Final Report: Employee Sentiment Analysis

May 7, 2025

Sean Kwon

**Project Overview**

The objective of this project was to analyze an unlabeled dataset of employee messages to assess sentiment and engagement. The dataset included various emails, and the goal was to categorize each message into sentiment labels (Positive, Negative, or Neutral) and further analyze trends, employee rankings, flight risk, and predictive modeling.

**Approach and Methodology**

* **Sentiment Labeling**: TextBlob was used to assign sentiment labels to each message. Messages were classified as Positive, Negative, or Neutral based on their polarity score.
* **Exploratory Data Analysis (EDA)**: Sentiment distribution was analyzed across the dataset and explored trends over time, including the sentiment distribution by month.
* **Employee Scoring and Ranking**: Each employee’s sentiment was aggregated into monthly scores, with positive messages contributing +1, negative messages contributing -1, and neutral messages contributing 0. Employees were ranked based on their sentiment scores.
* **Flight Risk Identification**: Employees who had sent 4 or more negative messages within a 30-day window were flagged as flight risks.
* **Predictive Modeling**: A Linear Regression model was built to predict sentiment scores based on message frequency and time-based features.

**Key Findings from EDA**

* **Sentiment Distribution**: The dataset exhibited a relatively balanced distribution of sentiment labels, with some months showing higher levels of Negative sentiment.
* **Sentiment Trends**: Sentiment trends showed fluctuations across months, indicating periods of higher employee engagement or dissatisfaction.
* **Employee Sentiment**: Employee sentiment scores varied significantly, with some employees consistently demonstrating more positive engagement, while others had more negative sentiment.

**Employee Scoring and Ranking Process**

* **Scoring Methodology**: The sentiment of each message was converted into a score (Positive = 1, Negative = -1, Neutral = 0). These scores were aggregated monthly for each employee.
* **Top Positive Employees**: Employees who consistently displayed high positive sentiment scores over time were identified. Visual representation of these rankings is provided.
* **Top Negative Employees**: Similarly, employees with consistently negative sentiment scores were identified. This allows pinpointing individuals who may need attention in terms of engagement or support.

**Flight Risk Identification**

* **Criteria**: An employee was flagged as a flight risk if they sent 4 or more negative messages within a rolling 30-day period. This identifies potential disengaged employees who may be at risk of leaving the company.
* **Outcomes**: The flight risk identification process flagged multiple employees, highlighting those with consistently negative sentiments.

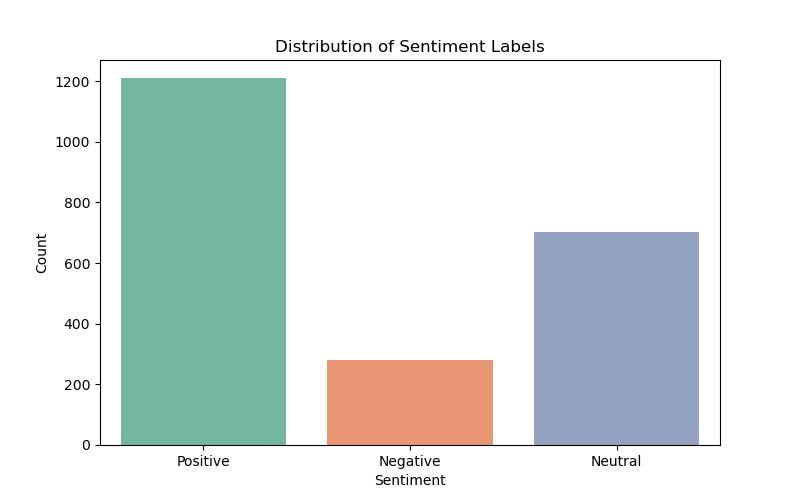
**Overview and Evaluation of Predictive Model**

* **Model**: A Linear Regression model was used to predict sentiment scores based on features such as month and message frequency.
* **Performance**:
  + **Mean Squared Error (MSE)**: The model had an MSE of 9.995, which suggests some room for improvement in terms of predictive accuracy.
  + **R-squared**: The model's R-squared value was -0.116, which indicates that the model's predictions were not well correlated with the actual sentiment scores.

The low R-squared value suggests that a more complex model might be needed to improve prediction accuracy.

**Visualizations and Tables**

**Sentiment Distribution:**



The sentiment distribution bar chart below shows the classification of employee messages into Positive, Negative, and Neutral sentiments. The Positive Sentiment category dominates the dataset, with over 1,200 messages categorized as positive. Neutral Sentiment follows, with a substantial number of messages in this category. The Negative Sentiment category is the least frequent, indicating fewer negative messages overall.

**Sentiment Trends Over Time:**

A graph of different colored lines

AI-generated content may be incorrect.

The sentiment trends over time chart below shows how the sentiment of employee messages fluctuated from January 2010 to December 2011. It tracks the message count for Positive, Neutral, and Negative sentiments each month. The chart highlights noticeable spikes in positive sentiment, with periods of lower engagement in negative and neutral sentiment. This provides insights into employee sentiment dynamics over time.

**Top Positive and Negative Employees:**

A close-up of several bars

AI-generated content may be incorrect.

The bar chart below illustrates the top positive and negative employees based on their sentiment scores. The Top Positive Employees have the highest number of positive messages, while the Top Negative Employees have sent the most negative messages. This ranking allows us to identify employees who have significantly influenced the overall sentiment in the dataset.

**Flight Risk Identification:**

A graph of different colored bars

AI-generated content may be incorrect.

The bar chart below shows the flight risk employees based on the count of negative messages sent over a rolling 30-day period. Employees flagged as flight risks have sent a significant number of negative messages, indicating potential disengagement.

The colors in the chart represent different levels of engagement, with darker colors indicating employees who have sent more negative messages (higher flight risk), and lighter colors indicating those with fewer negative messages. The chart displays the count of such messages for each employee flagged as a flight risk.

**Model Performance:**

A graph with a red line

AI-generated content may be incorrect.

The scatter plot below shows the actual sentiment scores compared to the predicted sentiment scores from the model. The red line represents a perfect prediction where actual and predicted values align. The spread of the points around this line indicates how well the model performed. A tight clustering around the line would suggest a good fit, while the scattered points indicate potential areas where the model could be improved.