Final Report (Employee Sentiment Analysis)

Project Overview and Methodology

This project involved a comprehensive analysis of an unlabeled dataset of employee messages to assess sentiment and engagement. The primary goal was to work from raw data, apply Natural Language Processing (NLP) and statistical techniques to derive insights, and build a predictive model. The project was executed through a series of distinct tasks, including sentiment labeling, exploratory data analysis, employee scoring and ranking, flight risk identification, and predictive modeling.

The methodology for each core task was as follows:

1. Data Processing: The initial test.csv dataset was loaded using the Pandas library in Python. A preliminary cleaning process was applied to the 'body' of the messages. This included converting all text to lowercase, removing special characters, and eliminating extra whitespace to create a standardized clean_body column for analysis.

2. Sentiment Labeling: To automatically label each message as "Positive," "Negative," or "Neutral," a Large Language Model (LLM) was employed using the Hugging Face Transformers library. Specifically, the sentiment-analysis pipeline, which defaults to a fine-tuned DistilBERT model, was used. A confidence threshold of 0.6 was set; if the model's confidence score for a "POSITIVE" or "NEGATIVE" label was below this threshold, the message was classified as "Neutral."

3. Statistical Analysis: Subsequent tasks relied heavily on data aggregation and analysis using Pandas. Monthly scores were calculated by mapping sentiment labels to numerical values (+1, -1, 0) and grouping by employee and month. Rolling window calculations were used for flight risk identification.

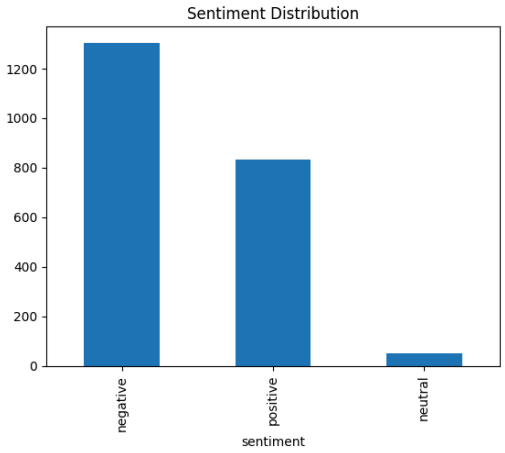
4. Predictive Modeling: A linear regression model was developed using the Scikit-learn library to predict monthly sentiment scores based on engineered features.

Exploratory Data Analysis (EDA)

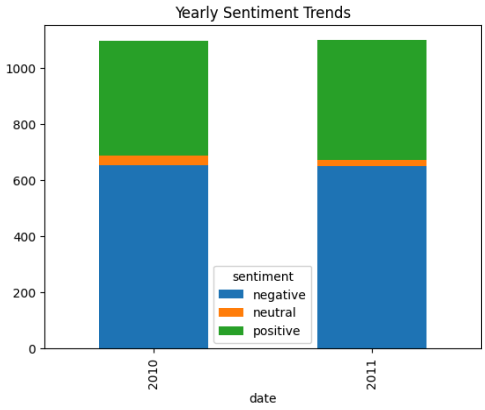
A thorough exploratory analysis was conducted to understand the structure, distribution, and trends within the dataset.

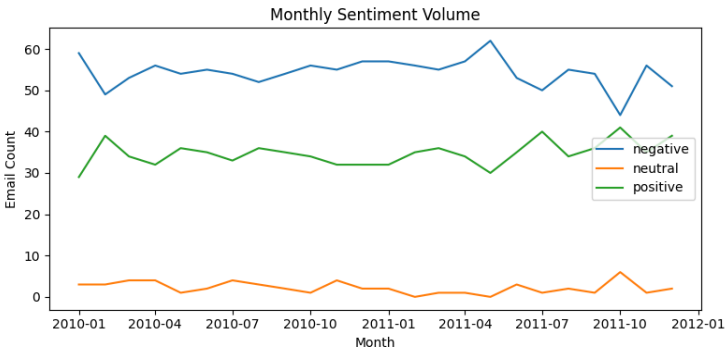
Data Structure: The dataset comprises 2,191 email records. An initial check confirmed there were no missing values in any of the key columns (Subject, body, date, from), ensuring a complete dataset for analysis. The date column was converted to a datetime format to enable time-series analysis.

Sentiment Distribution: The sentiment labeling process revealed a significant imbalance in the distribution of sentiments. The majority of messages were classified as Negative (1,304 messages), followed by Positive (834 messages). A very small fraction was labeled as Neutral (53 messages). This strong negative skew suggests a prevailing negative tone in the analyzed employee communications, which could be a key area for organizational focus. The low count of neutral messages is likely a result of the confidence threshold applied during the LLM-based labeling, which tended to assign a definitive positive or negative sentiment when possible.



Trends Over Time: Analysis of sentiment trends over the years 2010 and 2011 showed a consistent pattern. In both years, negative messages were the most frequent, followed by positive messages, with neutral messages remaining a small minority. There were no dramatic shifts in the overall sentiment balance between the two years, indicating a stable, albeit negative, communication environment during this period. Monthly volume analysis showed some fluctuation but no clear seasonal trend, with negative messages consistently outnumbering positive ones throughout most months.





Employee Scoring and Ranking Process

To quantify and compare employee sentiment on a regular basis, a monthly scoring and ranking system was implemented.

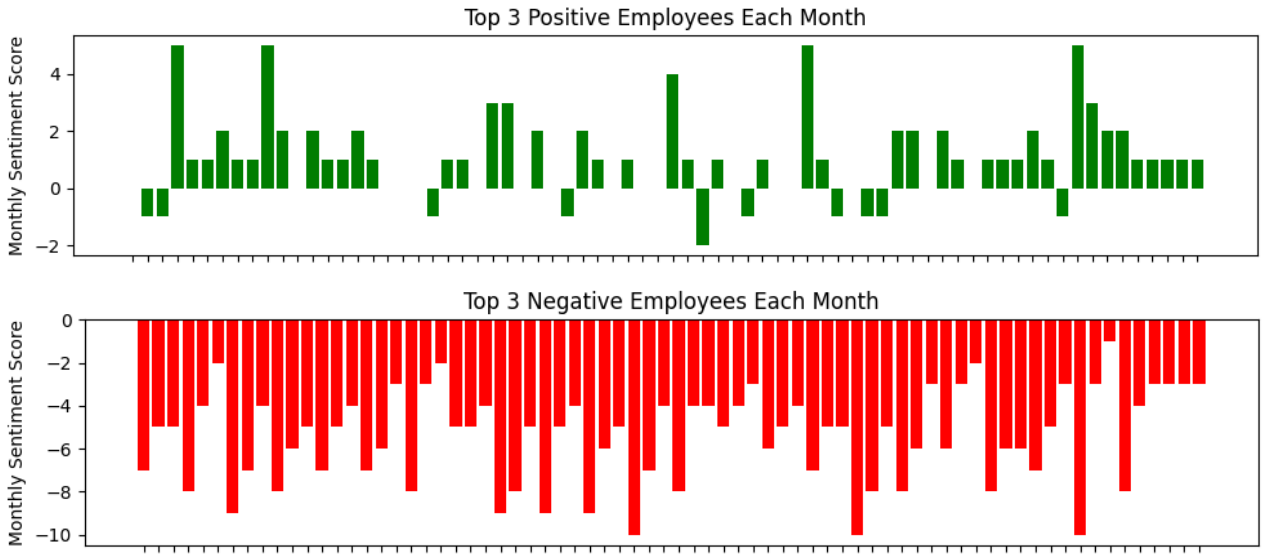
Scoring Mechanism: Each message was assigned a numerical score based on its sentiment label: Positive = +1, Negative = -1, and Neutral = 0. This simple system allows for straightforward aggregation.

Monthly Aggregation: Using the date information, scores were aggregated for each employee on a monthly basis. This was achieved by grouping the data by the 'from' (employee email) and 'year_month' columns and summing the sentiment_score. This process ensures that each employee receives a cumulative sentiment score for every month they were active, with the score resetting at the beginning of a new month.

Employee Ranking: Based on these monthly scores, two ranked lists were generated for each month:

- Top 3 Positive Employees: The three employees with the highest positive scores.
- Top 3 Negative Employees: The three employees with the lowest (most negative) scores.

Rankings were sorted first by the sentiment score (descending for positive, ascending for negative) and then alphabetically by email for tie-breaking.



Overall Performance: An analysis of these rankings over the entire two-year period identified the "Best Employee" (most frequent appearances in the Top 3 Positive list) and the "Worst Employee" (most frequent appearances in the Top 3 Negative list).

- Best Employee: johnny.palmer@enron.com (14 appearances)
- Worst Employee: patti.thompson@enron.com (12 appearances)

Flight Risk Identification

A key objective was to identify employees at potential risk of leaving the organization, based on sustained negative communication.

Criteria: An employee was flagged as a "Flight Risk" if they sent four or more negative messages within any rolling 30-day window. This method is more robust than a simple calendar month count, as it captures concentrated periods of negative sentiment irrespective of month boundaries.

Implementation: The analysis filtered for all negative messages, sorted them by date for each employee, and then applied a rolling 30-day window to count the occurrences. If any window for an employee contained 4 or more negative messages, they were added to the flight risk list.

Outcomes: The analysis identified the following ten employees as potential flight risks:

kayne.coulter@enron.com
rhonda.denton@enron.com
john.arnold@enron.com
bobette.riner@ipgdirect.com
johnny.palmer@enron.com
lydia.delgado@enron.com
patti.thompson@enron.com
sally.beck@enron.com
don.baughman@enron.com
eric.bass@enron.com

Evaluation of Criteria: It was noted that this list includes a significant portion of the most active employees. This outcome suggests that the threshold of 4 negative messages might be too sensitive for this dataset. **And more likely is my label model is too strict to give a positive.** For more actionable insights, it may be necessary to adjust the threshold (e.g., to 5 or 6 messages) or combine this metric with other indicators of disengagement.

Overview and Evaluation of the Predictive Model

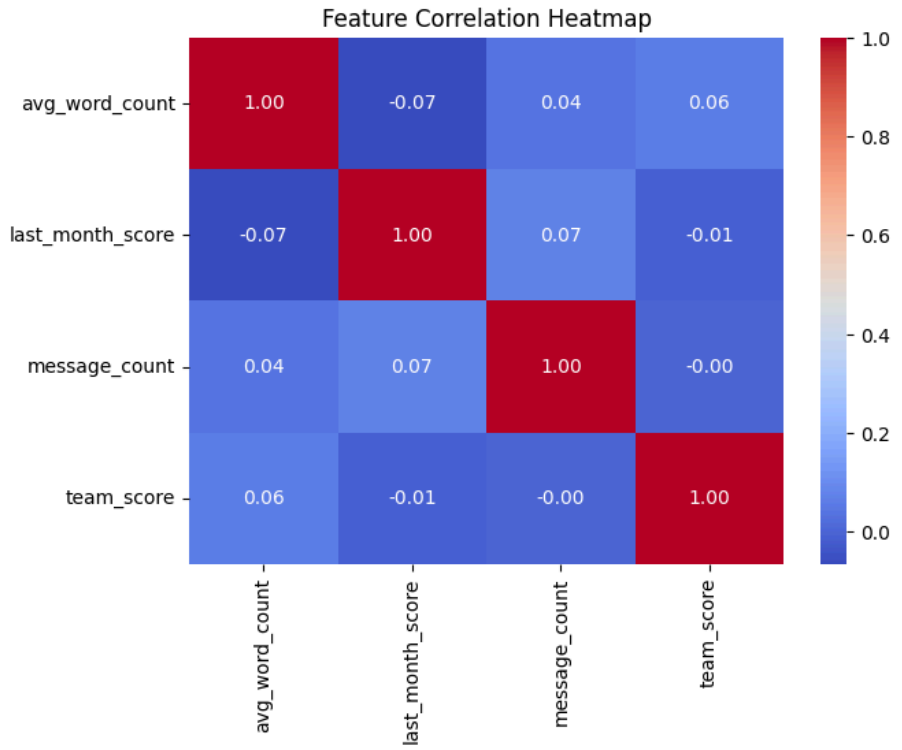
A linear regression model was developed to explore if monthly sentiment scores could be predicted based on communication patterns and other contextual factors.

Feature Engineering: The following features were engineered to serve as independent variables for the model:

- avg_word_count: The average word count of an employee's emails for a given month.

- last_month_score: The employee's total sentiment score from the preceding month, to capture momentum.
- message_count: The total number of emails sent by the employee in that month.
- team_score: The aggregated sentiment score of all employees for that month, serving as a proxy for the overall organizational mood.

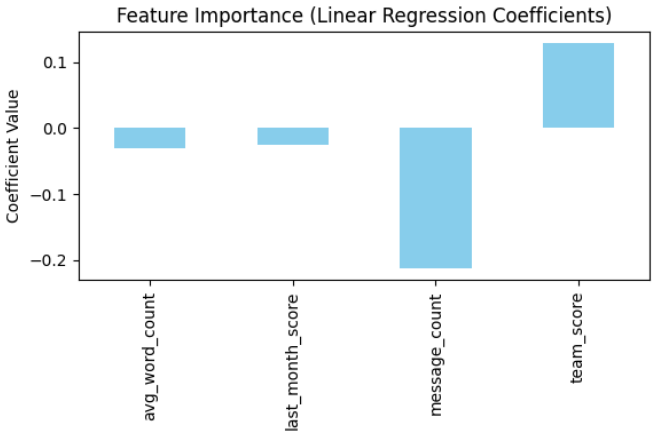
A correlation analysis confirmed that these features had low correlation with each other, making them suitable for a linear regression model.



Model Performance: The data was split into a training set (80%) and a testing set (20%). The model's performance on the test set was poor, as indicated by the following metrics:

- R-squared: 0.108. This extremely low value means the model only explains about 10.8% of the variability in monthly sentiment scores, giving it very limited predictive power.

- Mean Squared Error (MSE): 7.467. This error rate is high relative to the range of possible sentiment scores, indicating significant prediction inaccuracies.



Interpretation and Conclusion: The model's coefficients were all small, with message_count (-0.21) having the largest (negative) impact, suggesting that a higher volume of emails is weakly associated with a more negative score. The team_score (0.13) had a slight positive correlation, implying that an individual's sentiment is marginally influenced by the overall team sentiment.

I personally think the model's poor performance is likely due to two primary factors:

- **Data Sparsity:** Aggregating the data to a monthly level significantly reduced the number of samples available for training, making it difficult for the model to learn meaningful patterns.
- **Limited Predictive Features**: The available data (email text and metadata) provides limited information to predict a complex human attribute like sentiment. Factors outside of email communications (e.g., project deadlines, personal issues, management changes) likely have a much stronger influence on employee sentiment.

Given these limitations, a simple linear regression model with the current features is insufficient for accurately forecasting monthly employee sentiment scores.