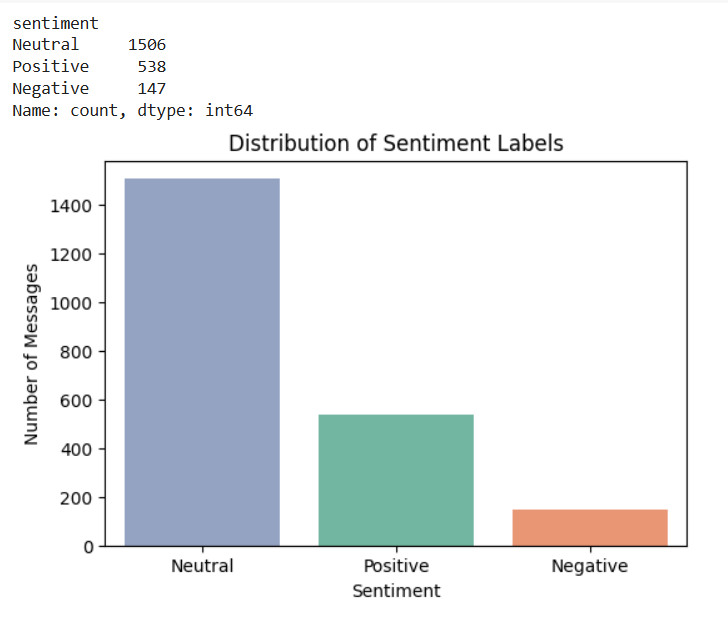
Sentiment Analysis and Employee Flight Risk Report

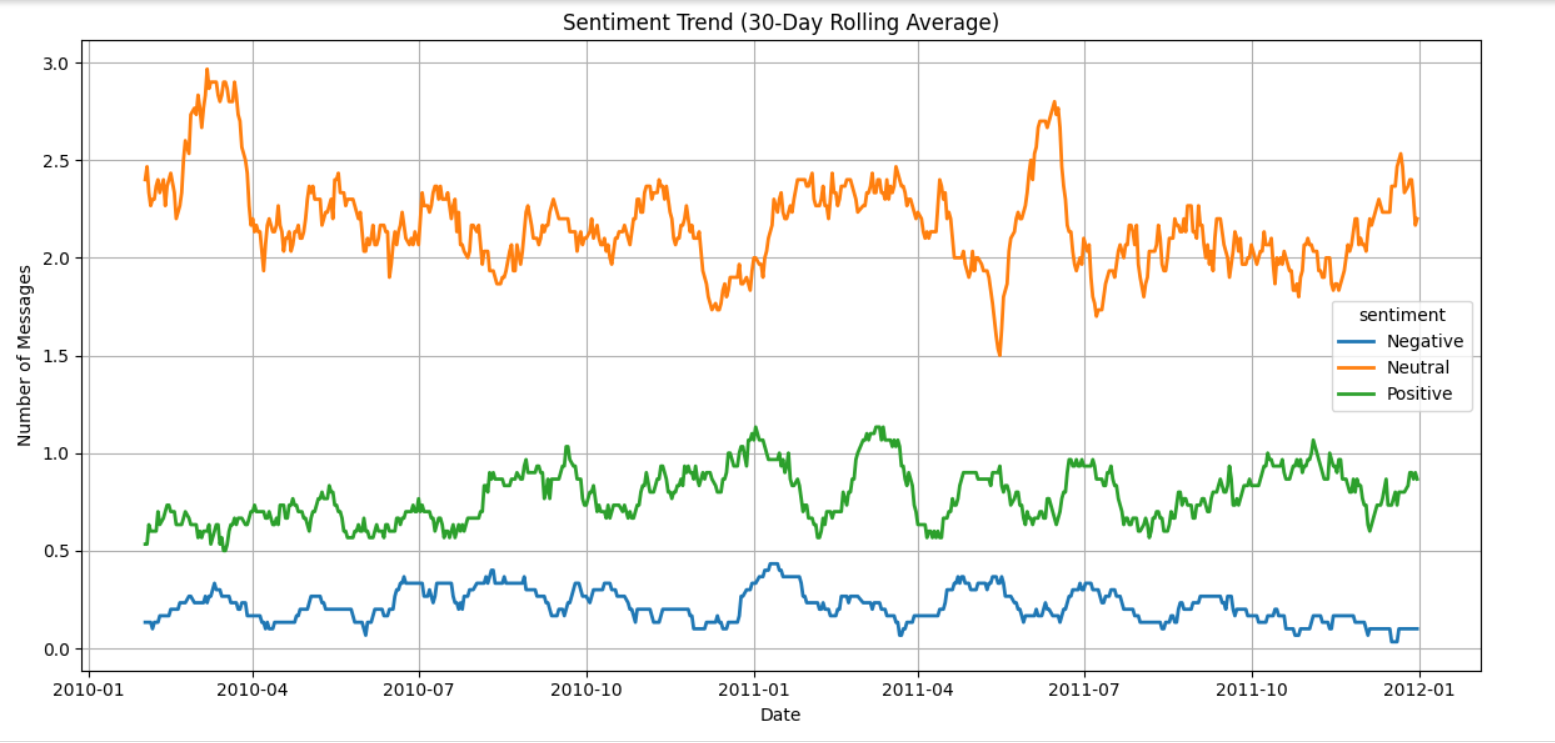
# 1. Approach and Methodology

This project aimed to analyze employee sentiment over time and identify patterns that could indicate flight risk.   
The approach included exploratory data analysis (EDA), employee sentiment scoring and ranking, flight risk identification using rolling negative sentiment counts,   
and the development of a linear regression model to predict future sentiment scores.

# 2. Key Findings from Exploratory Data Analysis (EDA)

EDA revealed fluctuations in sentiment across different employees and time periods. Some employees consistently displayed lower sentiment scores,   
and a few showed patterns of declining sentiment. A 'rolling count' of negative sentiment messages was used to identify concerning trends,   
particularly those with more than 4 negative messages in a 30-day window.



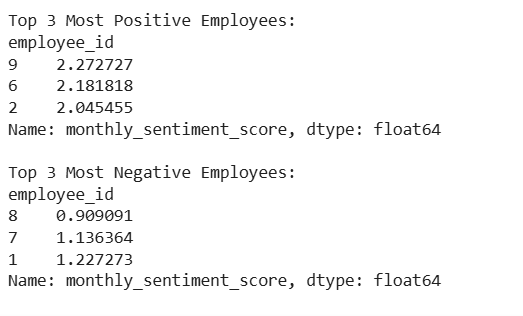


# 3. Employee Scoring and Ranking

Employees were scored using their average monthly sentiment scores. This allowed for a ranking of overall positivity or negativity in communications.   
The top 3 most positive and top 3 most negative employees were identified based on these averages.

• Positive and Negative Ranking was determined using:

* - Grouped average of `monthly\_sentiment\_score` by `employee\_id`
* - Sorted to highlight extremes (top and bottom 3)



# 4. Flight Risk Identification Criteria and Outcomes

An employee was considered at flight risk if they had 4 or more negative sentiment messages within a 30-day period.  
Using a rolling count method (grouped by employee), those crossing the threshold were flagged.  
The number of messages was visualized over time to track these at-risk individuals.



# 5. Predictive Model Overview and Evaluation

A linear regression model was built to predict `monthly\_sentiment\_score`. Features included year, month number, previous sentiment score, score change, and employee ID.  
The model was trained on 80% of the data and tested on the remaining 20%.

  
The model's performance was evaluated using R² and Mean Squared Error (MSE).   
A scatterplot of actual vs predicted values and a histogram of residuals were used for visual validation.