Employee Sentiment Analysis – Final Report

# 1. Approach and Methodology

This project leverages Natural Language Processing and statistical modeling to analyze employee sentiment from email messages.  
The entire workflow includes the following steps:  
  
1. Data Cleaning and Preparation: The raw dataset was cleaned and renamed for clarity.  
2. Sentiment Labeling: We used the pre-trained DistilBERT model (`distilbert-base-uncased-finetuned-sst-2-english`) via Hugging Face's Transformers pipeline to label each message as either Positive or Negative.  
3. Exploratory Data Analysis (EDA): We performed statistical summaries and visualizations to explore sentiment distribution, trends over time, and engagement patterns.  
4. Scoring and Ranking: Each employee's monthly sentiment score was calculated based on their labeled messages. Top and bottom employees were identified based on scores.  
5. Flight Risk Analysis: Employees sending 4 or more negative messages within any 30-day window were flagged as flight risks.  
6. Predictive Modeling: A linear regression model was trained to estimate monthly sentiment scores based on message volume and characteristics like length and word count.

# 2. Key Findings from EDA

- The dataset contains 2,191 email messages.  
- There are no missing values in any column.  
- Most messages are classified as Negative, highlighting potential employee dissatisfaction.  
- Monthly sentiment trends showed fluctuations, with spikes in negativity during certain periods.  
- Message volume varies significantly across employees, suggesting varying engagement levels.

# 3. Employee Scoring and Ranking

Each email was scored based on its sentiment: +1 for Positive, -1 for Negative. Scores were aggregated monthly per employee.  
  
Ranking was performed by sorting these monthly scores:  
- Top 3 Positive Employees: Highest cumulative scores  
- Top 3 Negative Employees: Lowest cumulative scores  
  
Ties were broken alphabetically. Rankings were visualized and tabulated for easy interpretation.

# 4. Flight Risk Identification

Criteria:  
- Any employee sending 4 or more Negative emails within a rolling 30-day period was flagged as a flight risk.  
  
Implementation:  
- Data was resampled by day per employee.  
- A rolling window sum was used to count negative messages.  
  
Outcome:  
- Multiple employees were flagged as at risk based on this rule, indicating early warning signals for HR intervention.

# 5. Predictive Model Overview and Evaluation

A Linear Regression model was developed using `scikit-learn` to predict sentiment scores.  
  
Features used:  
- Message Count per Month  
- Average Message Length  
- Average Word Count per Message  
  
Performance:  
- The model achieved an R² score of approximately 0.672, indicating moderate predictive power.  
- Results suggest that more detailed message features or additional metadata could further enhance prediction accuracy.