# **Final Report: Employee Sentiment Analysis**

## **Project Overview**

This project aimed to analyze employee email communications to assess their sentiment and engagement. We worked on an unlabeled dataset (test.csv) and used NLP techniques and statistical modeling to derive actionable insights.

## **Methodology**

### **Sentiment Labeling**

We labeled each email message as **Positive**, **Neutral**, or **Negative**.

* We used TextBlob to score polarity, mapping:
  + Positive: > 0.1
  + Negative: < –0.1
  + Neutral: in between

We acknowledged that thresholds should ideally be validated against domain data (FAQ Q1 and Q2), and future iterations should include multiple models or human validation.

### **Exploratory Data Analysis (EDA)**

* **Sentiment Distribution:**
  + Positive: 1,218 messages
  + Neutral: 703 messages
  + Negative: 270 messages
* **Time Trends:** We analyzed monthly sentiment trends to identify engagement changes over time.
* **Visualizations:** We included plots to show overall sentiment distribution and monthly sentiment evolution. Each plot was saved and interpreted to avoid "chart without story" issues (FAQ Q3).

### **Employee Score Calculation**

* **Monthly Sentiment Scores:**
  + Positive message: +1
  + Negative message: –1
  + Neutral: 0
* Scores were aggregated per employee per month, resetting each new month.

### **Employee Ranking**

* **Top Positive Employees (example month):**
  + [kayne.coulter@enron.com](mailto:kayne.coulter@enron.com)
  + [don.baughman@enron.com](mailto:don.baughman@enron.com)
  + [eric.bass@enron.com](mailto:eric.bass@enron.com)
* **Top Negative Employees (example month):**
  + [rhonda.denton@enron.com](mailto:rhonda.denton@enron.com)
  + [bobette.riner@ipgdirect.com](mailto:bobette.riner@ipgdirect.com)
  + [johnny.palmer@enron.com](mailto:johnny.palmer@enron.com)

These rankings help highlight highly engaged and disengaged employees.

### **Flight Risk Identification**

Employees sending ≥4 negative messages in any rolling 30-day period were flagged.

**Flagged Employees:**

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

This highlights potential retention risks.

### **Predictive Modeling**

* We built a **linear regression model** to predict monthly sentiment scores.
* **Features used:**
  + Message count
  + Average message length
  + Sentiment value mapped numerically
* **Performance:**
  + MSE: 4.10
  + R²: 0.55

**Interpretation:**  
 Approximately 55% of the variance in sentiment score can be explained by these features. The low MSE indicates decent prediction quality, but there is scope for richer features (FAQ Q6 & Q7).

**Visualization:** We included an **Actual vs Predicted** plot, saved for further analysis.

## **Key Findings**

* Majority of employee emails were positive, suggesting generally high engagement.
* A few employees with repeated negative sentiments may need HR follow-up.
* Monthly sentiment scores effectively revealed patterns that simple averages could not.

## **Recommendations**

* Periodically review flagged employees to reduce flight risk.
* Use sentiment trend insights to tailor engagement programs.
* Consider enhancing models with embeddings, behavioral signals, and domain-specific thresholds.

## **Visualizations**

All plots were saved in the visualizations folder, including:

* Sentiment distribution
* Monthly trends
* Actual vs Predicted plot for model evaluation

## **Conclusion**

The project demonstrated a structured approach to employee sentiment analysis combining NLP, EDA, scoring, risk identification, and predictive modeling. All decisions, thresholds, and interpretations were documented and justified per best practices (as advised in the FAQ document).