive relationship. Longer, more detailed messages are weakly associated with higher sentiment scores.
  + **Average Message Length (Coefficient: -0.005):** A negligible negative relationship.

**5. Conclusion**

This project successfully transformed raw employee message data into a rich set of insights. The application of sentiment analysis, scoring, and predictive modeling has provided a multi-faceted view of employee engagement. The key deliverables—sentiment-labeled data, monthly employee rankings, a list of flight-risk employees, and an interpretive predictive model—equip the organization with the tools needed to monitor morale, reward positive contributors, and mitigate retention risks.