Employee Sentiment and Engagement Analysis – Final Report

**Project Overview**

This project involves analyzing an unlabeled dataset of employee messages to assess sentiment and engagement. The task was to work on 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.

# 1. Methodology

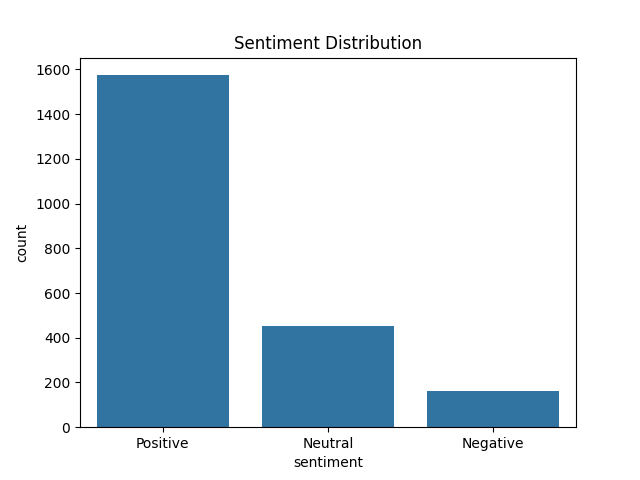
The analysis followed these key steps:

* Data Wrangling: Combined email subject and body into a unified text field.
* Sentiment Labeling: Applied VADER sentiment analysis to classify text as Positive, Neutral, or Negative.
* Exploratory Data Analysis (EDA): Analyzed sentiment distributions and monthly trends.
* Scoring: Assigned sentiment scores to messages to assess employee engagement over time.
* Ranking: Identified most and least positively engaged employees each month.
* Flight Risk Detection: Flagged employees with frequent negative messages within short timeframes.
* Predictive Modeling: Built a linear regression model to predict engagement score based on message volume and length.

# 2. Findings

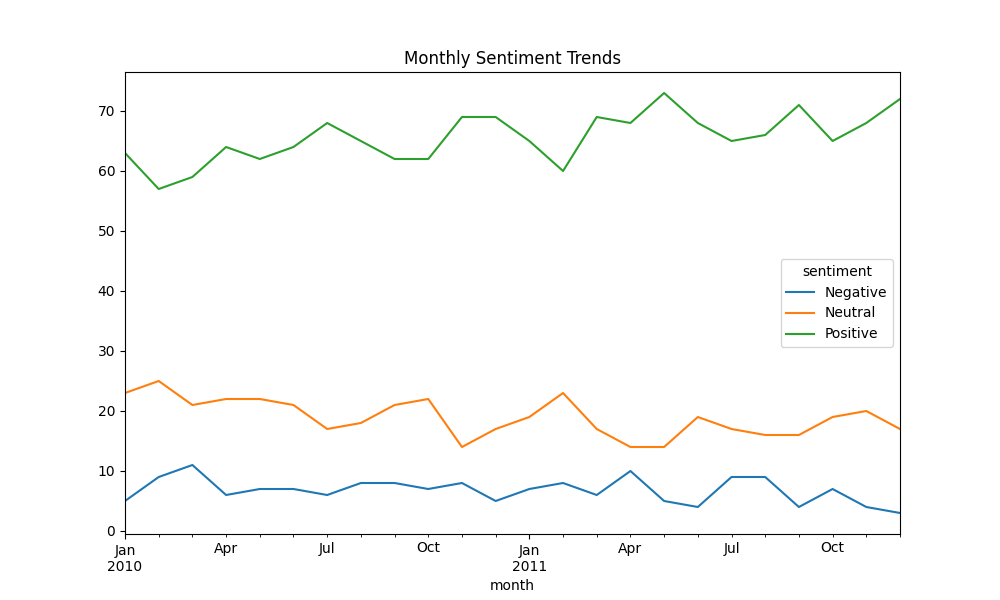
## 2.1 Sentiment Distribution

Most messages were Positive, indicating general employee satisfaction. Negative sentiments formed a significant minority, which is quite commendable but suggesting areas of concern.



## 2.2 Monthly Sentiment Trends

Sentiments fluctuate monthly, with dips during particular periods potentially indicating organizational stress or change.



## 2.3 Top and Bottom Engaged Employees

Top Positive: Employees with the highest monthly sentiment scores.  
Top Negative: Employees with the lowest scores, potentially experiencing disengagement.  
These rankings provide insight into employee morale over time.

## 2.4 Flight Risk Employees

Employees who sent 4 or more negative messages within 30-day windows were flagged as potential flight risks. These individuals may require support or managerial intervention to prevent a crisis.

# 3. Predictive Modeling and Insights

A Linear Regression model was trained to predict sentiment engagement scores using message count and average message length as features.  
Model Performance:  
- R² Score: 0.713  
- RMSE: 1.836  
  
This indicates that engagement score can be moderately predicted by communication patterns. Frequent and longer messages tend to correlate with stronger sentiment expressions.

# 4. Conclusion and Recommendations

The sentiment analysis pipeline provides meaningful insights into employee morale and engagement trends. It enables HR and management teams to identify top performers, disengaged individuals, and potential flight risks early.  
  
Recommendations:  
- Conduct regular sentiment monitoring to track employee well-being.  
- Intervene with employees showing declining sentiment trends.  
- Recognize and support high-engagement employees to retain talent.  
- Expand features for future models, such as email recipient sentiment and topic modeling.