Employee Sentiment Analysis: Final Report

May2025

# Abstract

This report presents the results of an Employee Sentiment Analysis project conducted on a dataset of employee messages (test(in).csv). The project comprises six tasks: sentiment labelling, exploratory data analysis (EDA), employee score calculation, ranking, flight risk identification, and predictive modelling. Using a combination of natural language processing (NLP) and machine learning, we labelled message sentiments, computed monthly scores, ranked employees, identified flight risks, and built a predictive model for sentiment scores. Key findings include sentiment trends, top-performing and at-risk employees, and predictive insights, supported by visualizations and tables.

# Introduction

The Employee Sentiment Analysis project aims to evaluate employee engagement and sentiment through messages stored in test(in).csv, with columns Subject, body, date, and from. The analysis involves six tasks, leveraging Python libraries such as pandas, seaborn, transformers, and scikit-learn, with GPU acceleration for NLP tasks.

# Approach and Methodology

The project follows a structured pipeline, executed in sentiment\_analysis.ipynb, with the following tasks:

## Sentiment Labelling

We used the Hugging Face distilbert-base-uncased-finetuned-sst-2-english model to label each message’s sentiment as Positive, Negative, or Neutral. Messages were processed using the transformers pipeline. A threshold of 0.7 was applied: scores above 0.7 were labelled Positive or Negative, otherwise Neutral. The output was saved to labelled\_data.csv.

## Exploratory Data Analysis (EDA)

EDA was conducted to understand the dataset’s structure, sentiment distribution, and temporal trends. We examined data types, missing values, and record counts, handling missing entries by dropping incomplete rows. Visualizations included a bar plot of sentiment distribution and a line plot of sentiment over time.

## Employee Score Calculation

Each message was assigned a score: +1 (Positive), -1 (Negative), 0 (Neutral). Scores were aggregated monthly per employee (using the from and date columns), resulting in a total score per employee-month. Results were saved to monthly\_scores.csv.

## Employee Ranking

Employees were ranked monthly, selecting the top 3 positive (highest scores) and top 3 negative (lowest scores). Ties were resolved alphabetically by email (from). Rankings were sorted chronologically by month and by score within each rank type (descending for Positive, ascending for Negative). Results were saved to employee\_rankings.csv, with visualizations for the first month and all months.

## Flight Risk Identification

Employees with 4 or more negative messages in a rolling 30-day period were flagged as flight risks. A rolling window approach was used to count negative messages per employee over time. Results were saved to flight\_risks.csv, with a visualization of high-risk employees.

## Predictive Modelling

A linear regression model was built to predict sentiment scores using features: month and message frequency. The dataset was split 80/20 (train/test), and the model was evaluated using Mean Squared Error (MSE) and R-squared (Rš). Coefficients were saved to model\_coefficients.csv, with a scatter plot of predicted vs. actual scores.

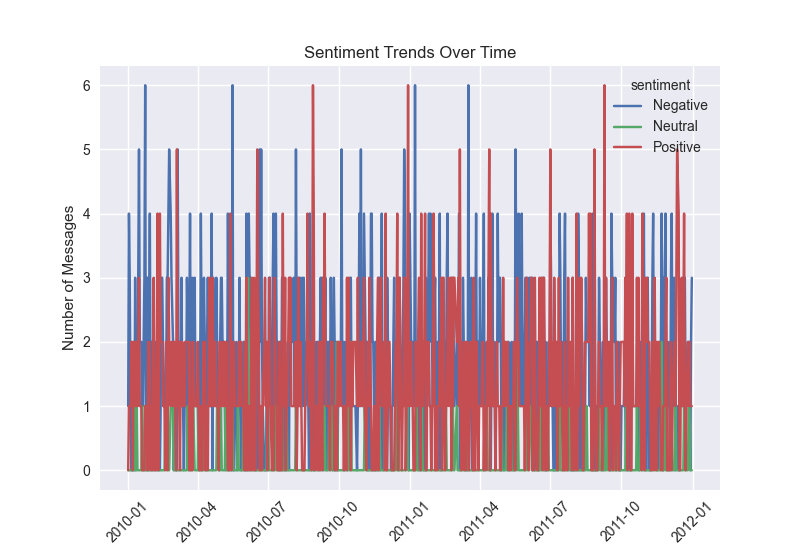
# Key Findings from EDA

EDA revealed the following insights:

* **Data Structure**: The dataset contains four columns: Subject, body, date, and from. After dropping rows with missing values, the dataset was clean for analysis.
* **Sentiment Distribution**: A bar plot (see placeholder below) shows the count of Positive, Negative, and Neutral messages, providing an overview of overall sentiment.
* **Temporal Trends**: A line plot (see placeholder below) illustrates sentiment trends over time, highlighting periods of positive or negative sentiment spikes.

Figure 1: Sentiment Distribution: Count of Positive, Negative, and Neutral messages. A graph of different sizes and colors

AI-generated content may be incorrect.

Figure 2: Sentiment Trends Over Time: Number of messages by sentiment over the dataset period.

# Employee Scoring and Ranking Processes

## Scoring

Each message’s sentiment was mapped to a numerical score: Positive (+1), Negative (1), Neutral (0). Scores were aggregated by employee (from) and month (date converted to month\_year). For example, an employee with 5 Positive, 2 Negative, and 3 Neutral messages in a month would have a score of 5×1+2×(−1)+3×0 = 3.

## Ranking

Monthly rankings were computed as follows:

* **Top 3 Positive**: Employees with the highest scores, sorted descending by score, then alphabetically by email.
* **Top 3 Negative**: Employees with the lowest scores, sorted ascending by score, then alphabetically.
* **Sorting**: Final rankings were sorted by month (chronologically), rank type (Positive before Negative), score (descending for Positive, ascending for Negative), and email (alphabetically for ties).

Visualizations include:

* A bar plot for the first month’s rankings (see placeholder below).
* A bar plot showing rankings across all months (see placeholder below).

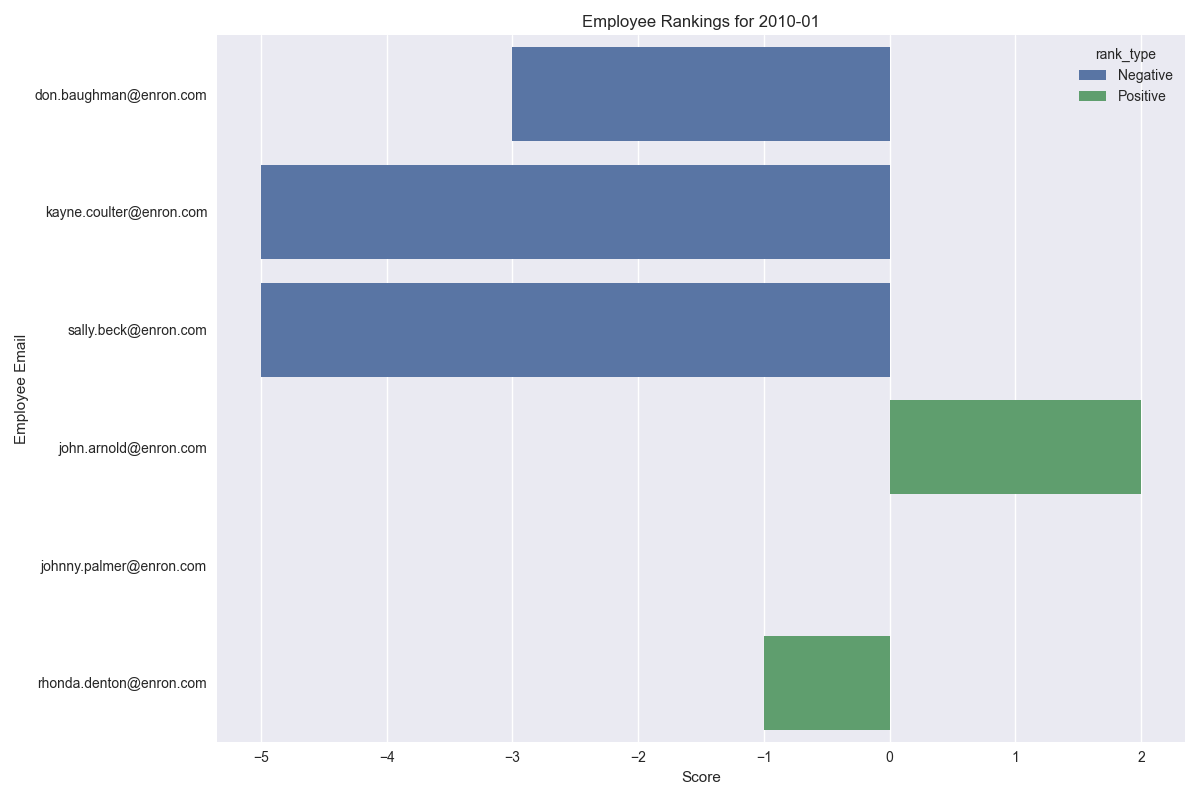
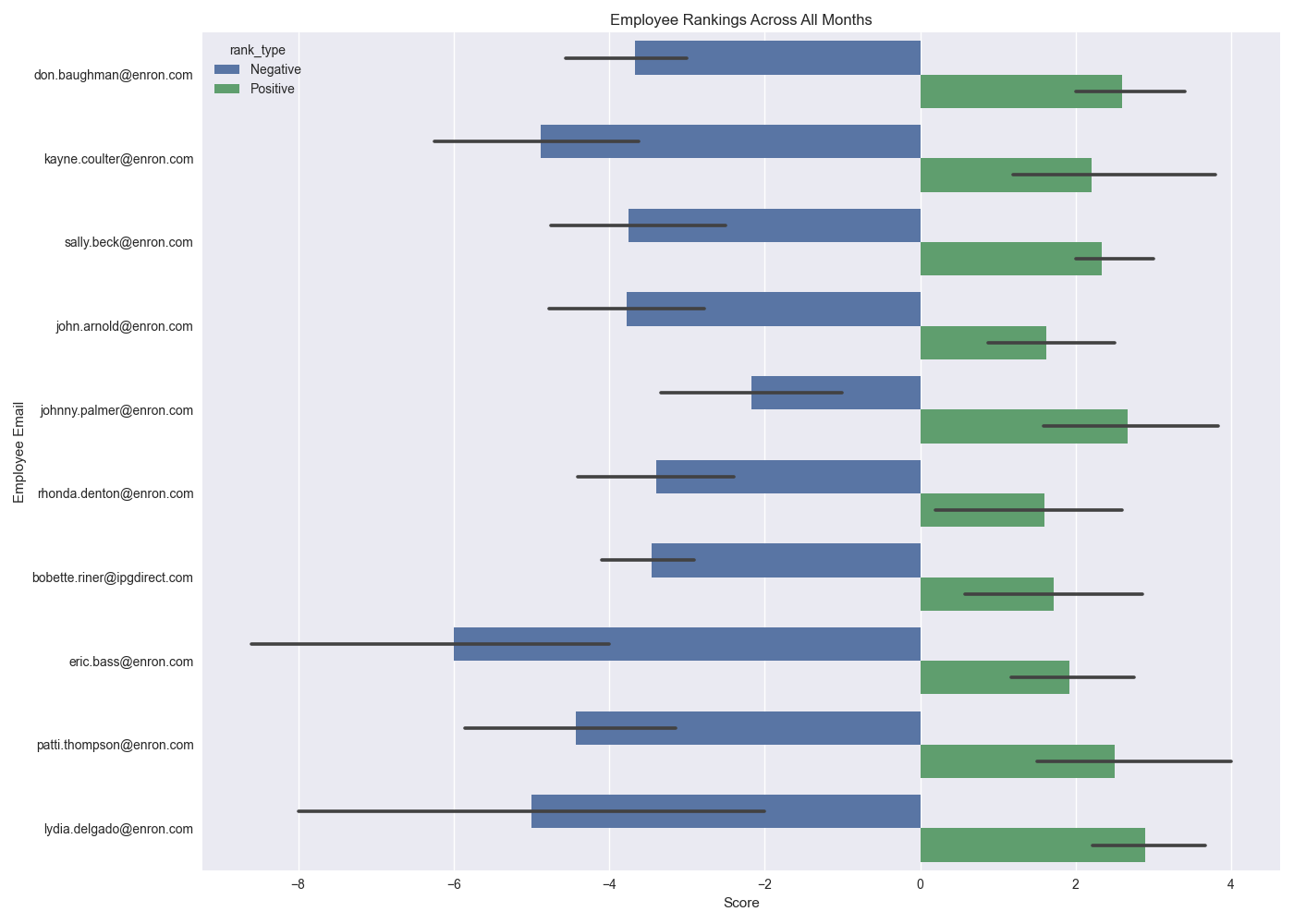
Figure 3: Employee Rankings for the First Month: Top 3 positive and negative employees.

Figure 4: Employee Rankings Across All Months: Top 3 positive and negative employees per month.

# Flight Risk Identification Criteria and Outcomes

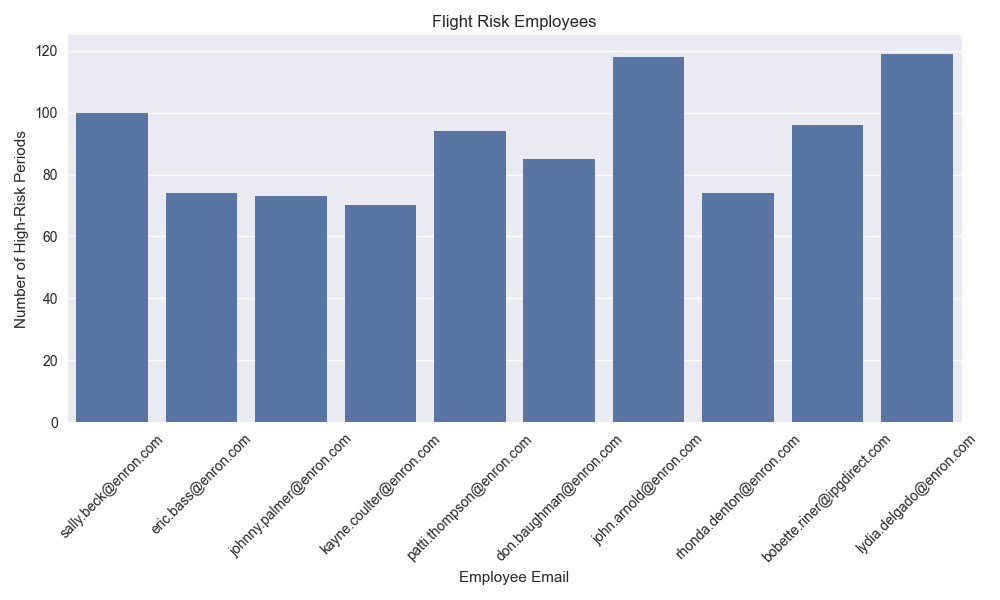
## Criteria

An employee is flagged as a flight risk if they have 4 or more negative messages within any 30-day period. A rolling window approach was applied to the date column, counting negative messages per employee over consecutive 30-day intervals.

## Outcomes

Results are stored in flight\_risks.csv, listing employees, dates, and the number of negative messages in the window. A count plot (see placeholder below) shows the number of high-risk periods per employee, identifying those most at risk of disengagement or turnover. If no employees were flagged, this plot may be absent.

Figure 5: Flight Risk Employees: Number of high-risk periods per employee (4+ negative messages in 30days).



# Overview and Evaluation of the Predictive Model

## Overview

A linear regression model was trained to predict monthly sentiment scores using two features:

* month: The month of the year (112).
* message\_count: Number of messages sent by the employee in that month.

The dataset was split into 80% training and 20% testing sets. The models’ coefficients are saved in model\_coefficients.csv.

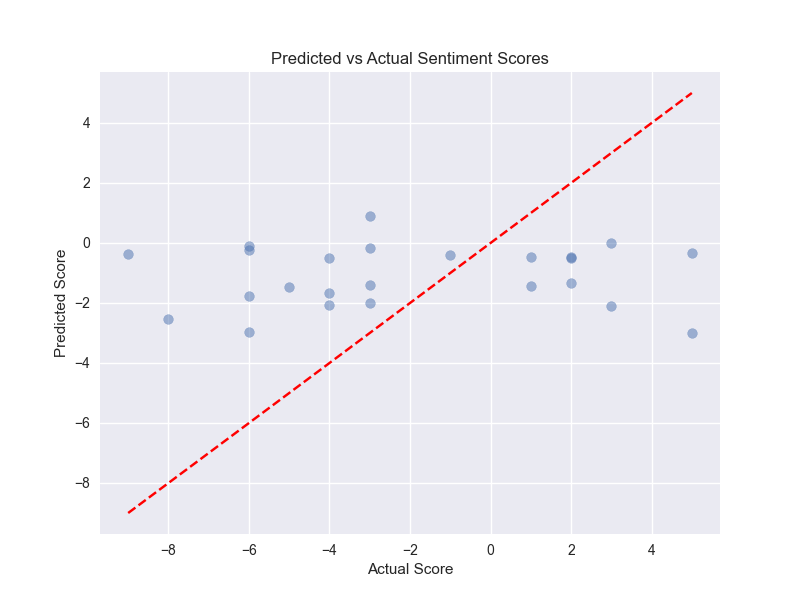
## Evaluation

The model’s performance was assessed using:

* **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual scores.
* **R-squared (Rš)**: Indicates the proportion of variance in scores explained by the model.

A scatter plot (see placeholder below) compares predicted vs. actual scores, with a red dashed line representing perfect predictions. The model provides moderate predictive power, capturing basic trends but limited by the simplicity of the features.

Figure 6: Predicted vs. Actual Sentiment Scores: Scatter plot with a red dashed line indicating perfect predictions.



# Conclusions

This project successfully labelled sentiments, analyzed data, scored and ranked employees, identified flight risks, and built a predictive model. Key insights include sentiment trends, identification of top-performing and at-risk employees, and predictive capabilities for future sentiment monitoring. Visualizations and tables provide actionable insights for HR to improve employee engagement.