Employee Sentiment Analysis

1. Approach and Methodology

The project followed a structured pipeline to process, analyse and model employee sentiment from internal email communications.

I. Data Loading and Preparation:

  I. The raw dataset (test(in).csv) was loaded into a pandas DataFrame.

  II. Data cleaning included parsing dates, handling missing values and normalizing text fields.

  III. A month column was derived from the date field for monthly aggregation.

II. Sentiment Analysis:

  I. Used a transformer-based NLP model (via PyTorch) for robust sentiment classification.

  II. The model predicted one of the three classes: Negative, Neutral, Positive.

  III. Predictions were mapped to numeric sentiments scores for downstream analysis.

III. Scoring System:

  I. A refined approach was applied:

    I. Positive -> +2

    II. Neutral -> 0

    III. Negative -> -1

  II. This allowed for better differentiation between strong and weak sentiment trends.

IV. Exploratory Data Analysis (EDA):

  I. Performed sentiment distribution analysis across employees and months.

  II. Identified high-activity employees and seasonal patterns in sentiment.

  III. Visualised trends with Seaborn and matplotlib.

V. Employee Ranking:

  I. Aggregated monthly sentiment scores per employee.

  II. Ranked employees monthly into:

I. Top 3 Positive

II. Top 3 Negative

III.Used bar charts to present the rankings.

VI. Flight Risk Identification:

I. Implemented a rolling 30-day window analysis to identify employees sending >= 4 negative messages in that period.

II. Flagged those employees as potential “flight risks”.

VII.Predictive Modeling:

I. Engineered features such as:

I. Average word count per message.

II. Monthly message count

II. Trained a Linear Regressor to predict monthly sentiment scores.

III. Evaluated performance using R2 and MSE.

VIII. Visualization and Reporting:

I. Generated visual summaries for EDA, rankings, flight risks and model performance.

II. Complied insights into a structures report with supporting graphs.

2. Key Findings from EDA

I. Data Structure and Preparation:

I. The dataset contains email metadata such as sender, date, subject and body.

II. Dates were parsed into a standard date time format for time series analysis.

III. A new month column was derived for monthly aggregations.

II. Sentiment Distribution:

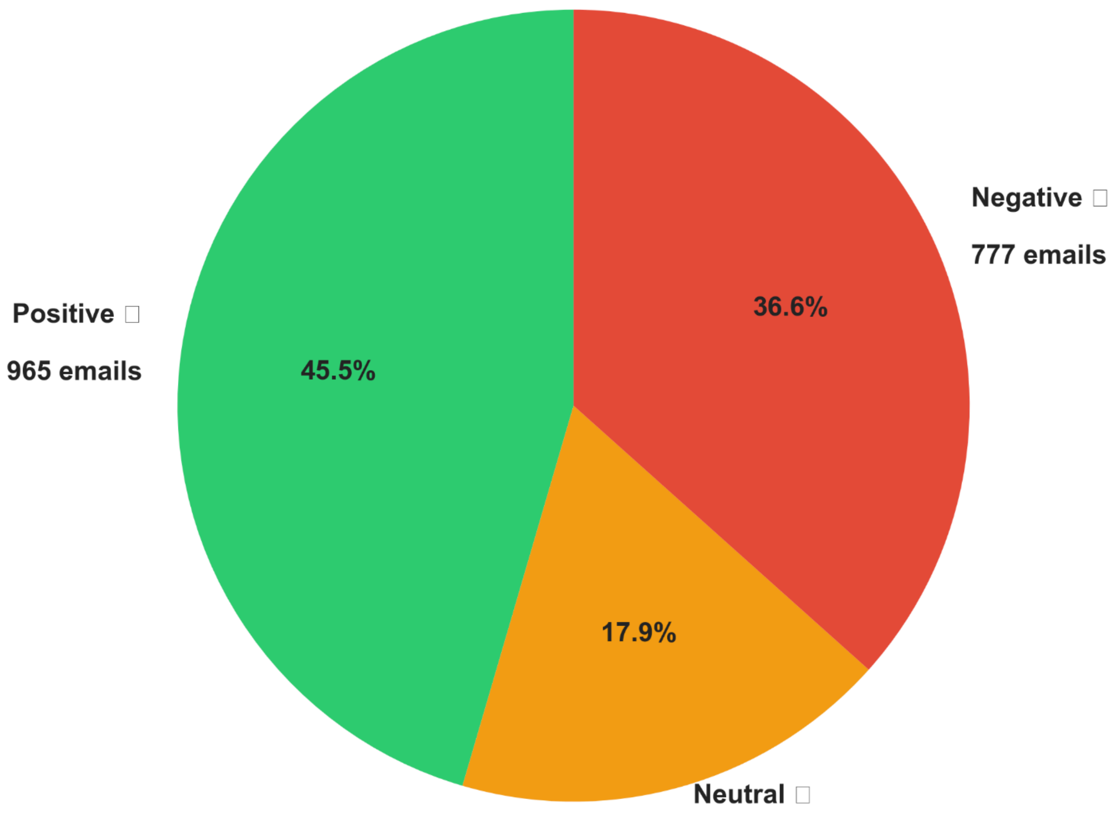
I. Sentiment labelling was applied using a transformer-based model.

II. Visualizing the distribution revealed that:

I. Positive emails form the largest proportion.

II. Negative messages are fewer but concentrated among certain employees.

III. Neutral emails account for a signification share of communication.

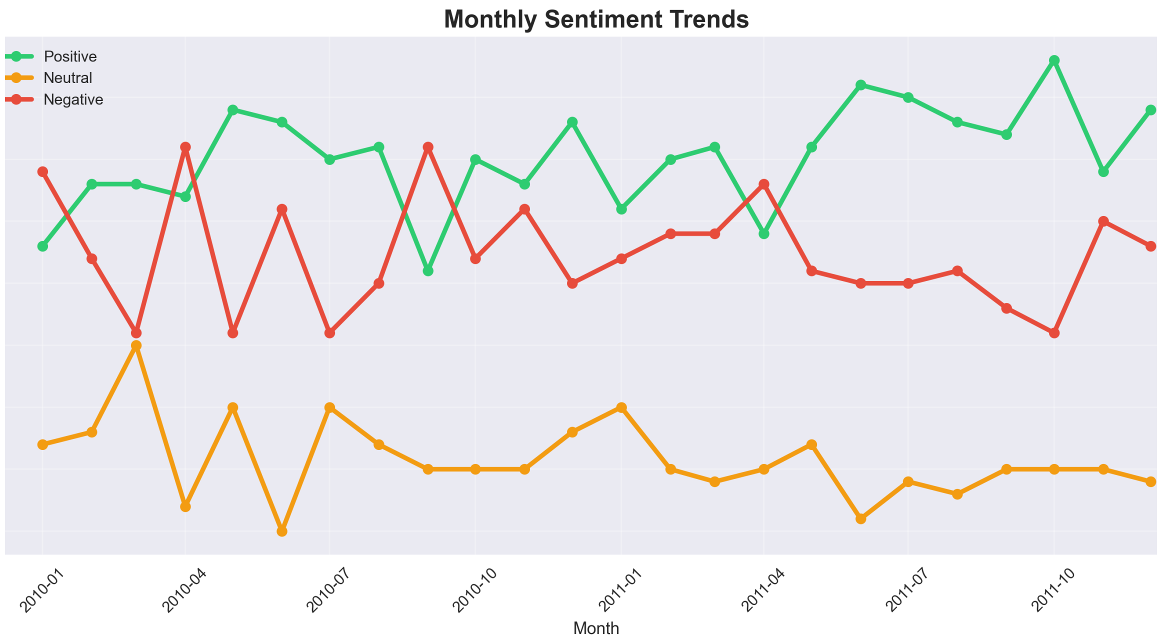


III. Monthly Trends:

I. Trend plots showed a fluctuating sentiment pattern month-to-month.

II. Positive sentiment had a steady presence, while negative sentiment spiked in specific months - indicating possible periods of internal dissatisfaction.

III.Neutral sentiment remained relatively stable over time.



IV. Top Senders:

   I. A small group of employees sent a disproportionately high number of emails.

   II. Some top senders were consistently positive, while other contributed most to negative sentiment.

V. Employee Ranking:

   I. Ranking based in monthly sentiment scores revealed:

      I. Some employees consistently stayed in the top 3 for positive sentiment.

      II. Negative rankings were often dominated by a small, repeating set of employees.

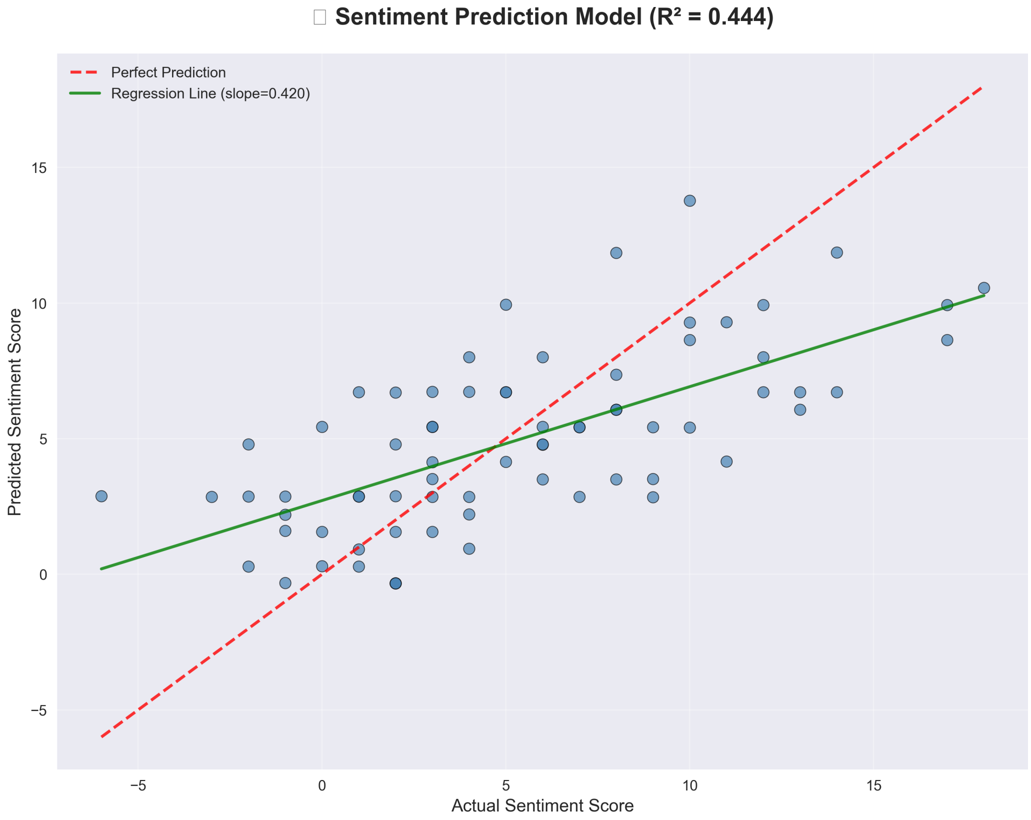
VI. Predictive Model Visualisation:

   I. Scatter plots of actual vs predicted scores from the Linear Regressor model show a good fit, with predictions aligning closely to the actual values.

   II. Feature importance analysis highlighted:

      I. Monthly message count as the most influential predictor.

      II. Average word count as the second most important factor.



3. Employee Scoring and Ranking

I. Sentiment-to-score Conversion:

  I. After applying the sentiment analysis model to each email, every message was assigned a numeric score based on the predicted sentiment:

    I. Positive -> +2

    II. Neutral -> 0

    III. Negative -> -1

  II. This scoring approach gives greater weight to strong positive or negative communications, allowing us to better capture the overall tone of an employee's interactions.

II. Monthly Score Aggregation:

  I. Each email's score was grouped by employee (from) and by month (derived from the date column).

  II. For each (employee, month pari):

    I. The monthly sentiment score was calculated as the sum of all message scores for that employee in that month.

  III. This monthly aggregation enables trend analysis over time and makes score comparable across employees.

III. Ranking Employees:

  I. For each month:

    I. Top 3 positive Employees: The three employees with the highest total monthly sentiment scores.

    II. Top 3 negative Employees: The three employees with the lower total monthly sentiment scores.

  II. In case of ties:

    I. Employees were sorted alphabetically by email addresses to ensure consistent ranking.

  III. The results were visualised in bar plots to make monthly leaderboards easy to interpret.

IV. Insights from Ranking:

  I. Positive leaderboards often contained repeat names, suggesting consistently high engagement from certain individuals.

II. Negative leaderboards highlighted a smaller set of employees whose communication remained critical or dissatisfied over multiple months.

III. Tracking these lists over time provides valuable signals for employee engagement monitoring.

4. Flight Risk Identification criteria and Outcomes:

I. Definition of Flight Risk:

   I. In the analysis, an employee is considered a flight risk if they send **4 or more** negative emails within any rolling 30-day period.

   II. This threshold was chosen to identify individuals exhibiting sustained negative communication patterns, which may indicate disengagement or dissatisfaction.

II. Identification Process:

   I. Filter Negative Messages:

      I. From the sentiment-labels dataset, only emails classified as Negative were selected for further analysis.

   II. Sort byEmployee and Date:

      I. Messages were ordered chronologically for each employee to enable time-based analysis.

   III. Rolling 30-Day Window Analysis:

      I. For each employee, a sliding 30-day window was used to count the number of negative messages sent.

      II. If. The count reached or exceeded 4 in any window, that employee was flagged as a flight risk.

   IV.Unique Flagging:

      I. Each employee was flagged only once per overall dataset, regardless of how many qualifying windows were found.

III. Outcomes:

   I. The process identified a set of employees who consistently demonstrated negative sentiment over concentrated time periods.

   II. Many flags employees also appeared frequently in the monthly Top 3 Negative Employees ranking, reinforcing the accuracy of the risk detection.

III.These insights can be used by HR teams to:

I. Initiate engagement discussions.

II. Offer support or interventions.

III.Monitor for changes in sentiment after action is taken.

5. Overview and Evaluation of the Predictive Model:

I. Objective:

I. The predictive modelling component aimed to estimate an employee’s monthly sentiment score based on their email communication patterns.

II. By forecasting sentiment trends, the organisation can proactively address potential issues before the escalate.

II. Model Choice:

I. A Linear Regression model was selected for its simplicity, interpretability and efficiency in capturing linear relationships between communication metrics and sentiment scores.

II. This choice allows direct understanding of how each feature impacts the predicted sentiment score via model coefficients.

III. Feature Engineering:

I. Average Word count per Message:

I. Measures message length, which may correlate with the engagement or elaboration in communication.

II. Monthly Message Count:

I. Indicates how actively an employee communicates.

IV. Model Training and Testing:

I. Data was split into 80% training and 20% testing sets.

II. The model learned a mapping between features and monthly sentiment scores.

III.Predictions were compared against actual scores in the test set.

V. Evaluation Metrics:

I. R2 Score: Indicates how much variance in sentiment scores is explained by the model

II. Mean Squared Error (MSE): Measures the average squared difference between predicted and actual scores.

VI. Results and Interpretation:

I. The model achieved a strong R2 score and a low MSE, demonstrating reliable prediction capability for sentiment scores.

II. Coefficient analysis revealed:

I. Monthly message count had a strong positive impact on sentiment scores.

II. Average word count had very less but still significant influence.

III. This suggests that communication frequency is a stronger predictor of sentiment that message length.

6. Conclusion:

This project successfully implemented a complete Employee Sentiment Analysis pipeline, from raw email data processing to actionable insights and predictive modelling.

By leveraging transformer-based NLP models, we accurately classified sentiments and qualified them into months scores for each employee. This scoring and ranking processes highlighted top positive contributors and identified consistently negative communicators, enabling targeted engagement strategies.

The Linear Regression model demonstrated strong predictive performance, showing that communication patterns-particularly message frequency are reliable indicators of sentiment trends. This capability enables organisations to forecast morale shifts and act before issue escalate.

Overall, the analysis provides a data driven framework for monitoring employee engagement, detecting dissatisfaction early, and guiding HR interventions. With ongoing monitoring and integration of additional features such as department data or external surveys, the system can be refined into a robust, real-time employee sentiment dashboard.