Final Report - Employee Sentiment Analysis

Made by : Amey Ghatol ( [ameyghatol@gmail.com](mailto:ameyghatol@gmail.com) )

### **1. Approach and Methodology**

The primary objective of this project was to analyze employee messages to extract insights about sentiment and engagement. The methodology was structured into clear, reproducible steps:

* **Data Preparation**: The dataset consisted of emails with fields like Subject, body, date, and from. The Subject and body were concatenated to form a unified message field. The from field was renamed to sender to avoid conflicts with Python keywords. The date column was parsed and validated.
* **Sentiment Labeling**: We used the VADER sentiment analyzer from NLTK to label each message as Positive, Negative, or Neutral. This was based on the compound sentiment score: scores ≥ 0.05 were labeled Positive, ≤ -0.05 as Negative, and those in between as Neutral.

### **2. Key Findings from the EDA**

* The dataset included 2191 messages from 10 unique employees.
* Most frequent message sender: lydia.delgado@enron.com with 284 messages.
* The sentiment distribution showed a roughly balanced spread across Positive, Negative, and Neutral, with a slight skew toward Neutral.
* Time-series analysis showed fluctuating sentiment levels across months, with certain periods dominated by either positive or negative communication.
* Message length and frequency varied significantly between employees; longer messages tended to carry stronger sentiment polarity.

### **3. Employee Scoring and Ranking**

* Each message was scored as:  
  + +1 for Positive
  + -1 for Negative
  + 0 for Neutral
* These scores were aggregated per employee per calendar month to compute a **monthly sentiment score**.
* **Top Positive Employees**: Employees with the highest total scores in a month, indicating strong positive engagement.
* **Top Negative Employees**: Employees with the most negative scores, possibly indicating dissatisfaction or stress.
* Rankings were generated monthly and sorted first by score and then alphabetically to break ties.

### **4. Flight Risk Identification Criteria and Outcomes**

* A **flight risk** is defined as any employee who sends **four or more negative messages** within a **rolling 30-day window**.
* This threshold helps identify sustained negativity over a short period, a potential sign of disengagement.
* Using this logic, a list of flight-risk employees was extracted. These employees warrant further attention from HR for proactive engagement or support.

### **5. Predictive Modeling Overview and Evaluation**

* A linear regression model was trained to **predict monthly sentiment scores** using:  
  + Message count per employee
  + Average word count per message
  + Average character count per message
* The dataset was split into training and testing sets (80/20).
* Model performance:  
  + **Mean Squared Error (MSE)**: Indicates average squared difference between actual and predicted scores.
  + **R² Score**: Showed how much variance in sentiment score is explained by the model.
* While the model showed moderate predictive capability, it highlighted message frequency and length as meaningful features. Future iterations could incorporate more nuanced NLP features (e.g., topic modeling, emotional tone).

### **Conclusion**

This project provided a structured pipeline for analyzing employee sentiment using real-world textual data. The insights generated can be pivotal for HR and leadership to identify top contributors, support at-risk employees, and build a more engaged workforce.