Employee Sentiment Analysis Project

Final Report

Andrew Cho

July 8, 2025

**Introduction**

This project analyzes employee email messages to determine sentiment, ranks employees by engagement, identifies flight risks, and builds a predictive model to analyze sentiment trends and engagement patterns.

**Limitations**

This project used TextBlob with default polarity thresholds (polarity > 0 as positive, < 0 as negative, = 0 as neutral). However, these thresholds are not tuned for corporate email contexts and may misclassify subtle sentiments. Additionally, TextBlob is trained primarily on informal datasets like reviews or tweets, which may not generalize well to formal internal communication. In production, validating thresholds against labeled corporate email data and testing multiple models would improve accuracy and contextual relevance.

**Task 1: Sentiment Labeling**

Approach: Used TextBlob to calculate polarity scores for each message in the `body` column.

Labeling Criteria:

- Polarity > 0 → Positive

- Polarity < 0 → Negative

- Polarity = 0 → Neutral

Justification: TextBlob offers a straightforward, interpretable, and reproducible rule-based approach that does not require labeled training data.

**Task 2: Exploratory Data Analysis (EDA)**

Key Findings:

- Dataset contained ~2191 messages with columns: Subject, body, date, from.

- **Sentiment Distribution:**

- Positive: 1218 messages

- Neutral: 703 messages

- Negative: 270 messages

- **Monthly trends:** Message volume was consistent each month.

- **Message length analysis**: Average message length varied slightly by sentiment label, with no major outliers.

**Visuals:** Includes saved plots of:

- Sentiment label distribution plot

- Messages per month plot

- Boxplot of message length by sentiment

**Task 3: Monthly Sentiment Scoring**

Methodology:

Assigned +1 for Positive, –1 for Negative, and 0 for Neutral messages.

Aggregated sentiment scores per employee per month to calculate monthly sentiment.

**Task 4: Employee Ranking**

Methodology:

Sorted monthly sentiment scores:

- Top 3 positive employees → highest scores, then alphabetical order

- Top 3 negative employees → lowest scores, then alphabetical order

Sample Rankings (January 2010):

**Top 3 Positive Employees:**

- kayne.coulter@enron.com (9)

- don.baughman@enron.com (5)

- eric.bass@enron.com (5)

**Top 3 Negative Employees:**

- rhonda.denton@enron.com (2)

- bobette.riner@ipgdirect.com (2)

- johnny.palmer@enron.com (0)

**Task 5: Flight Risk Identification**

Criteria: Employees with 4 or more negative emails within any rolling 30-day window.

Method: Filtered negative emails and used pandas rolling window count (30D) per employee to identify flagged employees.

Outcome: Employees flagged as flight risks:

- eric.bass@enron.com

- don.baughman@enron.com

- rhonda.denton@enron.com

- bobette.riner@ipgdirect.com

- lydia.delgado@enron.com

- patti.thompson@enron.com

- johnny.palmer@enron.com

- sally.beck@enron.com

- john.arnold@enron.com

**Task 6: Predictive Modeling**

Features Used:

- Message count per employee per month

- Average message length per employee per month

Model: Linear Regression using scikit-learn.

Evaluation Results:

- Mean Squared Error: 4.12

- R²: 0.55

Interpretation:

- \*\*message\_count coefficient (0.418):\*\* Positive correlation with sentiment score.

- \*\*avg\_message\_length coefficient (0.0026):\*\* Minimal impact.

**Conclusion and Recommendations**

Overall, employee sentiment trends remain largely positive or neutral. Employees with high negative message counts were flagged as flight risks and should be monitored with targeted retention interventions. Predictive modeling showed message frequency as a stronger indicator of sentiment compared to message length. Regular monthly sentiment analysis and proactive check-ins with at-risk employees are recommended for optimal engagement and retention.