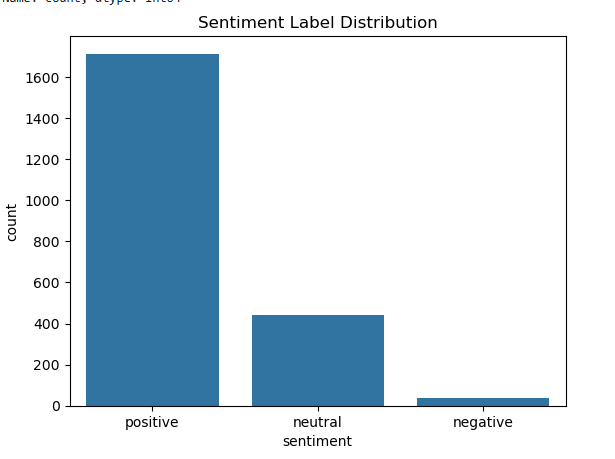
**Approach and Methodology**

Here are some key steps in my appoarch to this sentiment labeling assessment:

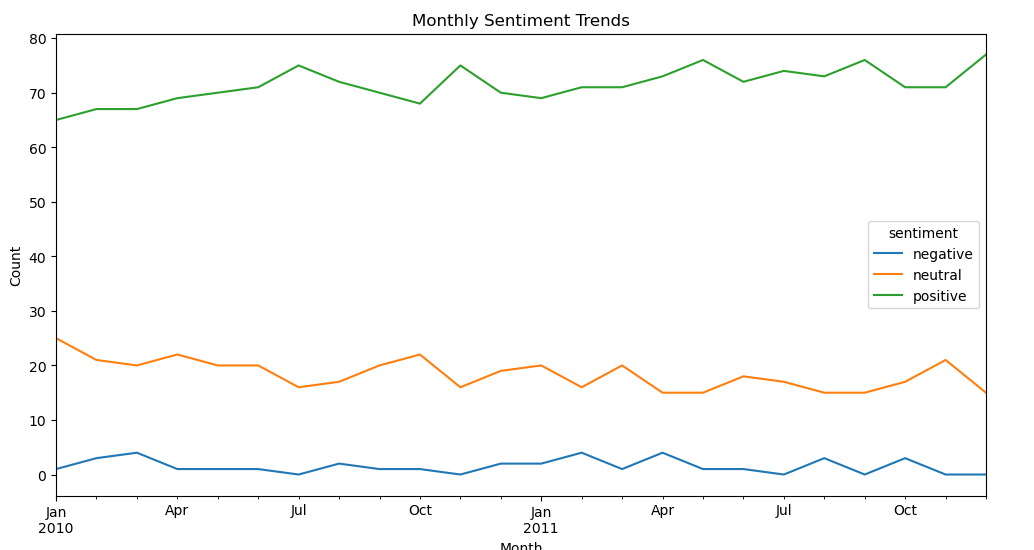
* **TF-IDF**: I used the body of emails, converted them to lowercase, and vectorized using **TF-IDF** to prepare for modeling.
* **Sentiment Labeling**: Employed the VADER sentiment analyzer, which is optimized for social text, to label messages as Positive, Neutral, or Negative.
* **Logistic Regression**: I trained a logistic regression model to complete the sentiment labeling for all the messages based on the message’s body, email address and title.
* **Feature Engineering**: Additional features such as from\_freq (email frequency by sender), message length, word count, and average length were calculated.
* **Predictive Modeling**: A linear regression model was built to predict sentiment scores using behavioral and content-based features.

**Key Findings from EDA**

****

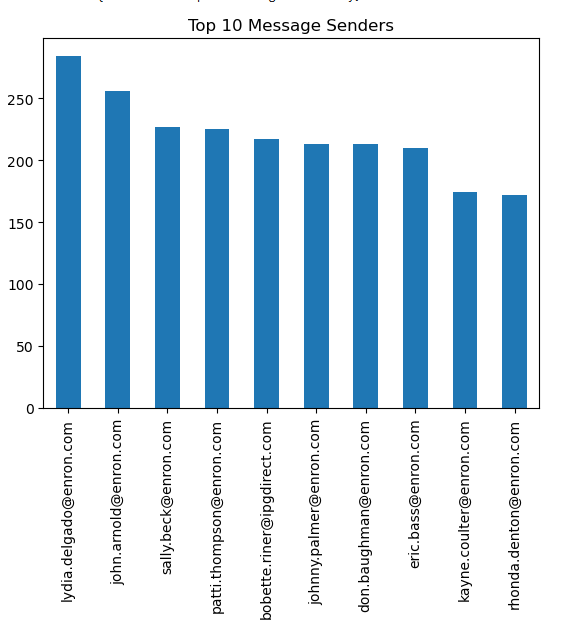
**Sentiment Label Distribution**

The graph reveals that the majority of employee messages(over 1,600) are classified as positive, indicating an overall optimistic sentiment within the communication data. This is followed by approximately 500 messages categorized as neutral. In contrast, only a small fraction of messages are labeled as negative. This distribution suggests a positive workplace atmosphere. Although the presence of any negative sentiment, may need further attention to identify underlying concerns or emerging issues.



**Monthly Sentiment Trends**

The graph indicates a slight upward trend in the volume of positive messages over time, while neutral messages show a very modest decline. Negative messages remain relatively constant throughout the observed period. Overall, sentiment trends appear stable across 2010 and 2011, suggesting consistent communication patterns and no significant shifts in employee sentiment during this timeframe.

 **A graph with blue lines

AI-generated content may be incorrect.**

**Top 10 message senders and average sentiment**

These two graphs show the top 10 message senders and average sentiment by employees.

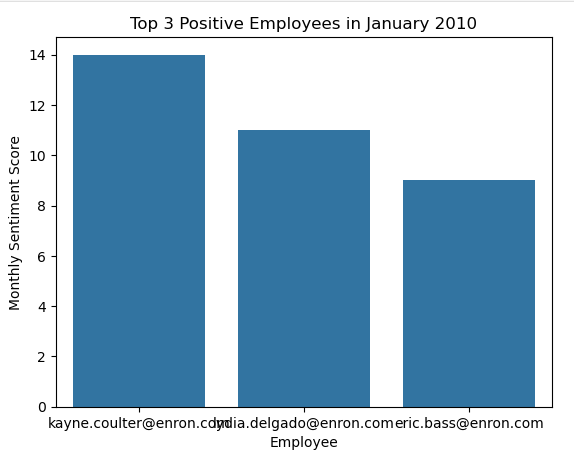
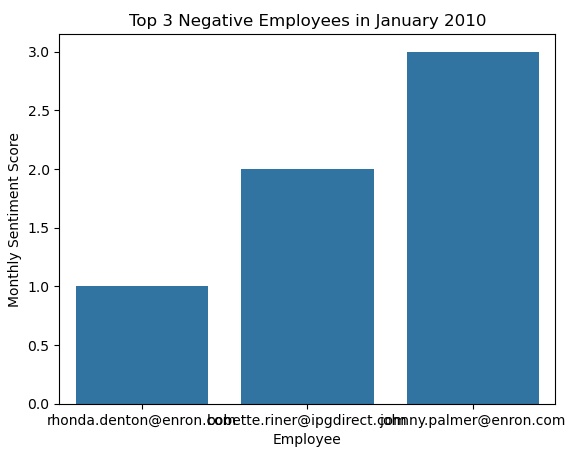
**Employee Scoring and Ranking Processes**

To quantify employee sentiment over time, each email message was assigned a sentiment score based on its label. These individual message scores were then aggregated on a monthly basis per employee to produce a cumulative monthly sentiment score. This aggregation allows us to track how each employee’s sentiment evolves over time, normalized by month.

Here are the steps and processes:

1. I began by defining a dictionary to map sentiment labels to numerical scores: assigning +1 for positive, 0 for neutral, and –1 for negative sentiments. Using this mapping, I created a new column in the DataFrame titled sentiment\_score, which captures the sentiment score for each message based on its label.
2. With the sentiment scores assigned, I computed each employee’s monthly sentiment score by grouping the data by employee email addresses and extracting the month from the message timestamps. This enabled aggregation of sentiment scores on a monthly basis.
3. To rank employees based on sentiment, I sorted the resulting monthly scores in descending order to identify the top positive contributors. Similarly, I sorted in ascending order to determine those with the most negative sentiment scores for each month. And then only extract the top three employee score by using head(3) function.

This approach provides an interpretable view of which employees are consistently expressing strong positive or negative sentiments, potentially indicating satisfaction or disengagement, respectively.

**Flight Risk Identification Criteria and Outcomes**

The criteria for identifying flight risk is by identifying if employee send 4 or more negative emails within any 30-day window, regardless of the sentiment score. Here is my implementation:

* + Filtered all emails with sentiment == 'negative'.
  + Grouped messages by employee and checked rolling 30-day counts.
  + Flagged employees meeting the threshold count of >= 4

**Outcomes**

From my implementions, I did not find any result of flight risk employee that has at least 4 negative messages with the same email address. So I adjusted the criteria to be at least 3 negative messages, I found 1 employee that may have risk of leaving.

**Predictive Modeling Overview**

For the predictive modeling task, a Linear Regression model was employed to analyze and forecast sentiment scores based on various behavioral features.

The selected features included:

* from\_freq (the number of emails sent)
* word\_count
* message\_length
* the month of the message

**Evaluation**

Key evaluation metrics used:

* R² Score
* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)

The final evaluation shows that the predicting values close to actuals, but it doesn't explain much of the overall variation in the data because of a weak R Squared score. So it works okay for point prediction, but not as well for understanding or generalizing the data’s behavior. The model might be underfitting resulting in a relatively lower R squared.

A graph with purple dots

AI-generated content may be incorrect. 