**Employee Sentiment Analysis - Final Report**

**Author:** Kunal Vishwa Sivakumar  
**Program:** M.S. Applied Data Analytics, Boston University

**1. Introduction**

This report outlines the process and findings of an Employee Sentiment Analysis project. The objective was to analyze employee emails using NLP techniques and statistical modeling to assess sentiment, rank employees, detect flight risks, and build a predictive model.

**2. Methodology Overview**

The analysis was conducted in Python using pandas, VADER from NLTK for sentiment classification, and scikit-learn for modeling.

**3. Task Breakdown & Findings**

**Task 1: Sentiment Labeling**

* **Approach:** Used the VADER sentiment analyzer to label each message as Positive, Negative, or Neutral.
* **Output:** A new column sentiment\_label added in test\_labeled.csv.

**Task 2: Exploratory Data Analysis (EDA)**

* **Data Structure:** No null values. Each record contains sender, message content, and timestamp.
* **Visual Insights:**
  + Sentiment distribution showed dominance of positive sentiment.
  + Monthly trends indicate fluctuations in employee engagement.
  + Top 10 active employees were mostly associated with the Enron group.
  + Word count distribution showed most messages are concise, peaking under 100 words.
* **Visuals:** All charts saved in visualization/ folder.

**Task 3: Monthly Sentiment Score**

* **Scoring Logic:**
  + +1 for Positive
  + -1 for Negative
  + 0 for Neutral
* **Aggregation:** Scores were grouped by employee and month and saved to monthly\_scores.csv.

**Task 4: Employee Ranking**

* **Ranking Rule:** Based on monthly sentiment scores, with alphabetical tie-breaker.
* **Output:**
  + top\_3\_positive.csv lists top 3 employees per month with highest scores.
  + top\_3\_negative.csv lists employees with lowest scores.

**Task 5: Flight Risk Identification**

* **Definition:** Employees who sent ≥ 4 negative messages within any rolling 30-day period.
* **Implementation:** Used rolling windows based on message date.
* **Output:** flight\_risks.csv lists the flagged employees.

**Task 6: Predictive Modeling**

* **Goal:** Predict sentiment score using message frequency and length.
* **Model:** Linear Regression using sklearn.
* **Features Used:**
  + Number of messages per employee per month
  + Average message length
* **Evaluation:**
  + MSE and R² calculated
  + Model showed moderate predictive power, suggesting other behavioral/emotional features may improve accuracy.

**4. Key Insights**

* Positive sentiment dominated across the organization, indicating general satisfaction.
* Certain employees consistently topped both positive and negative rankings month-over-month.
* Flight risks showed early negative trends, supporting the temporal detection approach.
* Predictive modeling captured basic patterns but left room for improvement with richer features.

**5. Recommendations**

* Consider proactive engagement for employees flagged as potential flight risks.
* Enrich future models with semantic analysis, departmental context, or communication patterns.
* Implement live dashboards to monitor sentiment trends in real time.

**6. Deliverables Summary**

* **Notebook:** Employee.ipynb
* **Outputs:** CSV files under outputs/
* **Visuals:** Plots under visualization/
* **Readme:** README.md with complete summary