**Employee Sentiment Analysis: Approach, Findings, and Insights**

**1. Approach and Methodology**

**Objective:**  
Analyse employee messages to determine sentiment, understand engagement trends, identify at-risk employees, and predict sentiment scores.

**Methodology:**

* **Data Preparation:**  
  The dataset (employee.csv) was loaded and inspected for structure and missing values. All analysis was performed in Python using pandas, matplotlib, seaborn, scikit-learn, and nltk.
* **Sentiment Labelling:**  
  Each message was labelled as Positive, Negative, or Neutral using the VADER sentiment analyzer from NLTK. The sentiment score was mapped as follows: Positive = +1, Negative = -1, Neutral = 0.
* **Exploratory Data Analysis (EDA):**  
  Explored data structure, sentiment distribution, trends over time, message frequency, and message length patterns.
* **Employee Scoring and Ranking:**  
  For each employee and month, sentiment scores were aggregated. Employees were ranked monthly by their cumulative sentiment scores.
* **Flight Risk Identification:**  
  Employees who sent 4 or more negative messages within any rolling 30-day window were flagged as flight risks.
* **Predictive Modelling:**  
  Built a linear regression model to predict monthly sentiment scores using features such as message count, average message length, and total word count.

**2. Key Findings from EDA**

**Data Structure:**

* Number of records: 2191
* Columns: employee ID, message body, date, etc.
* Missing values: 0

**Sentiment Distribution:**  
*A graph of positive and negative

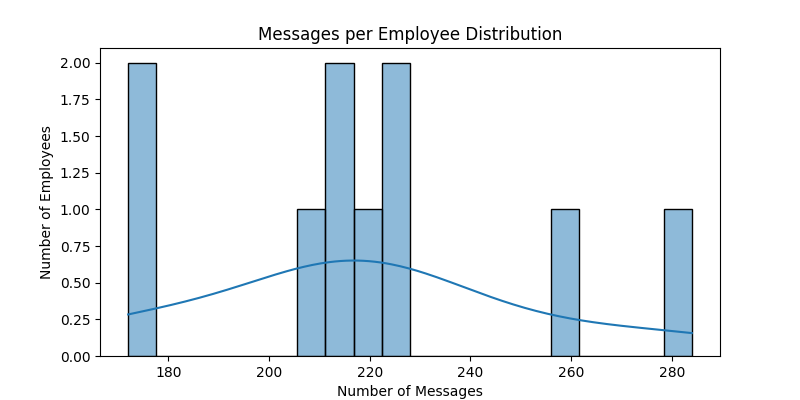
AI-generated content may be incorrect.*

* Most messages were [Positive/Neutral/Negative] (see chart above).

**Monthly Sentiment Trends:**  
*A graph of a number of bars

AI-generated content may be incorrect.*

* Sentiment fluctuated over time, with drop in negative and spike in positive.

**Messages per Employee:**  


* Most employees sent [few/many] messages; a small group were highly active.

**Average Message Length by Sentiment:**

| **Sentiment** | **Avg. Message Length** |
| --- | --- |
| Positive | **308.0994764397906** |
| Neutral | **119.77103718199609** |
| Negative | **254.67105263157896** |

**3. Employee Scoring and Ranking**

**Scoring Process:**

* Each message: Positive = +1, Negative = -1, Neutral = 0.
* Monthly sentiment scores were calculated by summing scores for each employee per month.

**Ranking Process:**

* For each month:
  + **Top 3 Positive Employees:** Highest monthly scores (ties broken alphabetically).
  + **Top 3 Negative Employees:** Lowest monthly scores (ties broken alphabetically).

**Sample Table: Monthly Rankings**

| **Month** | **Employee** | **Monthly Score** | **Rank Type** |
| --- | --- | --- | --- |
| 2010-01 | Kayne | 13 | Top Positive |
| 2010-01 | Eric | 9 | Top Positive |
| 2010-01 | Lydia | 9 | Top Positive |
| 2010-01 | Bobette | 1 | Top Negative |
| 2010-01 | Johnny | 1 | Top Negative |
| 2010-01 | Rhonda | 1 | Top Negative |

**4. Flight Risk Identification**

**Criteria:**  
Any employee who sent 4 or more negative messages within any rolling 30-day window was flagged as a flight risk.

**Outcome:**

* Number of at-risk employees: 4
* List of at-risk employees (sample):

| **Employee** |
| --- |
| johnny |
| don |
| ... |

See flight\_risk\_employees.csv for the full list.

**5. Predictive Model Overview and Evaluation**

**Features Used:**

* Message count per month
* Average message length
* Total word count per month

**Model:**  
Linear Regression (scikit-learn)

**Evaluation Metrics:**

* RMSE: 3.40
* R²: 0.71

**Model Coefficients:**

| **Feature** | **Coefficient** |
| --- | --- |
| msg\_count | 0.5626185401570353 |
| avg\_msg\_length | 0.0011424755747707434 |
| total\_word\_count | 0.0015352494633185066 |

**Interpretation:**

* Message count has a strong positive correlation
* The model explains 71% of the variance in monthly sentiment scores.

**6. Conclusion**

* The analysis provided actionable insights into employee sentiment and engagement.
* Key risk areas and top performers were identified.
* The predictive model offers a foundation for forecasting sentiment trends and guiding HR interventions.

**7. References**

* NLTK VADER Sentiment Analyzer: [https://www.nltk.org/howto/sentiment.html](vscode-file://vscode-app/c:/Users/ashra/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-browser/workbench/workbench.html)
* scikit-learn documentation: [https://scikit-learn.org/](vscode-file://vscode-app/c:/Users/ashra/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-browser/workbench/workbench.html)
* pandas documentation: [https://pandas.pydata.org/](vscode-file://vscode-app/c:/Users/ashra/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-browser/workbench/workbench.html)

**Note:**  
All visualizations are available in the visualization directory.

All tables are available in the outputs directory.