Employee Sentiment Analysis Report

This project involves analyzing an unlabeled dataset of employee messages to assess sentiment and engagement. The task is to work from raw data and derive insights using natural language processing (NLP) and statistical analysis techniques. The project is divided into several distinct tasks, each focusing on a different aspect of data analysis and model development.

# Approach and Methodology

This section describes the approach taken in the analysis. The methodology involves setup and data collection, and specific tasks including: sentiment labeling, exploratory data analysis (EDA), sentiment scoring of employee emails, ranking employees based on sentiment scores, identifying flight risk, and building a predictive model. In the setup and data collection phase, a module was created to automate the dataset file downloading process, but due to authentication access, it was unsuccessful. Future works will try to fix this module.

The dataset includes 2191 records with 4 columns. The “Subject” column contains the subject of each email, the “body” column contains the message, “date” contains the date sent and “from” contains the sender.

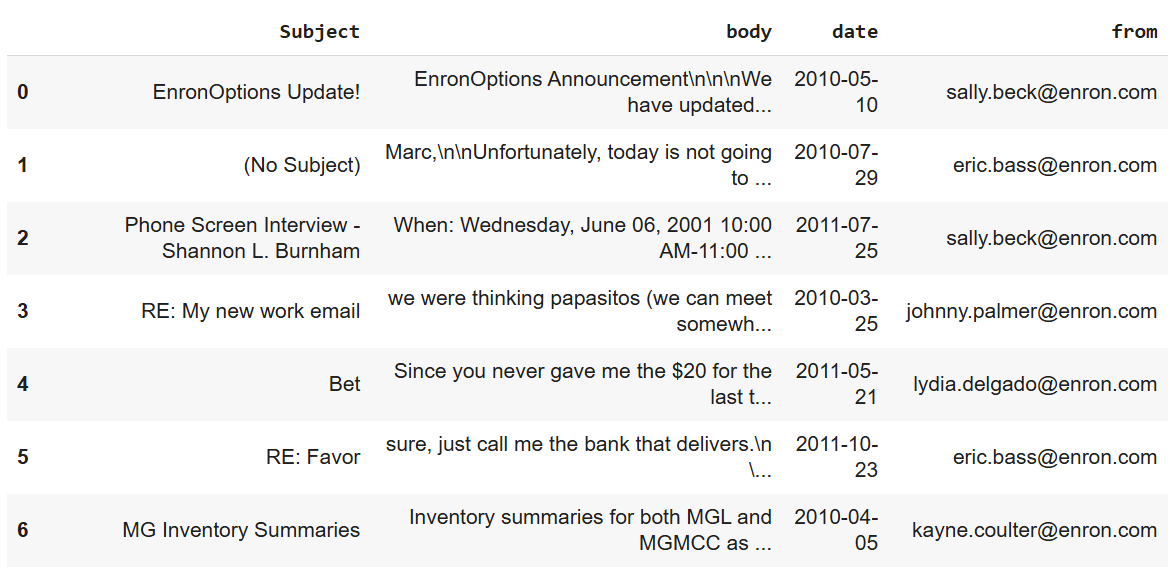


Figure 1: Overview of the dataset

The methodology, results and key findings of each task will be shown below.

## Sentiment Labeling

**Objective:** Label each employee message with one of three sentiment categories: Positive, Negative, or Neutral.

**Methodology**: Use an LLM to label the “body” column of the DataFrame as Positive, Negative, or Neutral, and add the labels as a new column to the DataFrame.

* The transformers library was successfully installed, and a sentiment analysis pipeline using the distilbert-base-uncased-finetuned-sst-2-english model was set up.
* A Python function *get\_sentiment* was created to utilize the sentiment analysis pipeline and return sentiment labels ('Positive', 'Negative', or 'Neutral') for input text.
* The *get\_sentiment* function was applied to the 'body' column of the DataFrame, generating a series of sentiment labels.
* A new column named 'sentiment' was added to the DataFrame, and it was populated with the generated sentiment labels.
* A manual review of a random sample of 20 rows indicated that the LLM's sentiment labels generally aligned with the content, and the accuracy was deemed acceptable for this general sentiment analysis task.

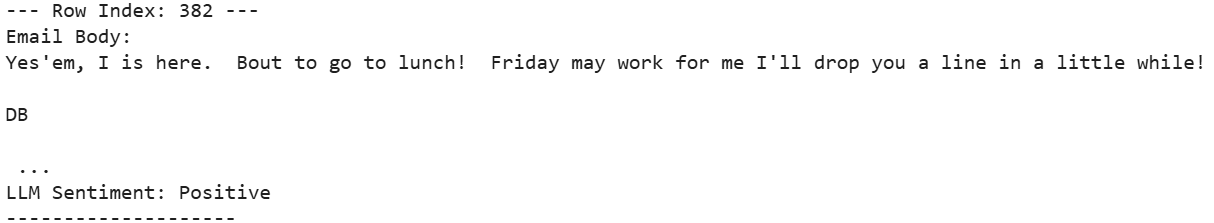


Figure 2: Sample of a record being labeled “Positive”

## Exploratory Data Analysis (EDA)

**Objective:** Understand the structure, distribution, and trends in the dataset through thorough exploration.

**Methodology**: Perform exploratory data analysis (EDA) on the provided DataFrame, focusing on data structure, sentiment distribution, and trends over time. Include visualizations to effectively communicate findings.

* Check the number of records, data types, and identify missing values in the DataFrame: The dataset contains 2191 records with no missing values in the analyzed columns (“Subject”, “body”, “date”, “from”, “sentiment”).
* Analyze sentiment distribution: The dataset is imbalanced, there are 270 records being labeled 'Negative', 1218 being labeled 'Positive' and 703 labeled 'Neutral'. The sentiment analysis reveals a predominance of negative sentiment in the emails.

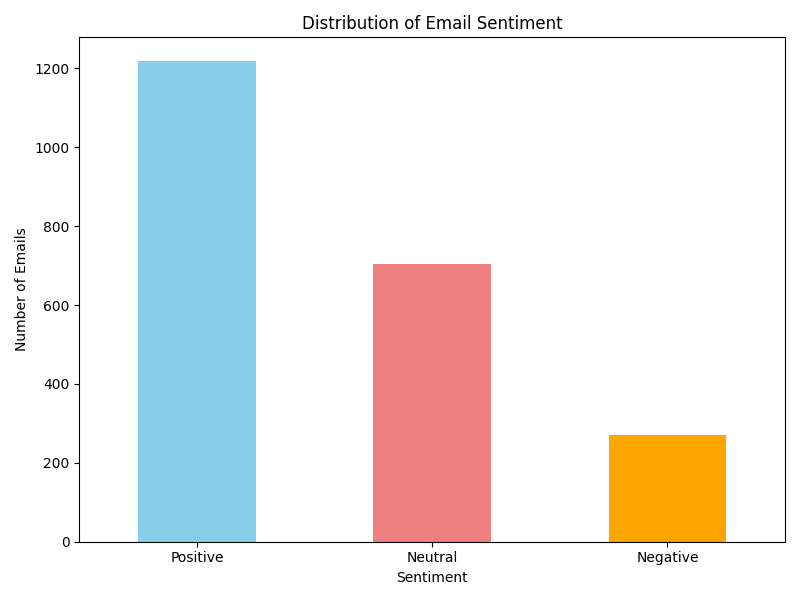


Figure 3: Distribution of Email Sentiment

* Analyze trends overtime: Sentiment distribution over time shows fluctuations in the number of emails per sentiment category across different years. The number of emails labeled 'Negative' and 'Neutral ' decreased slightly while that of 'Positive' increased.

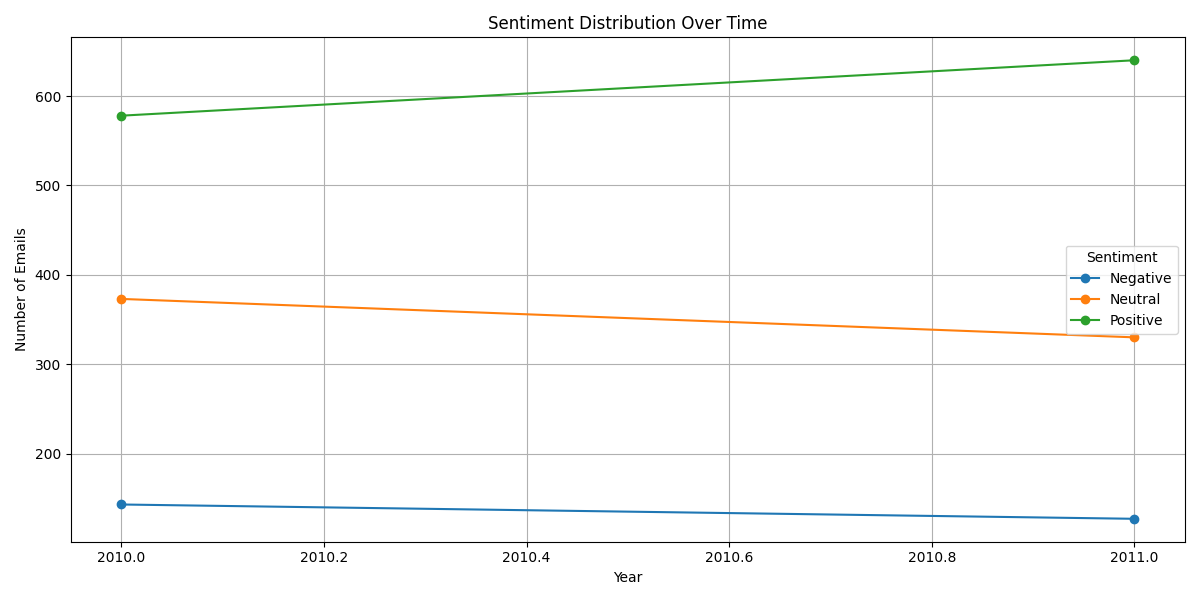


Figure 4: Sentiment Distribution Over time

* Analysis of frequent senders: identified individuals like Lydia Delgado, John Arnold, and Sally Beck as top contributors by email volume.

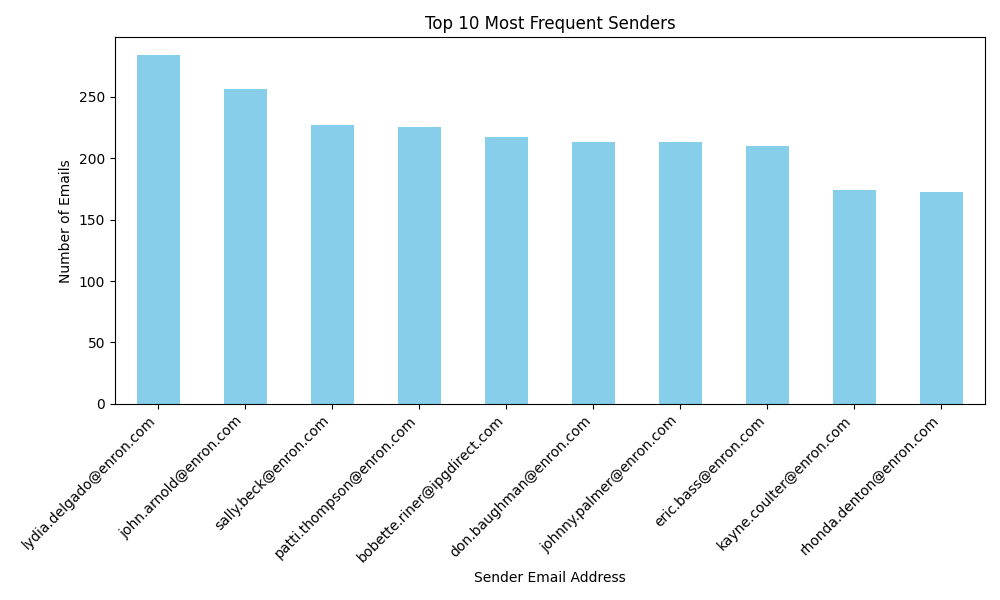


Figure 5: Top 10 most frequent senders

* Frequent subjects for negative sentiment include "(No Subject)" and "Re:", along with specific topics like "TradersNews".

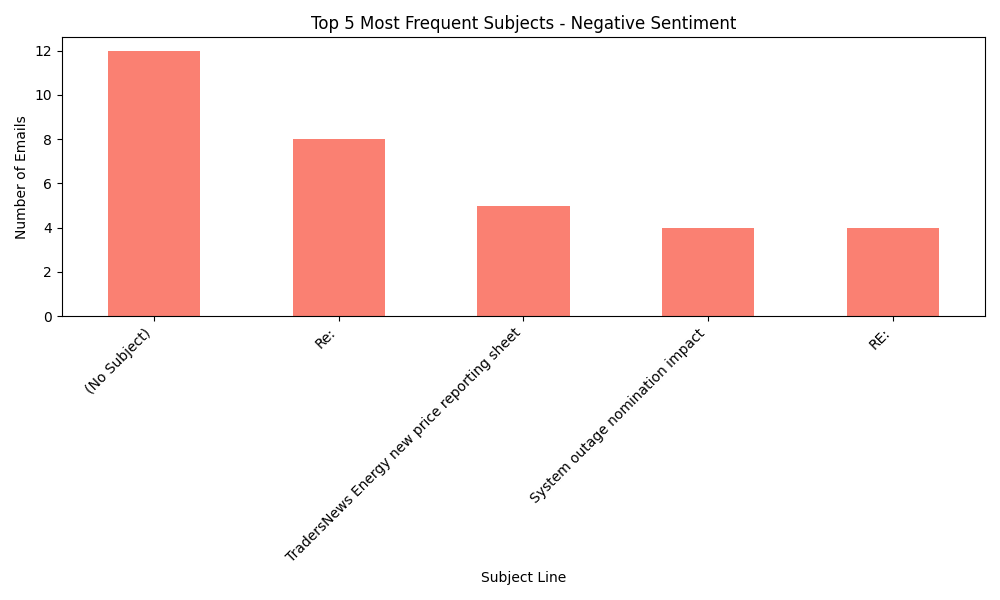


Figure 6: Top 5 most frequent subjects for emails with negative sentiment

* Frequent subjects for positive sentiment also include "(No Subject)" and "Re:", but also positive indicators like "Interview Schedule".

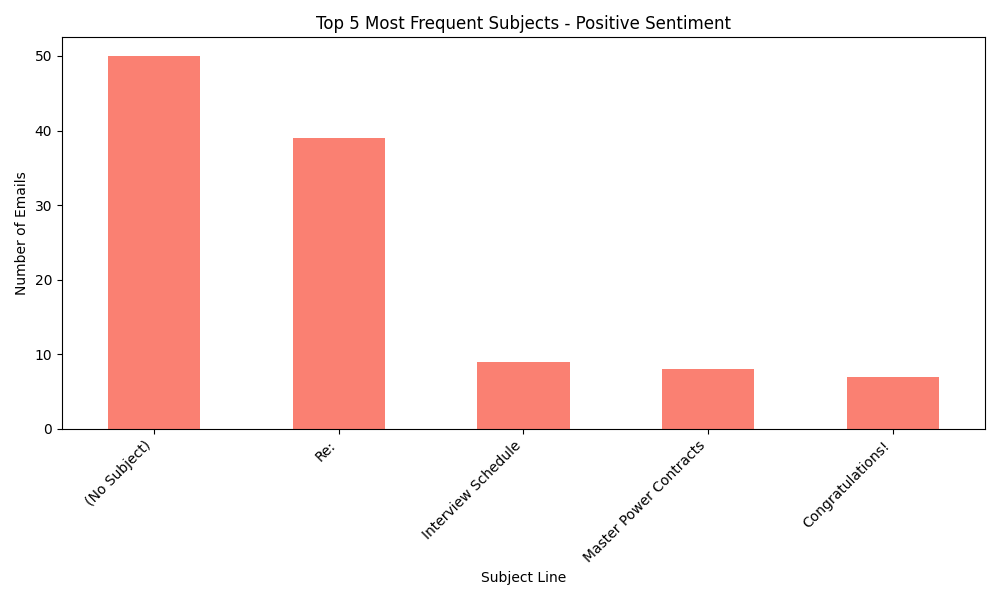


Figure 7: Top 5 most frequent subjects for emails with positive sentiment

* Frequent subjects for neutral sentiment also include "(No Subject)" and "Re:"

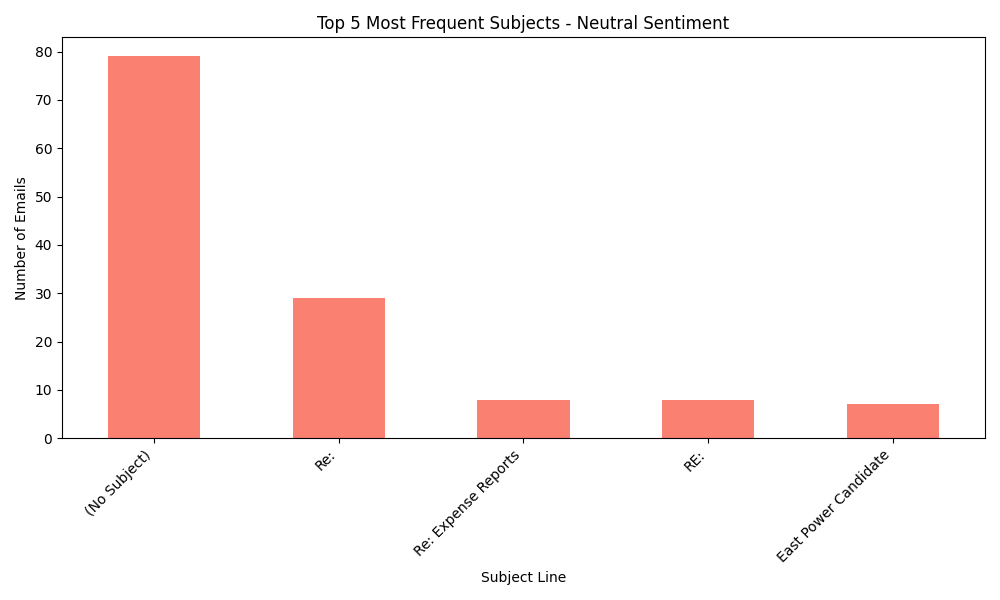


Figure 8: Top 5 most frequent subjects for emails with neutral sentiment

## Employee Score Calculation

**Objective:** Compute a monthly sentiment score for each employee based on their messages.

**Methodology**: Calculate a monthly sentiment score for each employee based on their messages, assigning +1 for Positive, -1 for Negative, and 0 for Neutral sentiment, and aggregate these scores on a monthly basis for each employee, ensuring the score resets at the beginning of each new month.

* Sentiment Score Assignment: A numerical score was assigned to each email based on its sentiment label derived from the LLM analysis. Positive sentiment emails were assigned a score of +1. Negative sentiment emails were assigned a score of -1. Neutral sentiment emails were assigned a score of 0. This score was added as a new column named *sentiment\_score* to the DataFrame.
* Extracting Month and Year: To facilitate monthly aggregation, the year and month were extracted from the date column and added as new columns, year and month, respectively.
* Grouping and Aggregation: The DataFrame was then grouped by the employee's email address (from), the year (year), and the month (month). Within each of these groups, the *sentiment\_score* was summed. This sum represents the total sentiment score for a given employee in a specific month and year.
* Monthly Score Reset: The grouping by year and month inherently ensures that the sentiment score is calculated independently for each month. As a result, the accumulated sentiment score for an employee naturally resets at the beginning of each new month and year combination.

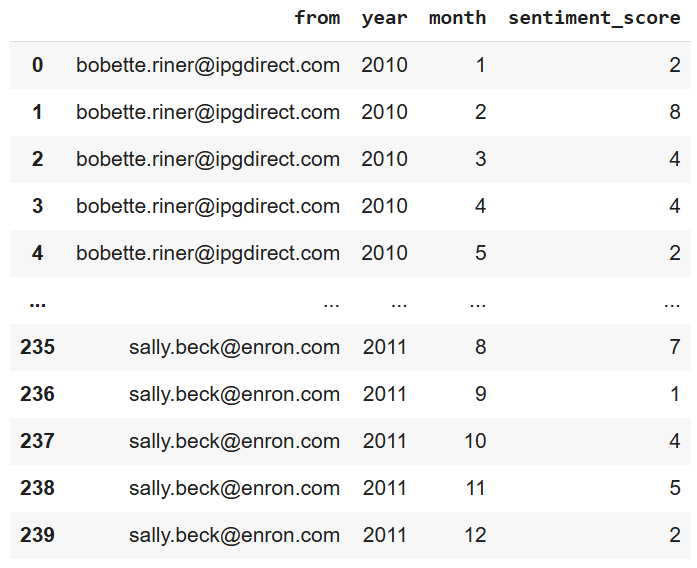


Figure 9: Final DataFrame calculating the monthly sentiment score of each employee

## Employee Ranking

**Objective:** Generate ranked lists of employees based on their monthly sentiment scores.

**Methodology**: Rank employees based on their monthly sentiment scores, identifying the top three positive and top three negative employees for each month, and present the rankings.

* **Monthly Sentiment Score Calculation:** As detailed in the "Employee Score Calculation" section, a numerical sentiment score was assigned to each email (+1 for Positive, -1 for Negative, 0 for Neutral). These individual scores were then aggregated by employee, year, and month to obtain a single monthly sentiment score for each employee.
* **Sorting for Ranking:** The DataFrame containing the monthly sentiment scores was sorted. The primary sorting was by year and then by month in ascending order. Within each month, the scores were sorted in descending order, placing the highest (most positive) sentiment scores at the top and the lowest (most negative) sentiment scores at the bottom.
* **Identifying Top Positive Employees:** For each month, the top three employees with the highest positive sentiment scores (scores greater than 0) were identified from the sorted data.

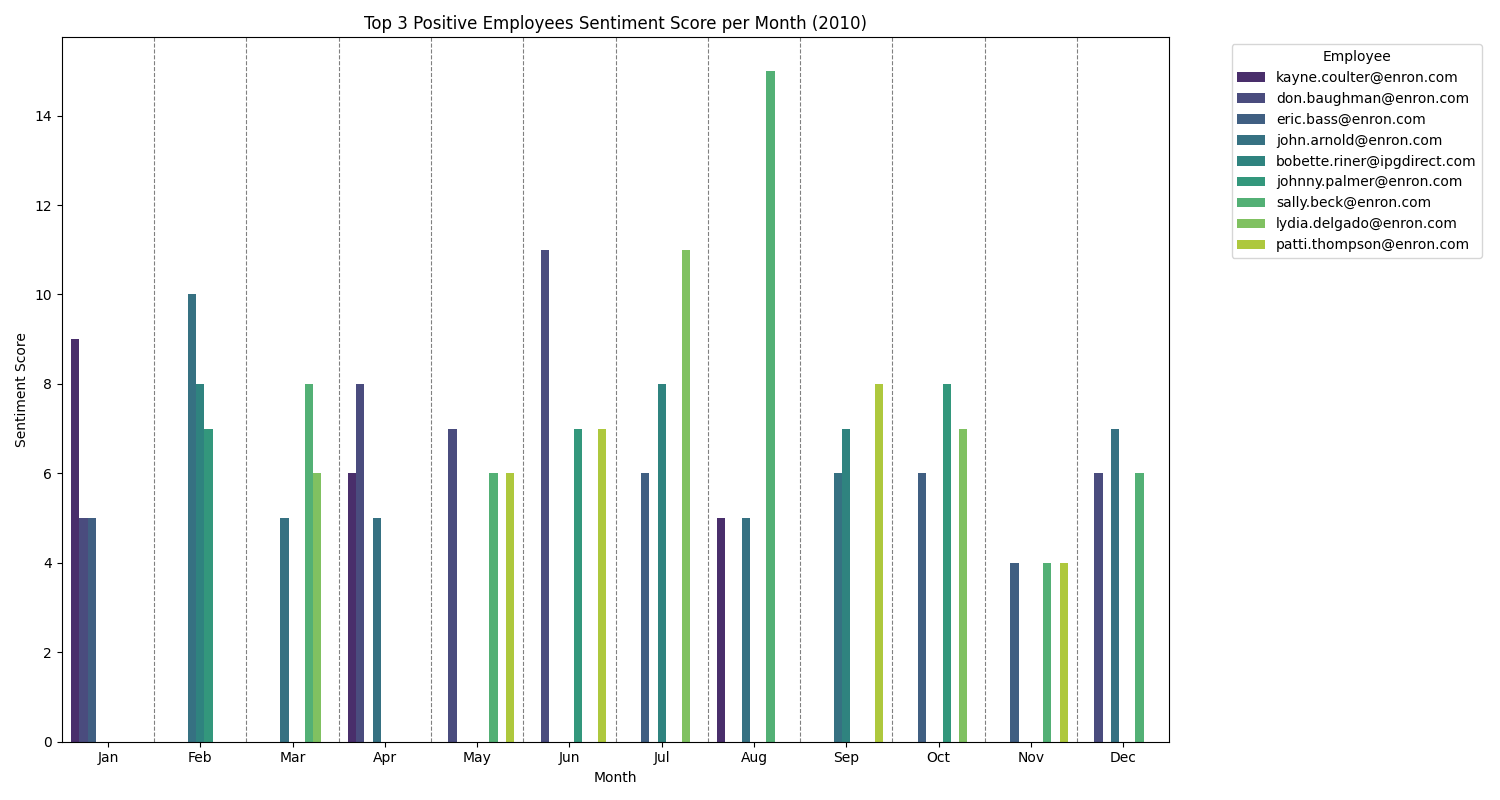


Figure 10: Top 3 employees with positive sentiment score of each month in 2010

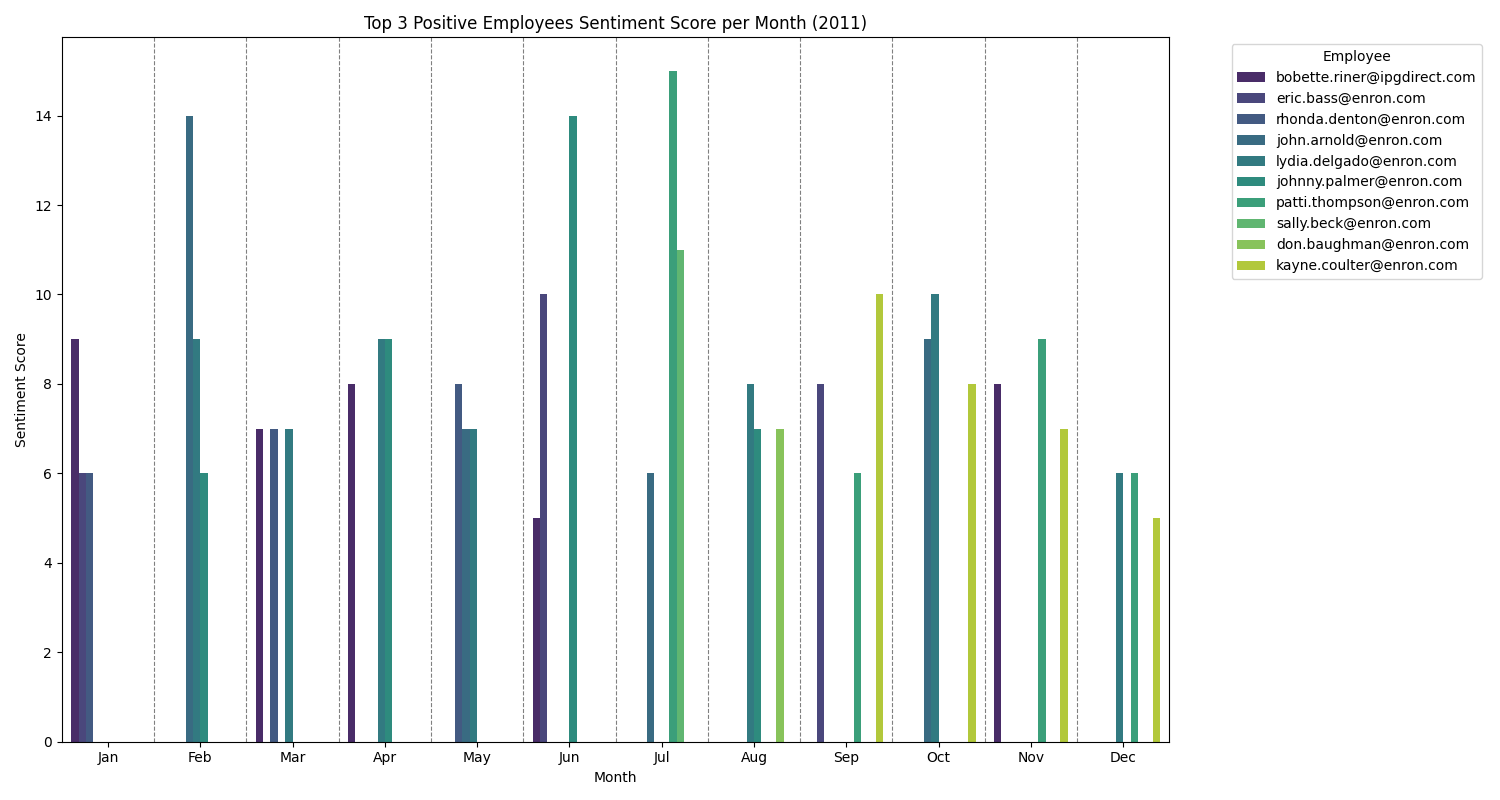


Figure 11: Top 3 employees with positive sentiment score of each month in 2011

* **Identifying Top Negative Employees:** For each month, the top three employees with the lowest (most negative) sentiment scores (scores less than 0) were identified from the sorted data.

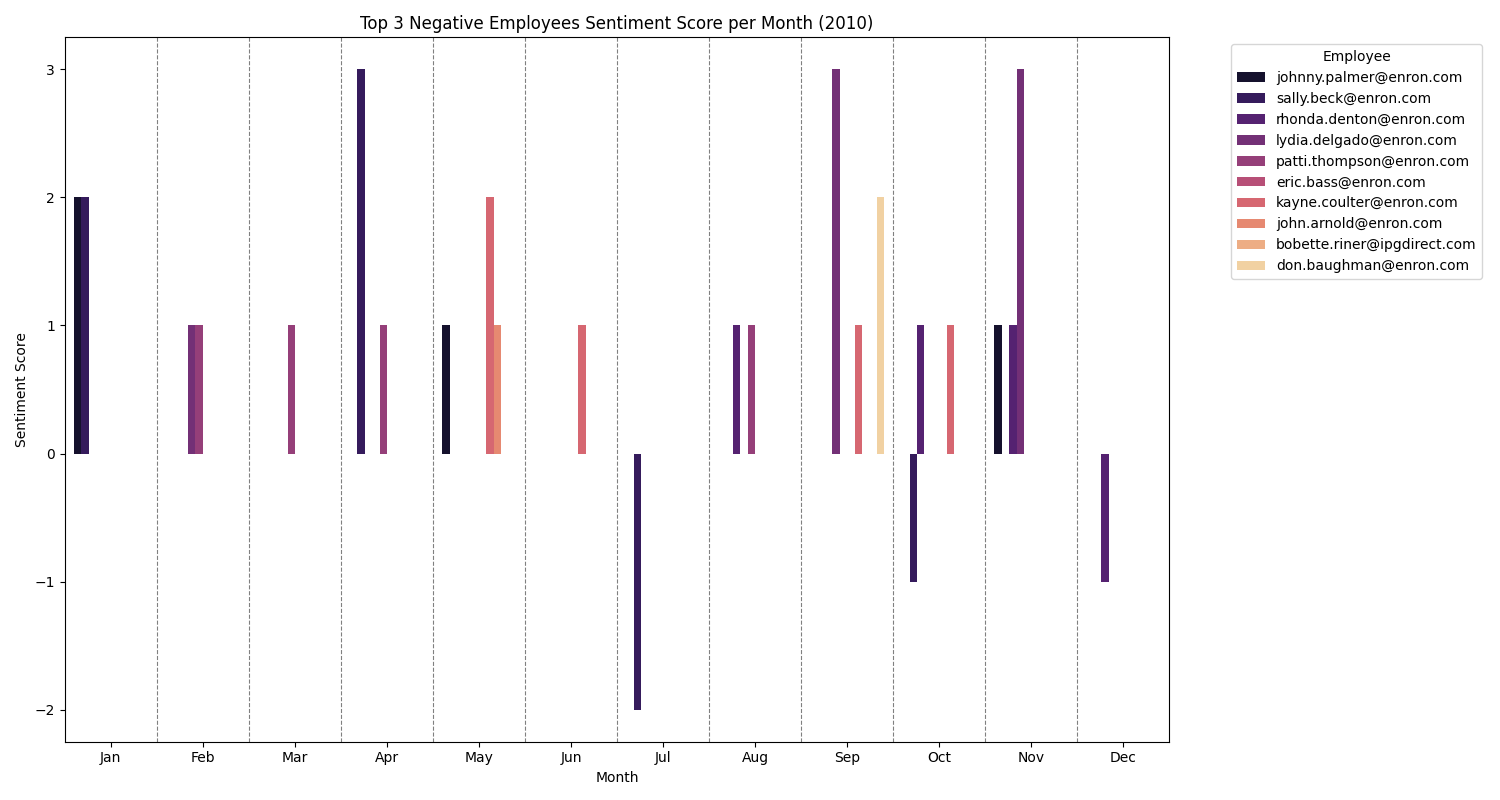


Figure 12: Top 3 employees with negative sentiment score of each month in 2010

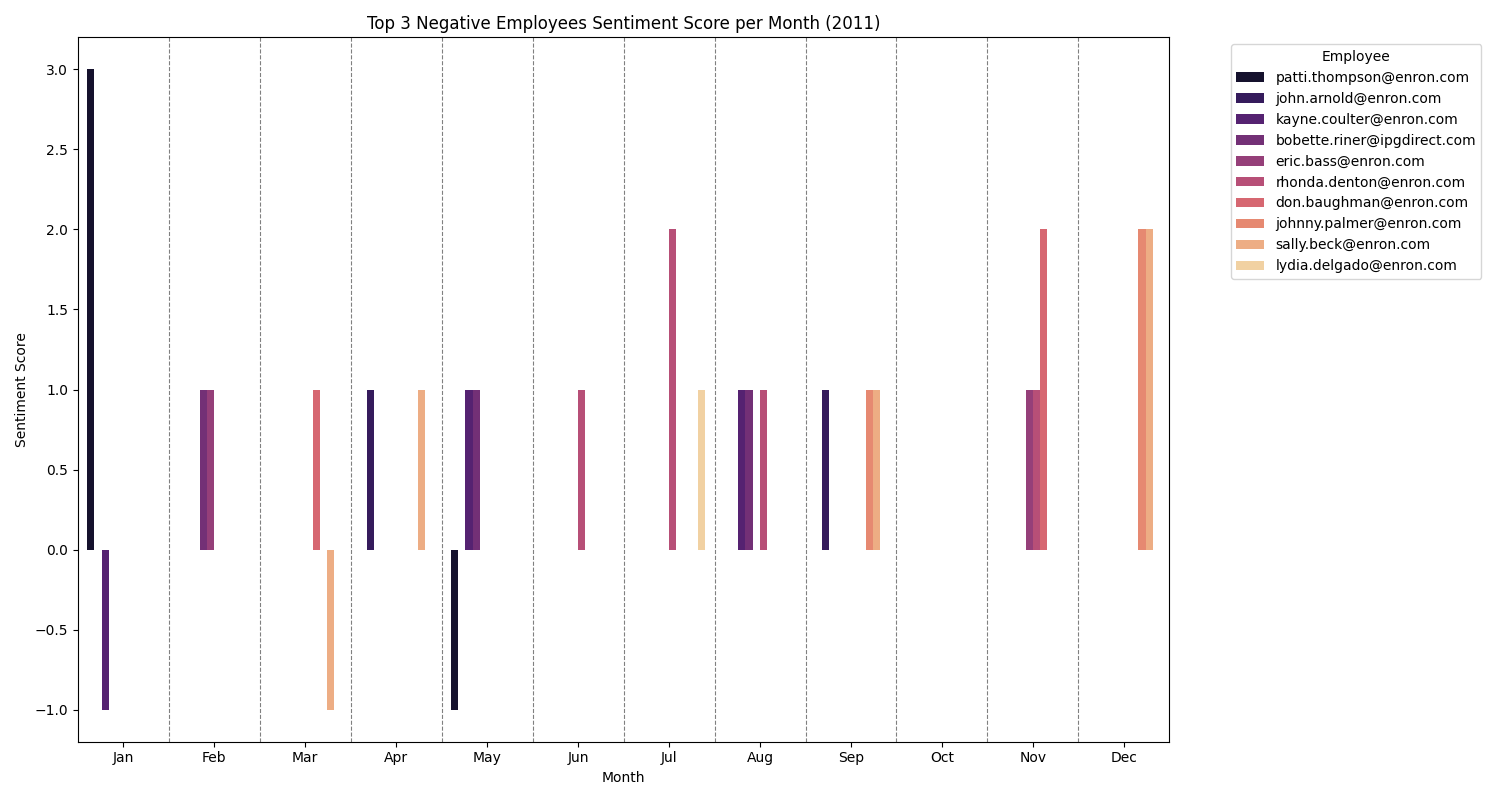


Figure 13: Top 3 employees with negative sentiment score of each month in 2011

This process ensured that the rankings for each month were independent and reflected the sentiment performance of employees within that specific month.

## Flight Risk Identification

**Objective:** Identify employees who are at risk of leaving based on their monthly sentiment scores. This task is critical for identifying potential issues in employee engagement and retention.

**Methodology:** Identify employees who are at risk of leaving based on their monthly sentiment scores. A Flight risk is any employee who has sent 4 or more negative mails in the span of 30 days (irrespective of the score). The 30-day period is rolling count of days, irrepective of months. Extract a list of these employees. Ensure that this flagging process is robust.

* Negative Email Sentiment: Create another dataframe containing only emails classified with a 'Negative' sentiment were considered for this analysis.
* Rolling 30-Day Window: A rolling 30-day window was applied to each employee's negative emails. For each negative email sent, the analysis looked back 30 days from the date of that email.
* Threshold for Negative Emails: An employee was flagged as at risk if they sent 4 or more negative emails within any of these rolling 30-day windows.
* Calculating Rolling Negative Count: The *rolling\_negative\_count* was calculated by first sorting each employee's negative emails by date. Then, for each email, the number of negative emails sent by that same employee in the preceding 30 days was counted.
* Identifying At-Risk Employees: Any employee who, at any point in time, had a *rolling\_negative\_count* of 4 or more was identified as an at-risk employee. The final list of at-risk employees includes all unique individuals who met this criterion at least once.

## Predictive Modeling

**Objective:** Develop a linear regression model to analyze sentiment trends and predict sentiment scores using a variety of independent variables that may influence sentiment scores. Due to the ambiguity the objective of the task using sentiment scores, whether of each email or the monthly sentiment score), we create two different models for both cases.

**Methodology:** Develop two different linear regression models to analyze sentiment trends and predict sentiment scores in two cases using a variety of independent variables that may influence sentiment scores such as message length, word count, average word count, etc.

**Case 1: Develop a linear regression model to analyze sentiment trends and predict sentiment score of each message**

* **Model Development Process:**

Feature Selection: The following features were selected as independent variables based on the assumption that they might influence the sentiment expressed in emails:

* + message\_length: The number of characters in the email body.
  + word\_count: The number of words in the email body.
  + monthly\_message\_count: The total number of emails sent by an employee within a specific month. These features represent quantifiable aspects of communication volume and verbosity, which could potentially correlate with sentiment.

Data Splitting: The dataset was split into training and testing sets. A 75/25 ratio was used, with 75% of the data allocated for training the model (X\_train, y\_train) and 25% for evaluating its performance on unseen data (X\_test, y\_test). This split helps in assessing the model's generalization ability.

Model Used: A Linear Regression model from the sklearn.linear\_model library was used. Linear regression is a simple yet powerful algorithm that models the linear relationship between independent variables and a dependent variable.

* **Model Evaluation Results:**

The trained linear regression model was evaluated on the testing set using the following metrics:

* The Mean Squared Error (MSE) of the model on the test set is approximately 0.452.
* The Root Mean Squared Error (RMSE) of the model on the test set is approximately 0.6726.
* The R-squared (R2) value of the model on the test set is approximately 0.07503, indicating that about 7.5% of the variance in sentiment scores is explained by the model.
* **Model Coefficients and Intercept:**
  + Intercept: The intercept of the model is approximately 0.2870. The intercept represents the predicted sentiment score when all independent variables (message length, word count, and monthly message count) are zero. However, interpreting this intercept in isolation is not meaningful in this context, as it is unrealistic for all features to be zero. It mainly serves as the baseline offset for the regression line. Coefficient Interpretation:
* message\_length Coefficient**:** The coefficient message\_length for is -0.00015. A one-unit increase in message length, while holding other variables constant, is associated with a decrease of approximately 0.00015 in the predicted sentiment score. This suggests a very slight negative relationship, meaning longer messages may be marginally associated with more negative sentiment.
* word\_count Coefficient: The coefficient for word\_count is approximately 0.00577. A one-unit increase in word count, holding other variables constant, is associated with an increase of about 0.00577 in the predicted sentiment score. This indicates a positive relationship: messages with more words tend to be slightly more positive in sentiment.
* monthly\_message\_count Coefficient:  The coefficient for monthly\_message\_count is approximately -0.00383. A one-unit increase in monthly message count, holding other variables constant, is associated with a decrease of about 0.00383 in the predicted sentiment score. This shows a negative relationship, suggesting employees who send more messages per month may show slightly more negative sentiment.

The linear regression model trained on the selected features provides some limited insight into the factors influencing email sentiment. While the coefficients suggest weak linear relationships between the features and sentiment, the low R-squared value indicates that these features alone are not strong predictors of sentiment. A significant portion of the variability in email sentiment remains unexplained by this model. This suggests that other factors not included in this analysis, or potentially non-linear relationships, play a more dominant role in determining email sentiment.

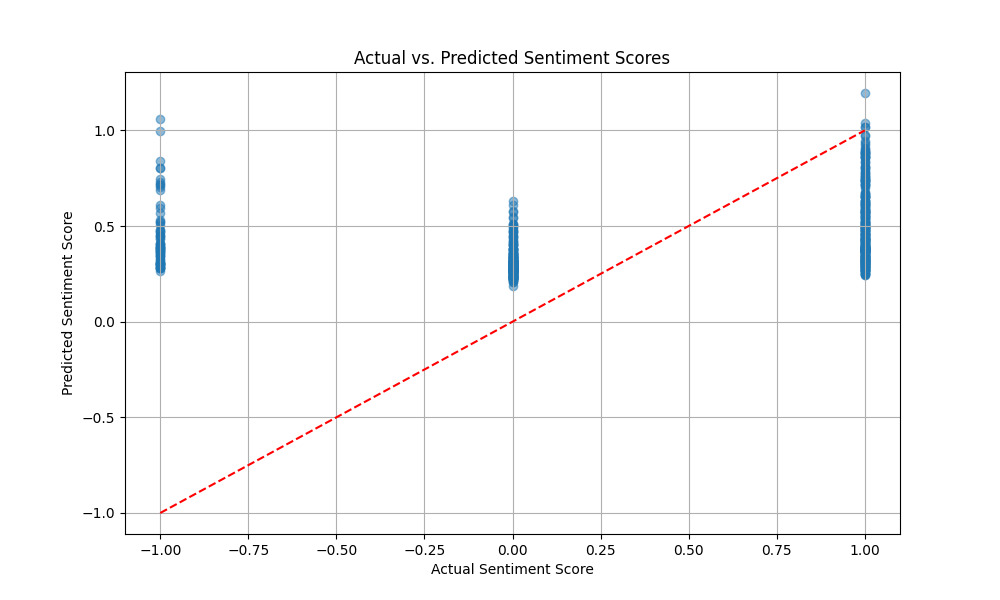


Figure 14: Plot depicting Actual Sentiment Score and Predicted Sentiment Score

The points are clustered at the discrete sentiment values (-1, 0, 1) on the x-axis. Predictions are mostly around 0.3–0.6 regardless of the actual class, showing the model struggles to differentiate between the three categories. The lack of alignment along the red diagonal indicates poor predictive performance and systematic bias (underpredicting positive and overpredicting negative values).

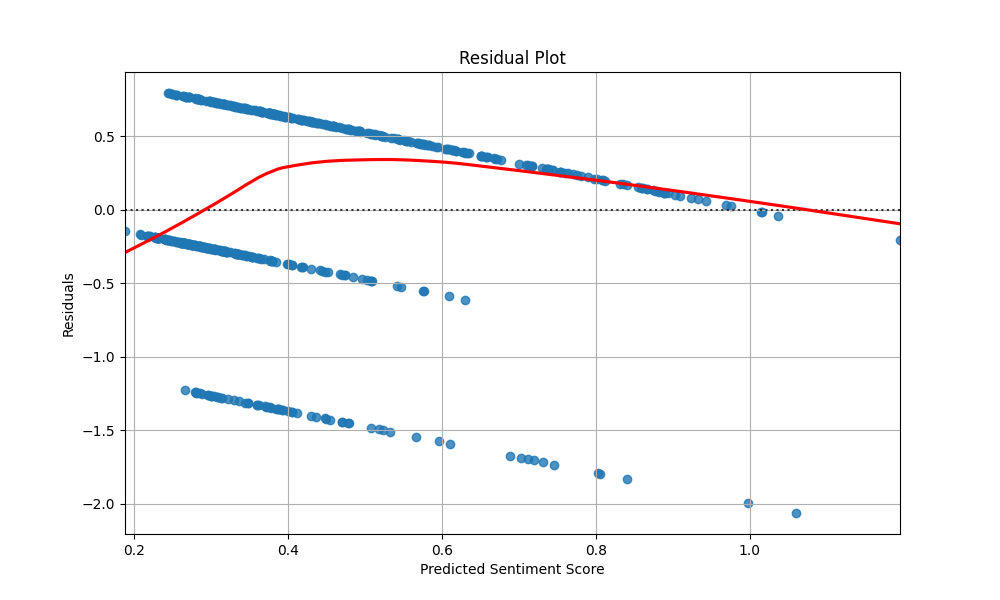


Figure 15: Residual plot

This plot shows residuals (y-axis) against predicted sentiment scores (x-axis), with a smoothed trend line in red. There are clear patterns and clusters in the residuals, rather than a random scatter around zero. Residuals become more negative as predicted scores increase, indicating heteroscedasticity (non-constant variance) and systematic prediction errors. This violates the linear regression assumption that residuals should be randomly distributed and suggests the model is mis-specified or unsuitable for this data.

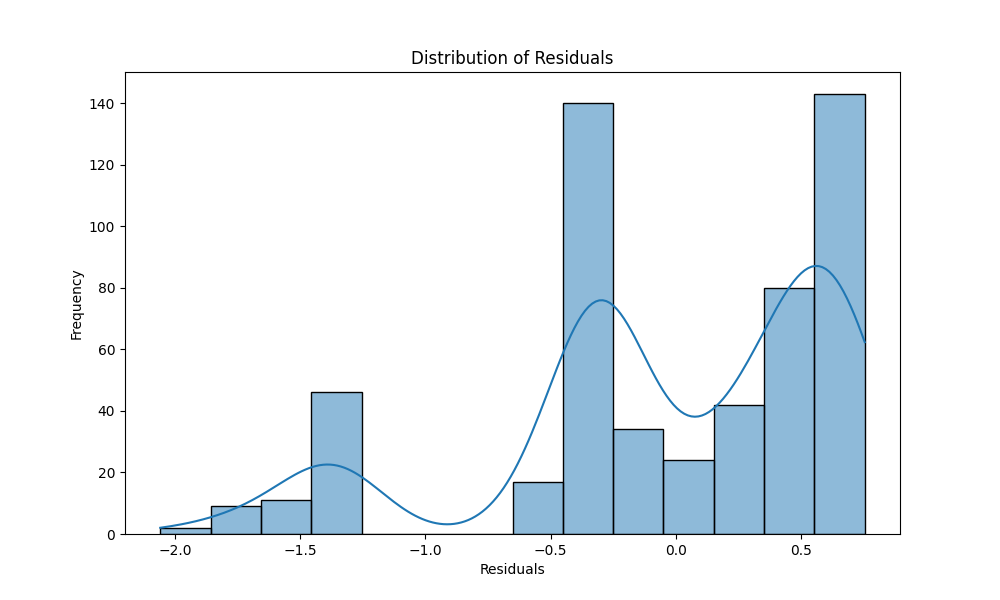


Figure 16: Distribution of Residuals

This histogram shows the distribution of residuals (actual - predicted values) with a kernel density overlay. Residuals are spread across both positive and negative values, showing systematic errors in both directions. The distribution is not centered around zero and is multimodal, which violates a key assumption of linear regression (that residuals should be normally distributed with mean ≈ 0).

This suggests that the model is not capturing the underlying data structure well and is making biased predictions. Based on the plots and the low R-squared value, the linear regression model developed with these specific features does not appear to be suitable for accurately predicting sentiment scores or analyzing sentiment trends effectively. The model's performance is quite poor, and a significant portion of the sentiment variability remains unexplained. The fundamental purpose of a linear regression model is to predict a continuous numerical output. The sentiment scores, -1 and 1, are not a continuous range. They are discrete, distinct categories. This type of problem is known as a classification problem, not a regression problem.

**Case 2: Develop a linear regression model to predict monthly sentiment scores using monthly aggregated features.**

* **Model Development Process:**

Feature Selection: Based on the monthly aggregated data, the following features were selected as independent variables to predict the monthly\_sentiment\_score:

* + monthly\_message\_count: The total number of emails sent by an employee in a given month.
  + average\_message\_length: The average character length of emails sent by an employee in a given month.
  + average\_word\_count: The average number of words in emails sent by an employee in a given month. These features were chosen as they represent quantifiable aspects of monthly communication volume and verbosity that might correlate with overall monthly sentiment.
  + average\_polarity: The average polarity score of emails sent by an employee in a given month.
  + average\_subjectivity: The average subjectivity score of emails sent by an employee in a given month.

Data Preparation: Before splitting, the selected features and the target variable (monthly\_sentiment\_score) were checked for missing values and appropriate data types. No missing values were found, and the data types (int64 and float64) were suitable for regression.

Data Splitting: The monthly aggregated dataset was split into training and testing sets using the train\_test\_split function from sklearn.model\_selection. A 75/25 ratio was used, with 75% of the data allocated for training and 25% for testing. A random\_state of 42 was set for reproducibility.

Model Used: A LinearRegression model from the sklearn.linear\_model library was chosen for this task. Linear regression models the linear relationship between the independent variables and the dependent variable.

* **Model Evaluation Results:**

The performance of the trained linear regression model was evaluated on the testing set (X\_test, y\_test) using common regression metrics:

* The MSE of 3.3493 represents the average squared difference between the actual and predicted monthly sentiment scores. A lower value indicates better fit.
* The RMSE of 1.8301 provides the average magnitude of the prediction errors in the units of the monthly sentiment score. This means, on average, the model’s predictions are off by about 1.83 units of monthly sentiment score, which is a noticeable but smaller error compared to the previous model version.
* The R2 value of 0.6831 shows that about 68.3% of the variance in monthly sentiment scores is explained by the model. This is a substantial improvement compared to the earlier version and indicates the chosen features collectively provide strong predictive power for monthly sentiment trends.
* **Model Coefficients and Intercept:**
* Intercept: The intercept represents the predicted monthly sentiment score when all independent variables (monthly\_message\_count, average\_message\_length, average\_word\_count , average\_polarity, average\_subjectivity) are zero. Similar to the individual message model, interpreting the intercept in isolation here is not particularly meaningful as these feature values are unlikely to be zero in a realistic monthly summary.
* monthly\_message\_count Coefficient: The coefficient is 0.4098. A one-unit increase in the monthly number of messages, holding other features constant, is associated with an increase of about 0.4098 in the predicted monthly sentiment score. This suggests a moderate positive relationship: months with more messages tend to have a more positive overall sentiment score.
* average\_message\_length Coefficient: The coefficient is 0.00456. A one-unit increase in the average message length (characters), holding other features constant, is associated with an increase of about 0.00456 in the predicted monthly sentiment score. This shows a weak positive relationship between longer average message length and monthly sentiment.
* average\_word\_count Coefficient: The coefficient is 0.0126. A one-unit increase in the average number of words per message, holding other features constant, is associated with a decrease of about 0.0126 in the predicted monthly sentiment score. This indicates a slight negative relationship, meaning more wordy messages might correlate with more negative sentiment overall.
* average\_polarity Coefficient: The coefficient is 8.3078. A one-unit increase in average polarity is associated with a large positive increase (≈ 8.31) in the predicted monthly sentiment score. This shows that message polarity is a very strong positive predictor of monthly sentiment.
* average\_subjectivity Coefficient: The coefficient is 0.4473. A one-unit increase in average subjectivity is associated with a decrease of about 0.4473 in the predicted monthly sentiment score. This implies a moderate negative relationship: months with more subjective content may have slightly more negative overall sentiment scores.

The Linear Regression model for predicting monthly sentiment scores shows stronger performance than the previous version. The inclusion of average\_polarity and average\_subjectivity has substantially improved explanatory power (R² ≈ 0.68), suggesting that message content characteristics are much more informative than purely structural features like message length or count. While monthly\_message\_count and average\_message\_length show weak to moderate positive effects, and average\_word\_count shows a weak negative effect, the strongest relationships come from average\_polarity (strong positive) and average\_subjectivity (moderate negative).

The relatively lower RMSE further supports that the model is producing more accurate monthly predictions overall.

This model indicates that sentiment-oriented features (polarity and subjectivity) are critical for understanding monthly sentiment patterns, whereas structural metrics alone are insufficient. The improved metrics show that the model can reasonably capture the variation in monthly sentiment scores, though further improvements may still be possible with additional non-linear models or other content-based features.

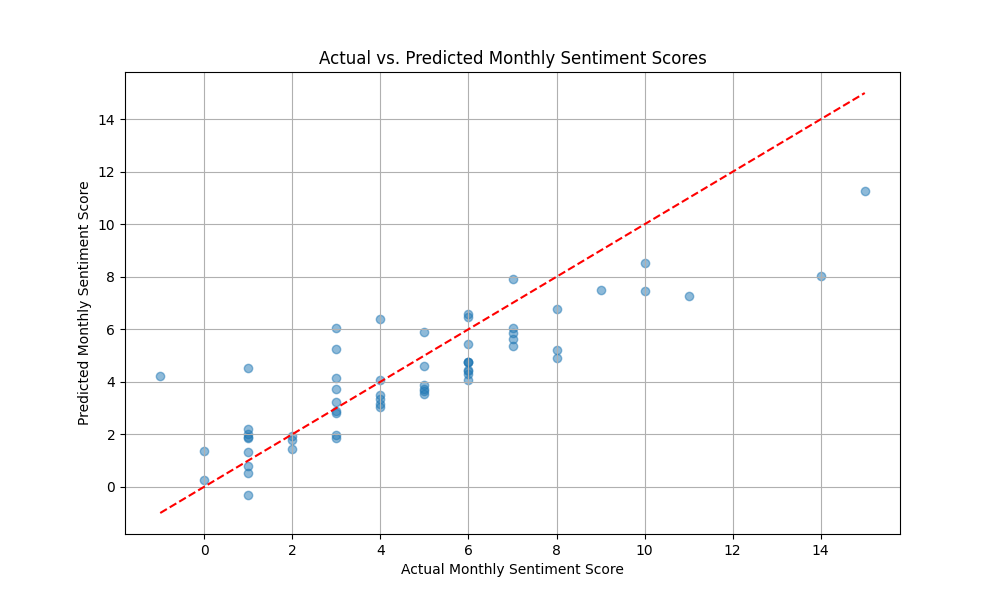


Figure 17: Plot depicting Actual Monthly Sentiment Score and Predicted Monthly Sentiment Score

The points generally follow the red diagonal line, indicating that predicted scores increase as actual scores increase. However, there is noticeable scatter around the line, especially at higher sentiment values, showing that while the model captures the overall trend, it still produces errors for some employees/months. This spread suggests moderate predictive accuracy but room for improvement. 

Figure 18: Residual Plot for monthly sentiment model

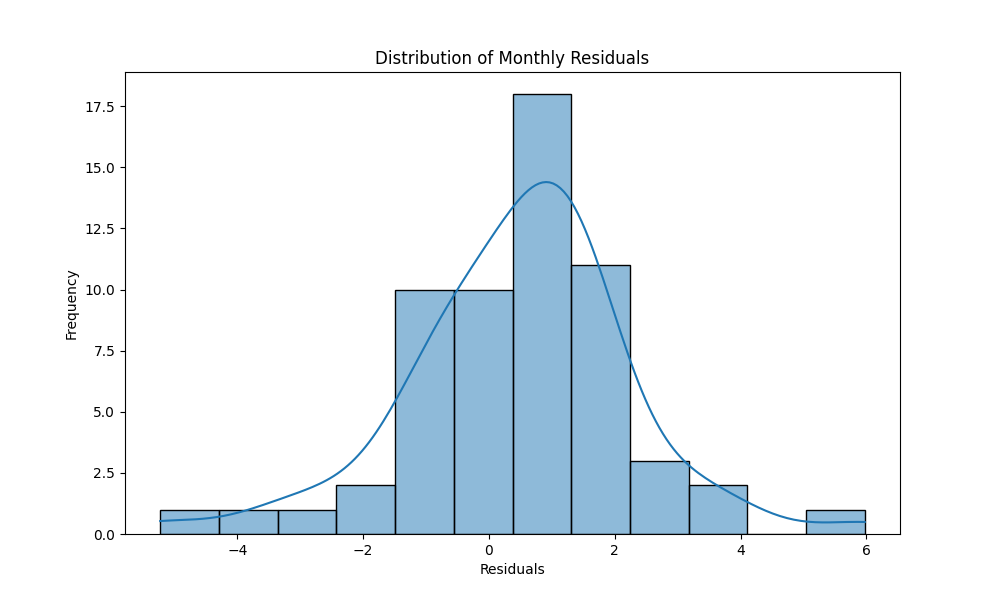
Residuals are scattered fairly evenly around the zero line across the range of predicted sentiment scores, with no clear funnel shape or curved pattern. This indicates that the model’s variance is relatively consistent (no heteroscedasticity) and that there is no major unmodeled non-linear relationship. A few larger residuals appear at higher predicted values, indicating slightly less accuracy in those cases. 

Figure 19: Distribution of Residuals of Monthly Sentiment Model

The residuals are approximately centered around zero and resemble a roughly normal distribution, which supports the linear regression assumption of normally distributed errors. Most residuals cluster near zero, but a few outliers on both ends suggest occasional over- or under-predictions by the model.

These results suggest that content-based features like polarity and subjectivity are much more informative predictors of monthly sentiment than purely structural metrics (message length, word count, message count).Overall, the linear regression model is reasonably effective at predicting monthly sentiment scores using this combination of features and represents a substantial improvement over the earlier version.

# Overview and Evaluation of the Predictive Model

One model predicts the sentiment score of individual message, one predicts the monthly aggregated sentiment score. While the second model perform slightly better, both model don’t show good results and performance. This indicate better more sophisticated modeling techniques or the inclusion of different types of features would be necessary to build a more effective predictive model.

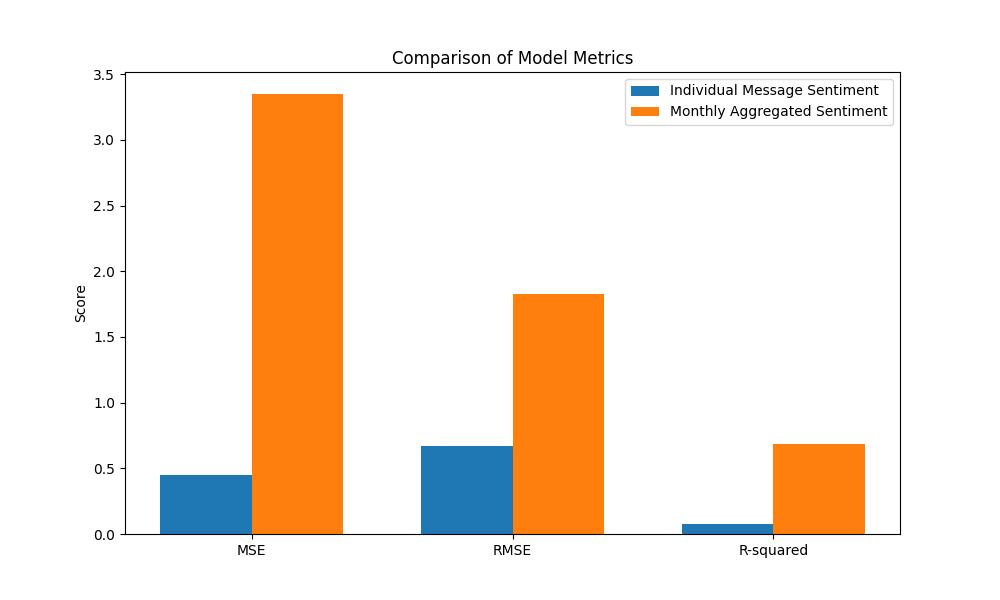


Figure 20: Comparison of metrics of two linear regression models