**Final Report: Employee Sentiment Analysis**

# 1. Introduction & Methodology

This project analyzes employee email communications to extract sentiment insights, rank employees based on positivity/negativity, and identify potential flight risks. We use a dataset of emails containing subject, body, sender, and date metadata. Sentiment analysis was performed using TextBlob, generating polarity scores (-1.0 = negative, 0 = neutral, 1.0 = positive).

The dataset was preprocessed by handling missing values, extracting text from email bodies, and converting date fields into structured time features. Employee-level aggregation was applied to compute sentiment scores over time.

# 2. Exploratory Data Analysis (EDA)

We conducted exploratory analysis to understand sentiment distributions, top senders, and temporal trends.

- Distribution of sentiment revealed a majority of neutral to mildly positive messages.

A graph of different colored squares

AI-generated content may be incorrect.  
- Top senders were frequent contributors to both positive and negative sentiment pools.  
- Sentiment trends showed fluctuations across months, highlighting spikes of negativity during certain periods.

A graph of a number of blue bars

AI-generated content may be incorrect.

# 3. Employee Ranking Logic

Employees were ranked based on average sentiment polarity aggregated per month.  
- Top 3 Positive Performers:

lydia.delgado@enron.com (score: 6)

patti.thompson@enron.com (score: 6)

kayne.coulter@enron.com (score: 5)

- Top 3 Negative Performers:

bobette.riner@ipgdirect.com (score: 1)

johnny.palmer@enron.com (score: 2)

eric.bass@enron.com (score: 3)

# 4. Flight Risk Criteria & Results

Employees with consistently negative polarity over multiple months were flagged as potential flight risks. The analysis identified a small subset of employees who exhibited recurring negative sentiment in their communications.

bobette.riner@ipgdirect.com

eric.bass@enron.com

john.arnold@enron.com

johnny.palmer@enron.com

lydia.delgado@enron.com

patti.thompson@enron.com

rhonda.denton@enron.com

sally.beck@enron.com\

# 5. Predictive Modeling

We trained a Linear Regression model to predict average monthly polarity using features:  
- Message count  
- Average word count  
- Other text-derived features

The dataset was split chronologically: the first 70% for training and the last 30% for testing.  
Results:  
- R²: -8.45 (indicating very poor explanatory power)  
- RMSE: 0.034 (low due to small polarity scale).

Interpretation: Linear regression failed to meaningfully predict sentiment trends. Future work should explore advanced NLP-based models such as transformers or recurrent networks.

The R² value is negative (−8.45), indicating poor predictive fit ,linear features are insufficient to explain sentiment trends. RMSE of 0.0345 is small, but relative to the narrow polarity range (−1 to 1), the model is not useful. Advanced NLP methods are recommended.

# 6. Visualizations

The following visualizations were produced and are included in the visualizations/ folder:  
- Sentiment distribution

A graph of different colored squares

AI-generated content may be incorrect.  
- Word count distribution

A graph of a number of blue bars

AI-generated content may be incorrect.  
- Monthly records

A blue and white graph with numbers

AI-generated content may be incorrect.

# 7. Conclusion & Next Steps

Key Insights:  
- Most employees maintain neutral-to-positive communication tone.  
- A small subset shows consistently negative sentiment, suggesting possible disengagement.  
- Predictive modeling with simple regression was not effective.  
  
Limitations:  
- TextBlob provides limited sentiment nuance.  
- Context of emails (sarcasm, domain-specific tone) is not captured.  
  
Next Steps:  
- Implement advanced NLP models (e.g., BERT).  
- Incorporate topic modeling for deeper insight.  
- Expand feature engineering with richer metadata (department, role, etc.).