# Employee Sentiment Analysis Report

Author: Kaushiki Tiwary

Date: June 18, 2025

## 1. Introduction

This project analyzes employee emails to understand sentiment trends, identify top positive/negative contributors, detect potential flight risks, and build a predictive model to forecast sentiment scores.

## 2. Methodology

- Data Cleaning: Extracted latest content from emails by removing quoted text and signatures.  
- Sentiment Analysis: Started with VADER, then switched to BERT for context-aware tagging.  
- Feature Engineering: Created features like body\_length, word\_count, is\_weekend, etc.  
- Scoring: +1 for positive, -1 for negative, 0 for neutral. Aggregated monthly per employee.  
- Flight Risk: Flagged employees with ≥4 negative emails in any rolling 30-day window.  
- Predictive Modeling: Used Linear Regression and Random Forest to predict sentiment scores.

## 3. Key Findings from EDA

- Negative sentiment spikes: April 2010 & April 2011.  
- Most emails are short and neutral.  
- Some employees consistently have high/low scores.

## 4. Employee Scoring & Ranking

- Sentiment scores were computed monthly.  
- Top positive employees consistently sent helpful, uplifting messages.  
- Negative scorers may require HR intervention.

## 5. Dashboard

**Interactive Sentiment Analysis Dashboard**

As part of this project, I developed a dynamic **interactive dashboard using ipywidgets in Jupyter Notebook**. The goal was to allow users (such as HR analysts or leadership) to **easily filter and inspect employee sentiment** across different months and categories.

**Key Features:**

* **Dropdown Filters**:

The dashboard includes **three dropdowns**:

* + **Employee**: To select any sender from the dataset.
  + **Month**: To focus on a specific time period.
  + **Sentiment Type**: To filter based on Positive, Neutral, or Negative emails.
* **Filtered View Output**:
  + A live **data table** displaying email metadata, sentiment score, and cleaned body for selected filters.
  + Useful for **inspecting behavior patterns** of individual employees.
* **Performance Summary**:
  + Total number of emails sent by the selected employee.
  + Sentiment distribution and **average email length**.
  + **Overall sentiment score** calculated based on the mapping rule: +1 (Positive), -1 (Negative), 0 (Neutral).
* **Top Performers**:
  + Displays the **Top 3 Positive** and **Top 3 Negative** employees from the entire dataset, based on sentiment scores.

**Why This Matters:**

This dashboard serves as a **HR insight tool**, providing instant visibility into employee communication patterns, behavioral sentiment trends, and potential **warning signals (like flight risk)**.

**Technologies Used:**

* ipywidgets for interactivity
* matplotlib and seaborn for visualization
* pandas for filtering and grouping
* BERT sentiment model for context-aware classification

## 6. Flight Risk Analysis

- Employees with ≥4 negative emails in any 30-day period were flagged.  
- Example: Employee X had 6 negative emails in March and was flagged.  
- This approach allows HR to take preventive action.

## 7. Predictive Modeling

- Models used: Linear Regression & Random Forest Regressor.  
- Random Forest performed better with higher R² and lower MSE.  
- Forecasted scores for the next 3 months.  
- Some deviation in predicted vs actual trend in earlier months.

## 8. Recommendations

- Monitor sentiment scores in monthly reports.  
- Address flagged flight-risk employees.  
- Enhance model with more HR signals like engagement surveys or leave records.

## Visualizations :

EDA :

1. Monthly Email Activity

A graph of blue and red squares

Description automatically generated

1. Top 10 Senders by Average Email

A graph of email length

Description automatically generated with medium confidence

1. Quarterly Email Volume

A graph showing the number of sales

Description automatically generated

1. Employee Vs Sentiment

A graph of a number of people

Description automatically generated

1. Length of the message vs the sentiment

A graph of a number of different colored squares

Description automatically generated with medium confidence

1. Quarterly Sentiment Trends

A graph of a bar chart

Description automatically generated with medium confidence

1. Monthly Sentiment Trends

A graph of a graph with numbers and a line

Description automatically generated with medium confidence

1. Total Negative Emails

A graph of red bars

Description automatically generated with medium confidence

1. Heat Map for Important features

A screenshot of a graph

Description automatically generated

1. Linear Regression

A graph with blue and orange lines

Description automatically generated

1. Random Forest

A graph with blue and orange lines

Description automatically generated

Confusion Matrix :   
  
A graph of confusion matrix

Description automatically generated

At the end, Random forest worked well with 71% r2 and 76% accuracy as compared to the other models such as linear regression and XG boost.