Final Report – Employee Sentiment Analysis

# 1. Introduction

This project analyzes employee emails to evaluate their sentiment and engagement using Natural Language Processing and predictive modeling. The analysis includes sentiment labeling, exploratory data analysis (EDA), employee scoring and ranking, flight risk detection, and sentiment trend prediction.

# 2. Methodology

## Task 1: Sentiment Labeling

TextBlob was used to assign sentiment polarity scores to each message. Based on these scores, labels Positive, Negative, or Neutral were assigned using a defined threshold. The results were added as a new column in the dataset.

## Task 2: Exploratory Data Analysis (EDA)

Explored data types, missing values, sentiment distribution, and trends over time. Charts such as sentiment distribution and wordclouds were created to visualize patterns in the data.

## Task 3: Monthly Sentiment Scoring

Each message was scored as +1 for positive, -1 for negative, and 0 for neutral. Scores were aggregated monthly per employee.

## Task 4: Employee Ranking

Employees were ranked based on their monthly sentiment scores. The top 3 most positive and negative employees were identified for each month.

## Task 5: Flight Risk Detection

Employees sending 4 or more negative messages in any 30-day rolling window were flagged as flight risks. The process was automated using datetime filtering and aggregation.

## Task 6: Predictive Modeling

A linear regression model was developed to predict sentiment scores using features like message length, word count, and average polarity. The model was evaluated using R² and Mean Squared Error (MSE).

# 3. Visual Results

Visualizations such as sentiment distribution, word clouds, monthly sentiment trends, and model performance were generated and included in the visualizations/ directory.

# 4. Summary & Recommendations

This project highlights trends in employee sentiment, identifies potential attrition risks, and builds a foundation for predictive modeling. It is recommended to further validate results using labeled datasets and domain-specific models. This system can be enhanced using ensemble models or fine-tuned LLMs.