

# Final LLM Assessment Project Report

**Name:** Shibikasri G

**Email:** shibiguna22@gmail.com

**Dataset:** test.csv

**Tool Used:** Jupyter Notebook (Python, scikit-learn, vaderSentiment)

## 1. Objective

The goal of this project was to analyze employee email data to evaluate sentiment and engagement. Tasks included labeling sentiment, performing EDA, calculating sentiment-based scores, ranking employees, identifying potential flight risks, and developing a predictive regression model to forecast sentiment trends.

## 2. Methodology

1. Loaded the raw dataset (*test.csv*) and performed preprocessing.
2. Used *VADER Sentiment Analyzer* to label messages as Positive, Negative, or Neutral.
3. Conducted Exploratory Data Analysis (EDA) to visualize sentiment distribution and monthly trends.
4. Calculated monthly sentiment scores for each employee (+1 for Positive, -1 for Negative, 0 for Neutral).
5. Ranked employees based on sentiment scores to find top positive and negative performers.
6. Identified flight risks — employees with  $\geq 4$  negative messages within 30 days.
7. Built a Linear Regression model using scikit-learn to analyze and predict sentiment trends based on message frequency and average message length.

## 3. EDA Findings

- The dataset contained employee emails with varied message frequencies across months.
- Positive messages dominated the sentiment distribution, indicating healthy engagement.
- A few employees consistently had higher negative sentiments over consecutive months.
- Message length and frequency showed mild correlation with sentiment polarity.

## 4. Employee Scoring & Ranking

- Monthly sentiment scores were computed for each employee.
- Top three positive and negative employees were identified based on aggregated scores.
- Rankings were used to highlight highly engaged and at-risk individuals.

## 5. Flight Risk Identification

- Employees sending  $\geq 4$  negative messages in a rolling 30-day window were flagged as flight risks.
- This helped isolate employees showing patterns of disengagement or dissatisfaction.

## 6. Predictive Modeling

- A Linear Regression model was trained using message length, sentiment scores, and frequency as features.
- The model achieved good interpretability, helping predict future sentiment trends.
- Metrics like Mean Absolute Error (MAE) and R<sup>2</sup> Score were used for evaluation.

## 7. Conclusion

This project successfully demonstrated sentiment analysis and predictive modeling using employee communication data. By combining NLP, visualization, and regression modeling, actionable insights were derived to improve engagement and retention.