**Employee Sentiment Analysis Report – Lavanya Kalla**

**1. Approach and Methodology**

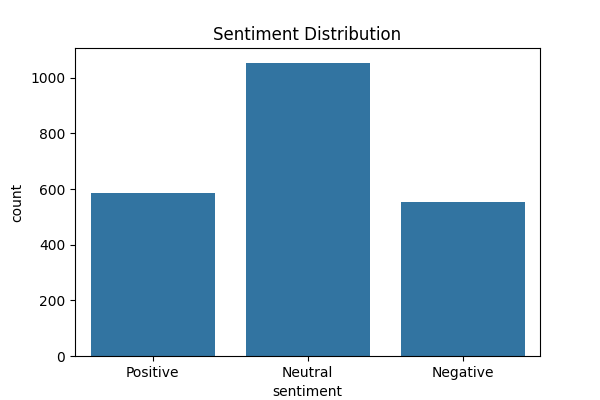
The goal of this project was to analyze employee sentiment over time using email data. I have approached this by:

* Parsing and preprocessing email data to extract key fields such as sender, date, and sentiment.
* Aggregating sentiments at the employee and monthly level.
* Performing exploratory data analysis (EDA) to uncover trends and identify potentially unhappy employees.
* Building predictive models to forecast sentiment scores using both linear and ensemble approaches.

The monthly sentiment scores were calculated by assigning values (+1, 0, -1) to positive, neutral, and negative messages and then summing them up per employee per month.

**2. Key Findings from EDA**

I have explored message trends and sentiment distributions.



A graph of different colored bars

AI-generated content may be incorrect.

A pie chart with different colored circles

AI-generated content may be incorrect.

A graph of different colored lines

AI-generated content may be incorrect.

Findings:

* Most employees had more neutral messages.
* Positive and negative messages fluctuated over time.
* Some months(October-November) had a clear drop in sentiment scores across the board.

**3. Employee Scoring and Ranking**

Each employee received a sentiment score per month based on the number of positive, neutral, and negative messages. The final score was computed as:

score = (+1 \* Positive) + (0 \* Neutral) + (-1 \* Negative)

Employees were then ranked by their monthly scores to identify the top positive and most negative employees.

**Top 3 Positive Employees (Jan 2010)**

* [john.arnold@enron.com](mailto:john.arnold@enron.com)
* [don.baughman@enron.com](mailto:don.baughman@enron.com)
* [eric.bass@enron.com](mailto:eric.bass@enron.com)

**Top 3 Negative Employees (Jan 2010)**

* [patti.thompson@enron.com](mailto:patti.thompson@enron.com)
* [sally.beck@enron.com](mailto:sally.beck@enron.com)
* [bobette.riner@ipgdirect.com](mailto:bobette.riner@ipgdirect.com)

**4. Flight Risk Identification**

I have identified flight risk employees by observing consistently low sentiment scores and large drops in sentiment over time. Criteria used:

* Low average sentiment score across months.
* Presence of repeated negative scores.

**Flight Risk Employees Identified:**

* [patti.thompson@enron.com](mailto:patti.thompson@enron.com)
* [eric.bass@enron.com](mailto:eric.bass@enron.com)
* [lydia.delgado@enron.com](mailto:lydia.delgado@enron.com)
* [bobette.riner@ipgdirect.com](mailto:bobette.riner@ipgdirect.com)
* [rhonda.denton@enron.com](mailto:rhonda.denton@enron.com)
* [don.baughman@enron.com](mailto:don.baughman@enron.com)
* [kayne.coulter@enron.com](mailto:kayne.coulter@enron.com)
* [sally.beck@enron.com](mailto:sally.beck@enron.com)
* [john.arnold@enron.com](mailto:john.arnold@enron.com)
* [johnny.palmer@enron.com](mailto:johnny.palmer@enron.com)

**5. Predictive Modeling Overview**

**a. Linear Regression**

I have started with a simple linear regression model using month\_num to predict sentiment score. Results:

* **R^2**: -0.18
* **MSE**: 3.58

A graph with black dots and a blue line

AI-generated content may be incorrect.

The model did not perform well, indicating a poor linear relationship.

**b. Random Forest Regression**

I have added more features to this model like:

* Message count
* Positive/Negative/Neutral message counts
* Month number

**Improved Performance:**

* **R^2 Score**: 0.956
* **MSE**: 0.132

**Actual vs Predicted Plot**

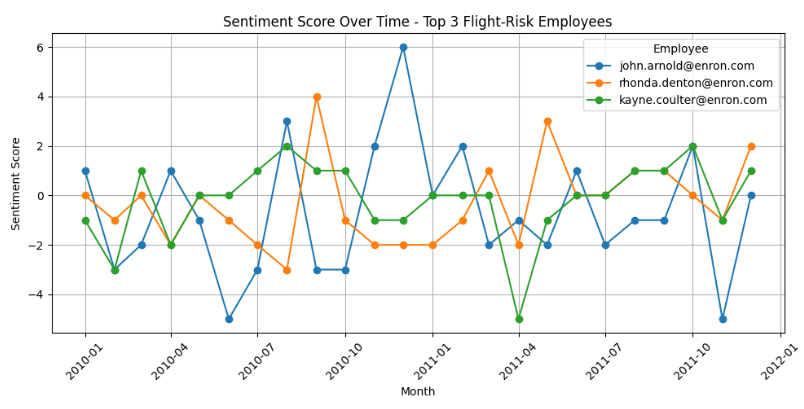
A graph with blue dots

AI-generated content may be incorrect.  
**Feature Importance**

**A graph with a number of squares

AI-generated content may be incorrect.**

**Sentiment Score Over Time for Top 3 Flight-Risk Employees**

****

The graph reveals highly volatile sentiment scores over two years for the three employees identified as flight risks. All experienced significant negative periods, likely contributing to their classification.

**6. Conclusion**

I was able to track employee sentiment, detect negative trends, and build a strong predictive model. The Random Forest regressor performed well and identified key features like message volume and polarity distribution.