Employee Sentiment Analysis Report

1. Introduction & Objective

The purpose of this project is to analyze employee email/text messages to evaluate sentiment and engagement levels. Using natural language processing (NLP) and statistical analysis, we identify sentiment trends, calculate employee scores, rank employees, detect potential flight risks, and build a predictive model to understand the drivers of sentiment.

The dataset provided (test.csv) was unlabeled. The tasks included:

- Sentiment labeling (Positive, Negative, Neutral)
- Exploratory Data Analysis (EDA)
- Employee sentiment score calculation
- Employee ranking
- Flight risk identification
- Predictive modeling using regression

2. Methodology

Sentiment Labeling

- **Approach:** Used **NLTK VADER SentimentIntensityAnalyzer** for classifying each message.
- **Criteria:**
  - Compound score ≥ 0.05 → Positive
  - Compound score ≤ -0.05 → Negative
  - Otherwise → Neutral
- Output: Added a sentiment column with values {Positive, Negative, Neutral}.

Exploratory Data Analysis (EDA)

- Checked dataset structure, missing values, and distributions.
- Visualized sentiment distribution across all messages.

- Analyzed organizational monthly sentiment trend.
- Created summary tables and charts.

Employee Score Calculation

- Each message scored: Positive = +1, Negative = -1, Neutral = 0.
- Scores aggregated **monthly per employee**.
- Generated monthly_scores.csv containing employee × month × score.

Employee Ranking

- For each month:
    - **Top 3 Positive Employees:** Highest scores.
    - **Top 3 Negative Employees:** Lowest scores.
- Sorted by score, then alphabetically.
- Presented via tables and bar plots.

Flight Risk Identification

- Defined as: Employees with **≥ 4 negative messages in any rolling 30-day window**.
- Implemented rolling sum of negative messages grouped by employee.
- Extracted flagged employees as potential flight risks.

Predictive Modeling

- **Model:** Linear Regression using scikit-learn.
- **Features:**
    - Message count per month
    - Average message length
    - Positive, Negative, Neutral message counts
- **Target:** Monthly sentiment score.
- **Evaluation Metrics:** RMSE, R^2.
- **Interpretation:** Feature coefficients analyzed for significance.

3. Key Findings

EDA Insights

- Sentiment distribution: Majority of messages were Neutral, followed by Positive and then Negative.
- Organizational trend: Sentiment scores fluctuated month-to-month, reflecting employee morale variations.

Employee Scoring & Ranking

- Produced monthly ranking of top 3 positive and negative employees.
- Example (latest month):
    - **Top Positive Employees:** [Employee_A, Employee_B, Employee_C]
    - **Top Negative Employees:** [Employee_X, Employee_Y, Employee_Z]

Flight Risk Employees

- Employees flagged with ≥4 negative messages in 30 days.
- Example: [Employee_M, Employee_N]

Predictive Model Results

- **RMSE:** ~[value from run]
- **R^2:** ~[value from run]
- **Important Features:** Negative message count and positive message count were the strongest predictors of sentiment score.

4. Conclusion & Recommendations

- Employee sentiment varies significantly month to month, with identifiable top performers and struggling employees.
- Flight risk detection provides early signals for HR intervention.
- Predictive modeling confirms that communication tone (positive vs negative counts) is highly influential in determining sentiment scores.

Recommendations:

1. Monitor flagged flight risk employees proactively.
2. Encourage positive communication initiatives.
3. Use predictive insights to guide HR engagement strategies.