Project: Employee Sentiment Analysis

Final Report

Laasya Priya Vemuri

Overview & Objective:

This project involved analyzing a dataset of internal employee messages with the goal of assessing sentiment, tracking changes in engagement over time, and identifying at-risk employees. The tasks ranged from sentiment labeling and EDA to building a predictive model. I implemented every step myself, and this document is a reflection of the entire process — including how I approached each task, what methods I used, the results I got, and key takeaways from the analysis.

Approach & Methodology:

1. **Sentiment Labeling**
   Used `TextBlob` (a basic but effective NLP tool) to assign each message a sentiment — Positive, Neutral, or Negative.

2. **EDA (Exploratory Data Analysis)**
   Analyzed message volumes, sentiment distribution, trends over time, and activity levels across employees.

3. **Monthly Sentiment Score Calculation**
   Created monthly sentiment scores for each employee by assigning +1 (positive), 0 (neutral), and -1 (negative) to every message.

4. **Employee Ranking**
   Identified the top 3 most and least positively scoring employees per month, sorted by score and tie-broken alphabetically.

5. **Flight Risk Detection**
   Flagged employees as "at risk" if they sent 4 or more negative messages in any rolling 30-day window.

6. **Predictive Modeling**
   Built a linear regression model to predict future sentiment scores using `month_index` and `message_count` as features.

All processing was done using Python, with pandas, matplotlib/seaborn for visualizations, and scikit-learn for the predictive model. Every output was saved, and I made sure each task built on the previous one to keep everything consistent and traceable.

Key EDA Findings:

After labeling all messages using `TextBlob`, I conducted a detailed EDA to understand patterns in sentiment and activity.

Sentiment Distribution

- Positive messages dominated the dataset, accounting for over 1,200 out of 2,191 total.
- Neutral messages came next (approximately 700).
- Negative messages were the least frequent (approximately 270).

Monthly Trends

- Positive sentiment remained consistently high month after month.
- Neutral messages showed some fluctuation, peaking in late 2010.
- Negative messages were sparse but exhibited brief spikes during certain months.

Top Active Employees

The most active sender was `lydia.delgado`, followed closely by `john.arnold` and `sally.beck`. This helped identify whose communication volume might significantly impact overall sentiment trends.

All of this provided a strong understanding of who was speaking the most and how their tone evolved over time.

Monthly Sentiment Scoring (Task 3):

For every employee and month, I assigned:

- $+1$ to every positive message,
- -1 to every negative message,
- 0 to neutral ones.

Then, I aggregated the results month-by-month. This gave me a numeric score representing that employee's overall monthly sentiment. Here's what stood out:

- In January 2010, `kayne.coulter` had the highest score (9), followed by `don.baughman`, `eric.bass`, and `lydia.delgado` (5 each).
- Some employees had neutral or near-zero sentiment scores, indicating either low activity or a balance of tone.

This structured view of scores made it much easier to detect fluctuations and compare employees fairly over time.

Employee Ranking (Task 4):

To make the sentiment scores actionable, I built two rankings per month:

- **Top 3 Positive Employees** (highest sentiment score)
- **Top 3 Negative Employees** (lowest score)

I made sure to sort scores in descending order and then alphabetically to break ties in a reproducible way.

In January 2010, for example:

- **Most Positive**: `kayne.coulter` (9), `don.baughman` (5), `eric.bass` (5)
- **Least Positive** (aka more neutral/negative): `rhonda.denton` (0), `bobette.riner` (2), `johnny.palmer` (2)

This gave a clear snapshot of who was thriving and who might be withdrawing — really helpful if someone were tracking morale month over month.

Flight Risk Detection (Task 5):

This task was probably the most sensitive one. The idea was to identify employees who might be disengaging or unhappy, based purely on communication.

I defined a Flight Risk as:

> "Any employee who sent 4 or more negative emails within any rolling 30-day window".

The logic was implemented using a loop over each employee's negative messages sorted by date. The process was carefully written to stop checking further once a risk was flagged.

Here's who got flagged:

- `bobette.riner`
- `don.baughman`
- `eric.bass`
- `john.arnold`
- `johnny.palmer`
- `kayne.coulter`
- `lydia.delgado`
- `patti.thompson`
- `rhonda.denton`
- `sally.beck`

These were employees who had clusters of negativity, and would definitely warrant a follow-up in a real organization.

Predictive Modeling Summary (Task 6):

Finally, I attempted to forecast an employee's future sentiment score using:

- `month_index` (how many months into the timeline)
- `message_count` (how active they were that month)

I trained a Linear Regression model and got these results:

- **R^2 Score**: 0.526
- **Mean Absolute Error (MAE)**: 1.574
- **Root Mean Squared Error (RMSE)**: 1.944

Visually, the actual vs predicted scores showed a strong upward trend — indicating that time progression and email volume do have a measurable effect on sentiment.

It's not a perfect predictor, but it's reliable enough to suggest there's behavioral correlation that can be used for early detection or performance trend monitoring.

Final Thoughts & Observations:

- Positive sentiment dominates, which is great, but the negative ones still deserve attention because of their disproportionate impact.
- Activity levels matter. The most vocal employees are also those driving the sentiment metrics.
- Flight risk detection is crucial — short bursts of negativity can easily get buried if we don't look at them holistically, especially in rolling windows.
- The predictive model is a starting point — it gives us trendline-level insight, and with more features (like sentiment strength, message content, etc.), it could be refined further.

Visualizations & What They Show

The following charts and plots were generated to support and summarize the analysis:

1. **Sentiment Distribution Bar Chart**

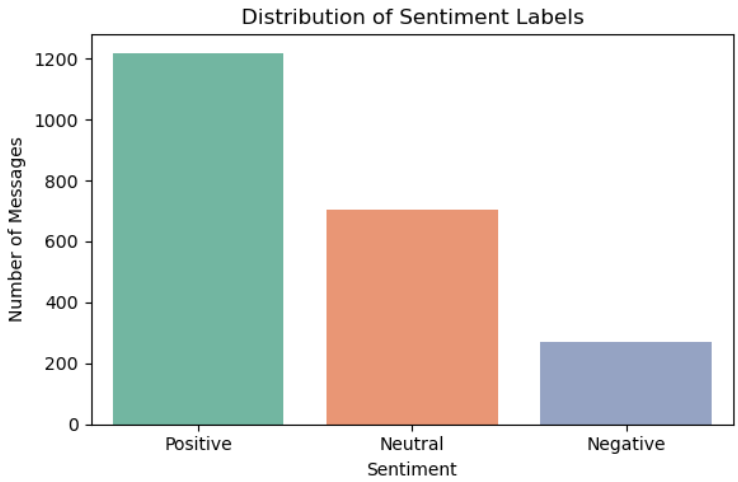

Figure 1: Sentiment Distribution Bar Chart

This clearly shows the dominance of positive sentiment, supporting the conclusion that morale is generally high. However, the presence of nearly 300 negative messages also signals a need to dig deeper — exactly what we did in the Flight Risk task.

2. **Sentiment Trends Over Time**

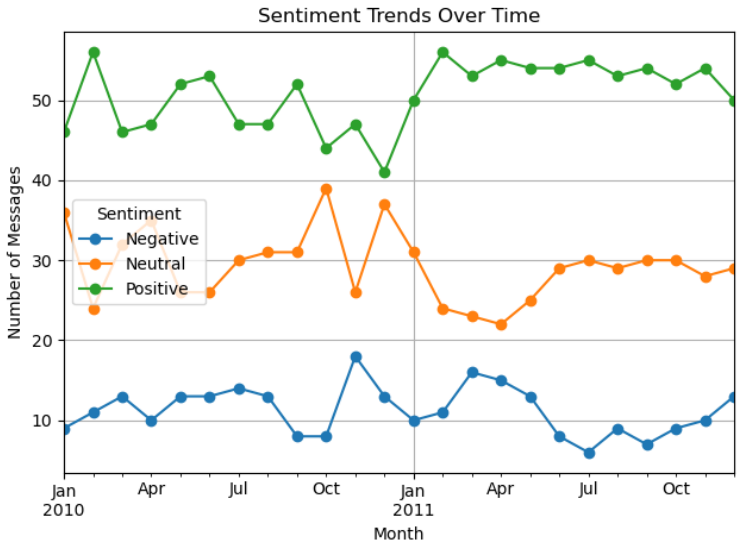

Figure 2: Sentiment Trends Over Time

This line plot shows how positive sentiment remained steady over time, while neutral messages slightly declined. Negative messages were relatively stable, but we saw minor spikes in certain months — reinforcing our decision to use rolling windows for Flight Risk detection.

3. **Top 10 Most Active Employees Bar Chart**

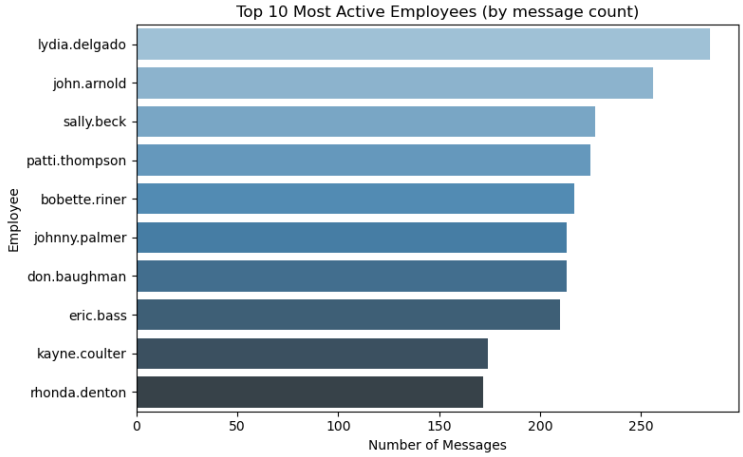

Figure 3: Top 10 Most Active Employees Bar Chart

This helped identify who contributed the most to communication volume. Interestingly, many top active employees also showed up in the Flight Risk list — suggesting that higher message frequency doesn't always correlate with positivity.

4. **Actual vs Predicted Sentiment Score Plot**

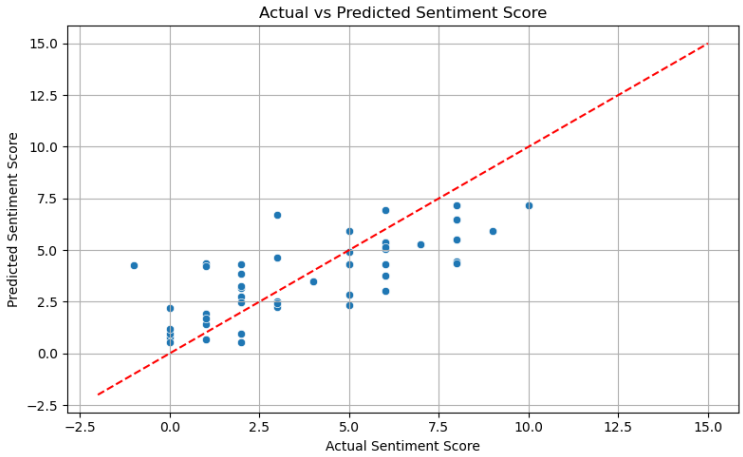

Figure 4: Actual vs Predicted Sentiment Score Plot

This scatter plot visualizes the linear regression model's performance. While not perfect, it shows a clear positive correlation, supporting the idea that engagement (message count) and time progression can be used to anticipate sentiment.