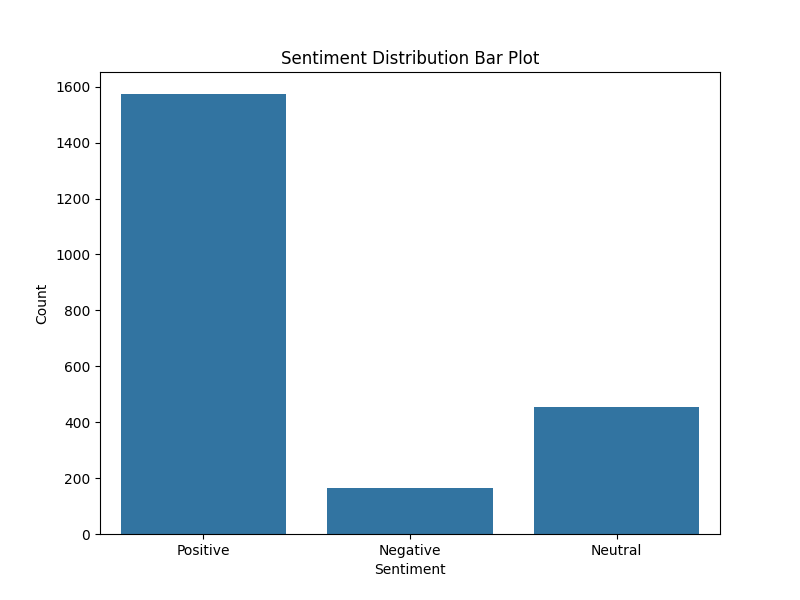
Employee Sentiment Analysis Final Report

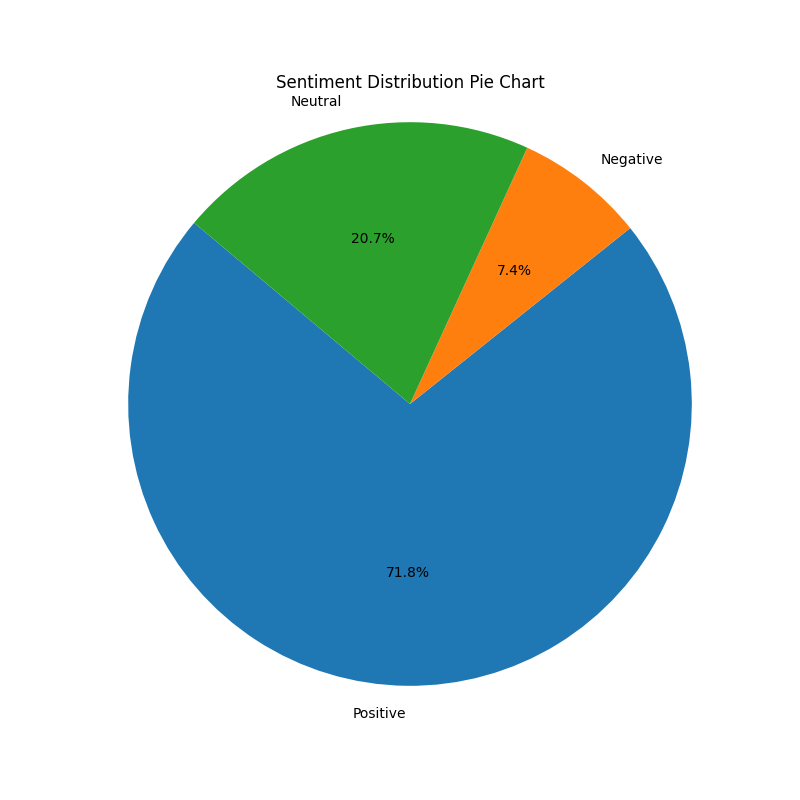
# Introduction

This report details the analysis of an employee message dataset to assess sentiment and engagement. The project involved sentiment labeling, exploratory data analysis, employee score calculation, ranking, flight risk identification, and predictive modeling.

# Data Analysis

The dataset was analyzed using VADER for sentiment labeling, categorizing messages as Positive, Negative, or Neutral. Exploratory data analysis provided insights into sentiment distribution and basic data structure.





# Employee Sentiment and Ranking

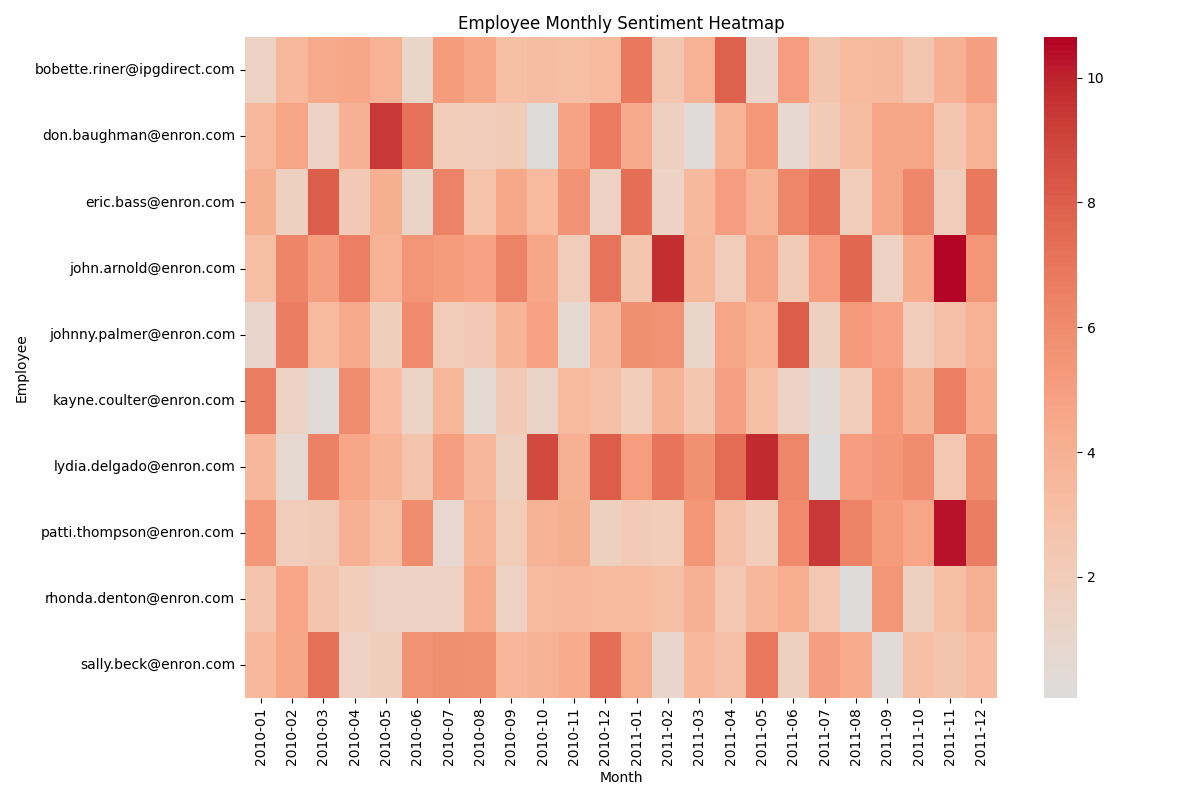
Monthly sentiment scores were calculated for each employee. Employees were ranked based on these scores to identify top positive and negative sentiment.

**Top 3 Positive Employees (Sample):**

from year\_month sentiment\_score year\_month\_str  
120 kayne.coulter@enron.com 2010-01 12 2010-01  
168 patti.thompson@enron.com 2010-01 8 2010-01  
24 don.baughman@enron.com 2010-01 6 2010-01  
73 john.arnold@enron.com 2010-02 12 2010-02  
97 johnny.palmer@enron.com 2010-02 8 2010-02

**Top 3 Negative Employees (Sample):**

from year\_month sentiment\_score year\_month\_str  
0 bobette.riner@ipgdirect.com 2010-01 2 2010-01  
96 johnny.palmer@enron.com 2010-01 2 2010-01  
192 rhonda.denton@enron.com 2010-01 3 2010-01  
49 eric.bass@enron.com 2010-02 2 2010-02  
145 lydia.delgado@enron.com 2010-02 2 2010-02

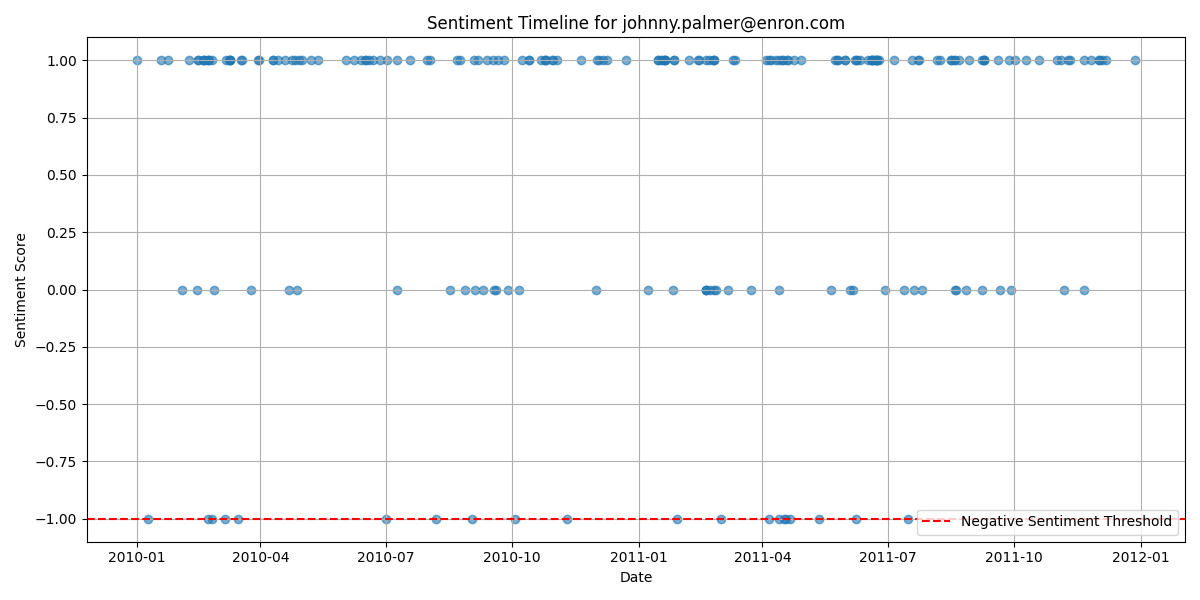


# Flight Risk Identification

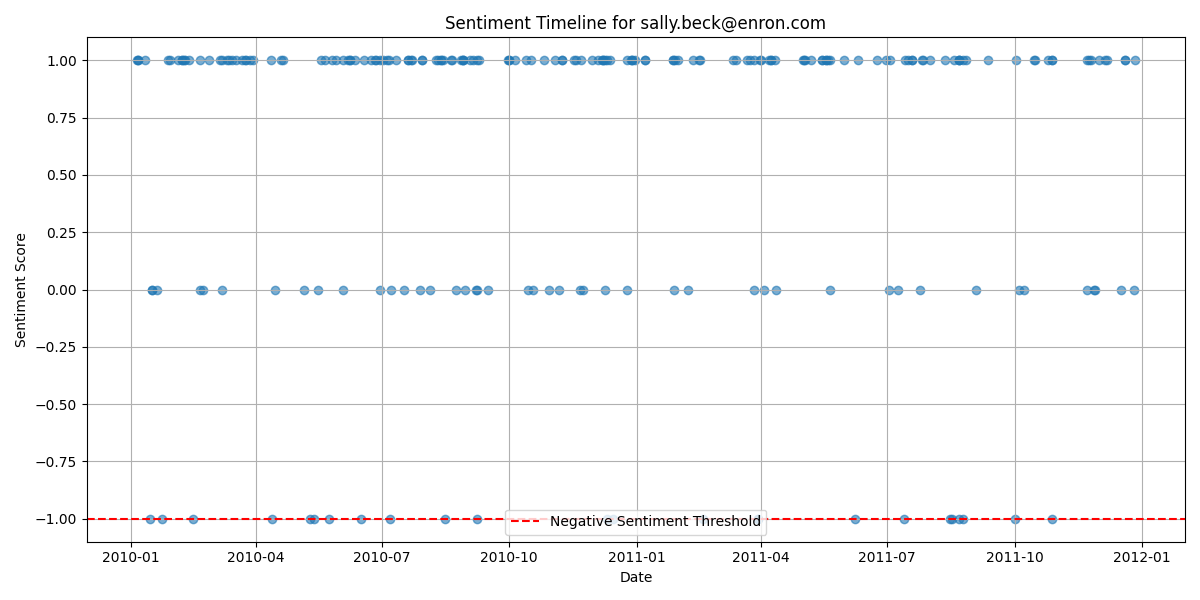
Employees were identified as potential flight risks if they sent 4 or more negative emails within a 30-day rolling window.

**Employees identified as flight risks:**

* johnny.palmer@enron.com



* sally.beck@enron.com



# Predictive Modeling

A linear regression model was developed using PyTorch to predict sentiment scores based on features like message length, word count, and message frequency. The model was trained using 5-fold cross-validation.

**Cross-Validation Results:**

Average MSE: 1.5107

Standard Deviation of MSE: 0.4054

**Model Evaluation on Test Set:**

Mean Squared Error (MSE): 0.8397

R-squared (R2): -0.9199

Actual vs. Predicted Sentiment plot not found. Please ensure it is generated.

# Model Interpretation

The model coefficients indicate the influence of each feature on the sentiment score:

* message\_length: -0.0148
* word\_count: 0.0969
* message\_frequency: 0.0017
* Intercept: -0.3135

The negative R-squared value suggests that the linear model with these features does not effectively predict sentiment score. Further feature engineering or different model types may be needed for better predictive performance.

# Conclusion

This project provided insights into employee sentiment, identified potential flight risks, and explored predictive modeling. The visualizations offer a clear overview of sentiment trends. Future work could enhance the predictive model and incorporate more advanced NLP techniques.