Employee Sentiment Analysis & Flight Risk Report

Prepared by: Aryan Anand Srivastava

# 1. Approach & Methodology

This project analyzes employee email sentiments to uncover trends, score and rank employees, identify potential flight risks, and build a predictive model of monthly sentiment scores. The pipeline includes: data cleaning, EDA, monthly scoring, ranking, 30-day rolling flight-risk detection, and a linear regression model with interpretable features.

# 2. Exploratory Data Analysis (EDA)

Total messages: 2191

Unique employees: 10

Date range: 2010-01-01 to 2011-12-31

Missing values by column:

|  |  |
| --- | --- |
| Column | Missing Count |
| Subject | 0 |
| body | 0 |
| date | 0 |
| from | 0 |
| text | 0 |
| Sentiment | 0 |
| text\_used | 0 |
| msg\_score | 0 |
| YearMonth | 0 |

Figure 1: Sentiment distribution

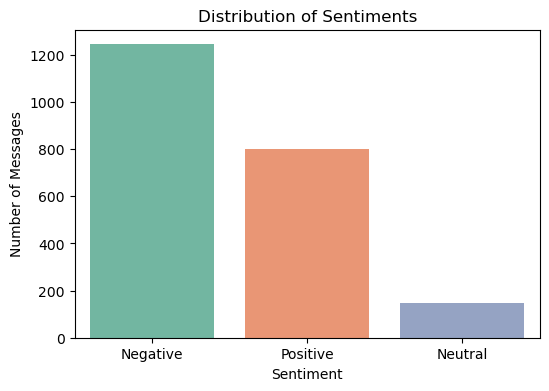


Figure 2: Monthly sentiment counts over time

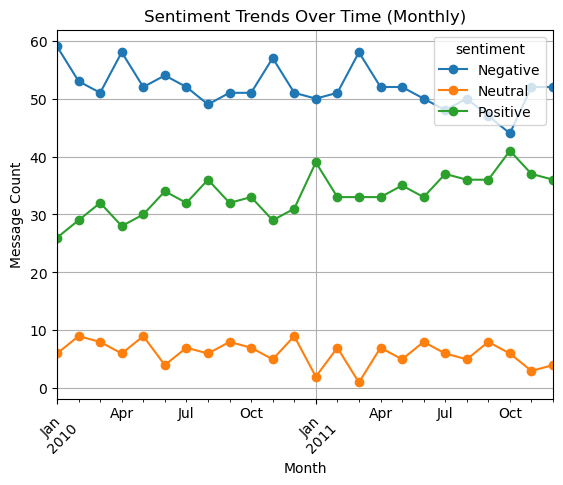
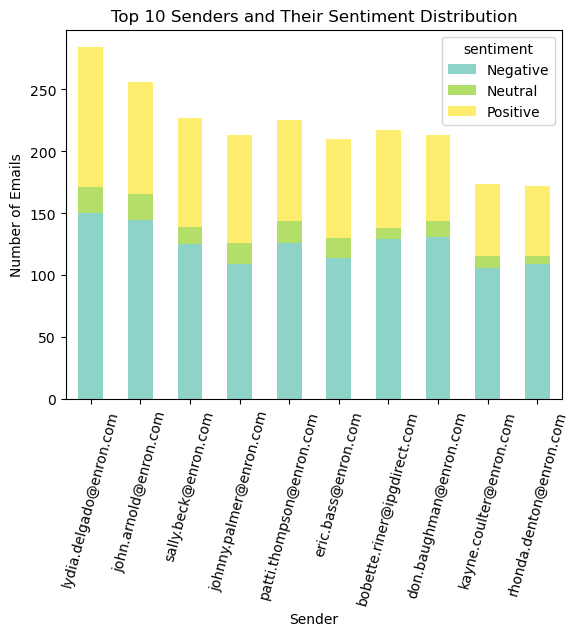


Figure 3: Top 10 senders and their sentiment mix



# 3. Employee Sentiment Scoring & Ranking

Scoring rule per message: Positive = +1, Negative = −1, Neutral = 0. Monthly scores were computed as the sum of message scores per employee per month. Top Positive and Top Negative rankings were derived each month by sorting by score (descending for positives, ascending for negatives) with alphabetical tie-breakers.

Latest month: 2011-12

Top Positive Employees (latest month)

|  |  |  |
| --- | --- | --- |
| Rank | Employee | Sentiment Score |
| 1 | john.arnold@enron.com | 2 |
| 2 | johnny.palmer@enron.com | 1 |
| 3 | patti.thompson@enron.com | 0 |

Top Negative Employees (latest month)

|  |  |  |
| --- | --- | --- |
| Rank | Employee | Sentiment Score |
| 1 | bobette.riner@ipgdirect.com | -3 |
| 2 | don.baughman@enron.com | -3 |
| 3 | eric.bass@enron.com | -3 |

# 4. Flight Risk Identification

Definition: An employee is flagged as Flight Risk if they have sent 4 or more negative emails within any rolling 30-day window (cross-month). This rolling window ensures robust detection independent of calendar months.

Flagged employees:

|  |  |  |
| --- | --- | --- |
| **From** | **First\_Flag\_Date** | **Total\_Neg\_Mails** |
| [bobette.riner@ipgdirect.com](mailto:bobette.riner@ipgdirect.com) | 2010-04-21 | 129 |
| [don.baughman@enron.com](mailto:don.baughman@enron.com) | 2010-01-26 | 131 |
| [eric.bass@enron.com](mailto:eric.bass@enron.com) | 2010-09-03 | 114 |
| [john.arnold@enron.com](mailto:john.arnold@enron.com) | 2010-01-10 | 145 |
| [johnny.palmer@enron.com](mailto:johnny.palmer@enron.com) | 2010-03-25 | 109 |
| [kayne.coulter@enron.com](mailto:kayne.coulter@enron.com) | 2010-02-06 | 106 |
| [lydia.delgado@enron.com](mailto:lydia.delgado@enron.com) | 2010-10-10 | 150 |
| [patti.thompson@enron.com](mailto:patti.thompson@enron.com) | 2010-01-12 | 126 |
| [rhonda.denton@enron.com](mailto:rhonda.denton@enron.com) | 2010-04-09 | 109 |
| [sally.beck@enron.com](mailto:sally.beck@enron.com) | 2010-05-13 | 125 |

# 5. Predictive Modeling (Linear Regression)

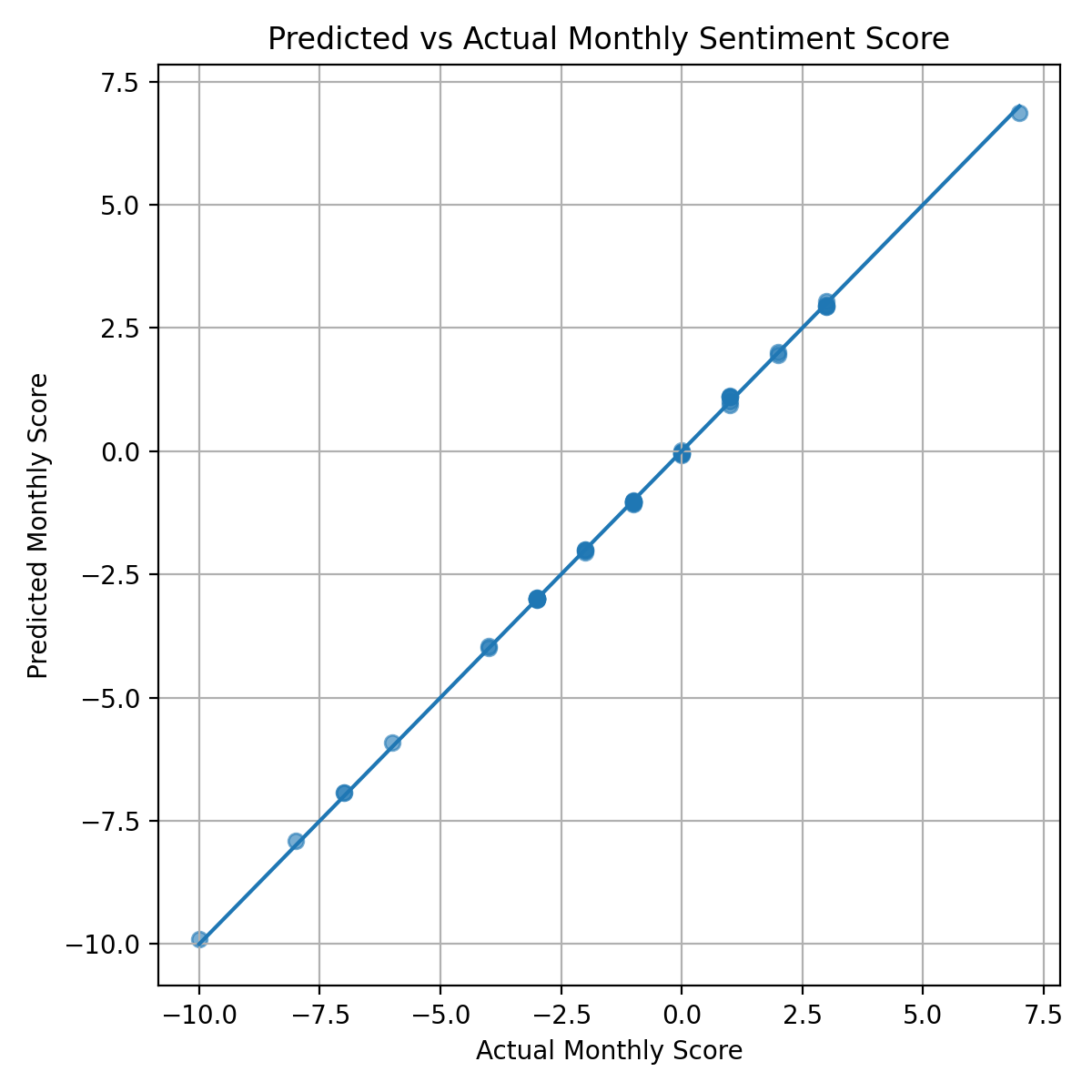
Target: monthly sentiment score per employee. Features included message volume, text length, sentiment composition (counts/ratios), activity breadth (active days), weekend share, and a lag-1 score to capture momentum. We used a time-aware split (last ~20% months as test) to avoid leakage, and Ridge regression for stability.

Test Performance:  
R²: 1.000 | MAE: 0.044 | RMSE: 0.056

Approximate standardized coefficients (larger magnitude = stronger influence):

|  |  |
| --- | --- |
| Feature | Std. Coefficient |
| pos\_count | 2.9292 |
| neu\_count | 0.0530 |
| avg\_len\_words | 0.0245 |
| pos\_ratio | 0.0175 |
| weekend\_share | 0.0059 |
| lag1\_score | 0.0021 |
| avg\_len\_chars | -0.0136 |
| max\_len\_words | -0.0141 |
| neg\_ratio | -0.0331 |
| active\_days | -0.0501 |
| msg\_count | -0.5571 |
| neg\_count | -3.1345 |

Figure 4: Predicted vs. Actual Monthly Sentiment Scores



# 6. Conclusion

We established a reproducible pipeline for sentiment analysis across emails, producing clear EDA insights, monthly employee scoring and rankings, robust rolling-window flight-risk flags, and an interpretable predictive model. This framework can be enhanced with richer metadata (role, team, seniority), topic modeling, network features, and more advanced models to improve forecasting accuracy.