**Final Project Report:**

**1. Introduction**

This report details the full analysis of the employee email dataset (test.csv). The primary objective was to evaluate employee sentiment and engagement by labeling messages, analyzing trends, scoring and ranking employees, identifying potential flight risks, and developing a predictive model. The following sections outline our methodology, key findings, and final recommendations.

**2. Approach and Methodology**

**2.1. Sentiment Labeling**

The first step was to assign a sentiment label (Positive, Negative, or Neutral) to each email. A custom lexicon-based model was developed for this task. A predefined list of positive (e.g., "thanks", "success", "approved") and negative (e.g., "problem", "issue", "delay") keywords was used to score each email. An email was labeled 'Positive' if it had a net positive score, 'Negative' if it had a net negative score, and 'Neutral' otherwise. This approach was chosen for its speed, transparency, and reproducibility.

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

After labeling, an EDA was conducted to understand the dataset's underlying trends. This involved:

* **Data Structure Examination:** Reviewing the dataset for size, completeness, and data types.
* **Sentiment Distribution:** Analyzing the proportion of positive, negative, and neutral emails.
* **Temporal Analysis:** Plotting email volume and sentiment trends over time to identify patterns.

**2.3. Employee Scoring**

To quantify sentiment on a monthly basis, a scoring system was implemented.

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

These scores were aggregated for each employee at the end of every month to produce a monthly\_sentiment\_score.

**2.4. Employee Ranking**

Using the monthly scores, two ranked lists were generated for each month:

* **Top Three Positive Employees:** Sorted by the highest monthly score, then alphabetically.
* **Top Three Negative Employees:** Sorted by the lowest (most negative) monthly score, then alphabetically.

**2.5. Flight Risk Identification**

A critical objective was to identify employees at a potential flight risk. According to the project definition, a flight risk is any employee who sends 4 or more negative emails in a rolling 30-day period. This was calculated by filtering for negative emails and applying a rolling 30-day count for each employee.

**2.6. Predictive Modeling**

Finally, a linear regression model was developed to analyze factors influencing monthly sentiment scores.

* **Features Selected:** The independent variables included message\_count (monthly message frequency), average\_word\_count (average message length), and negative\_message\_ratio.
* **Validation:** The model was trained on 80% of the data and evaluated on the remaining 20% using R-squared (R2) and Mean Squared Error (MSE) as performance metrics.

**3. Key Findings and Results**

**3.1. EDA Findings**

The EDA revealed a healthy communication environment. The dataset contained 2,191 emails, with the majority being Neutral (61.8%), followed by Positive (33.1%) and a small minority of Negative (5.1%).

*[Visualization: Bar chart showing the Overall Distribution of Email Sentiment]*

Email volume increased significantly in 2011, but the proportion of negative messages remained consistently low, suggesting that periods of high activity did not negatively impact overall morale.

*[Visualization: Line chart showing Monthly Email Volume by Sentiment]*

**3.2. Employee Ranking Outcomes**

The ranking process consistently identified key communicators. The table below shows a sample ranking for May 2011.

**Table 1: Sample Employee Rankings for May 2011**

|  |  |  |
| --- | --- | --- |
| **Rank** | **Top 3 Positive Employees** | **Score** |
| 1 | sally.beck@enron.com | 4 |
| 2 | lydia.delgado@enron.com | 2 |
| 3 | bobette.riner@ipgdirect.com | 1 |

|  |  |  |
| --- | --- | --- |
| **Rank** | **Top 3 Negative Employees** | **Score** |
| 1 | don.baughman@enron.com | -1 |
| 2 | eric.bass@enron.com | -1 |
| 3 | john.arnold@enron.com | 0 |

|  |  |  |
| --- | --- | --- |
| **Rank** | **Top 3 Negative Employees** | **Score** |
| 1 | don.baughman@enron.com | -1 |
| 2 | eric.bass@enron.com | -1 |
| 3 | john.arnold@enron.com | 0 |

**3.3. Flight Risk Identification Outcomes**

The analysis identified **two employees** who met the flight risk criteria at some point during the two-year period:

* eric.bass@enron.com
* rhonda.denton@enron.com

This finding is critical, as it highlights individuals who experienced a concentrated period of negative communication, which may indicate high stress or disengagement.

**3.4. Predictive Model Evaluation**

The linear regression model provided valuable insights into the drivers of sentiment.

**Table 2: Model Performance Metrics**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| R-squared (R2) Score | 0.40 |
| Mean Squared Error (MSE) | 1.83 |

An **R-squared value of 0.40** indicates that the model explains 40% of the variance in sentiment scores. The model's coefficients confirmed that the negative\_message\_ratio was the most significant factor influencing an employee's score.

**4. Conclusion and Recommendations**

This analysis provided a data-driven framework for understanding employee sentiment. While overall morale appears positive, the results show that specific individuals exhibit patterns of negative communication that warrant attention.

* **Recommendation 1:** We recommend that HR or management connect with the individuals identified in the "Top Negative" and "Flight Risk" reports to offer support and address any underlying issues.
* **Recommendation 2:** The scoring and flight risk models should be used as ongoing tools to proactively monitor employee well-being and identify potential issues before they escalate.