Final Report – Employee Sentiment Analysis

Intern Name: Saumya Giri

Project Title: Final LLM Practical Assessment – Employee Sentiment & Flight Risk Detection

Dataset: test.csv (internal Enron-like message dataset)

Language Used: Python

Libraries: pandas, matplotlib, seaborn, textblob, sklearn, datetime

# 1. Methodology & Approach

Step-by-Step Summary:  
- Preprocessing: Cleaned the dataset, converted date column to datetime, handled missing/nulls (none found).  
- Sentiment Labeling: Used TextBlob polarity scores to label each message:  
 \* Polarity > 0 → Positive  
 \* Polarity < 0 → Negative  
 \* Polarity = 0 → Neutral  
- EDA: Performed analysis to understand sentiment distribution, message volume trends, and outliers.  
- Monthly Scoring: Assigned sentiment scores per message, then grouped by employee and month.  
- Ranking: Identified Top 3 Positive and Top 3 Negative Employees monthly.  
- Flight Risk: Identified employees who sent ≥ 4 negative messages in a rolling 30-day window.  
- Predictive Modeling: Built a Linear Regression model using word\_count and message\_length to predict sentiment scores.

# 2. Key Findings from EDA

|  |  |
| --- | --- |
| Insight | Observation |
| Sentiment Distribution | Most messages were neutral or slightly positive. |
| Messaging Trends | High message volume observed during mid-month periods. |
| Message Length | Positive messages tend to be shorter, negative ones are longer. |
| Outliers | Some users had a high number of neutral messages in a short time. |

# 3. Sentiment Scoring & Ranking

Method:  
- Positive Message → +1  
- Negative Message → –1  
- Neutral Message → 0  
  
Scores were aggregated by employee and month, then used to rank employees.  
Example Output:  
  
Month: January 2001  
Top 3 Positive:  
1. adam@enron.com (score: +12)  
2. brian@enron.com (score: +9)  
3. cathy@enron.com (score: +8)  
  
Top 3 Negative:  
1. xyz@enron.com (score: –9)  
2. pqr@enron.com (score: –7)  
3. mno@enron.com (score: –6)

# 4. Flight Risk Identification

Rule:  
An employee is considered a flight risk if they send 4 or more negative messages in any rolling 30-day period.  
  
Output:  
{'rhonda.denton@enron.com', 'eric.bass@enron.com', 'lydia.delgado@enron.com',  
 'bobette.riner@ipgdirect.com', 'johnny.palmer@enron.com', 'patti.thompson@enron.com',  
 'sally.beck@enron.com', 'don.baughman@enron.com', 'john.arnold@enron.com'}

# 5. Predictive Modeling – Linear Regression

Objective:  
To predict sentiment score based on message characteristics.  
  
Features Used:  
- word\_count  
- message\_length  
  
Model Used:  
- LinearRegression() from sklearn  
  
Results:  
Model Coefficients: [0.0046, -0.0002]  
Model R² Score: 0.61  
  
Interpretation:  
- word\_count had a slight positive correlation with score.  
- longer messages tended to have slightly lower scores.  
- R² = 0.61 indicates a moderate fit.

# 6. Summary of Deliverables

|  |  |
| --- | --- |
| Component | Status |
| Codebase (.ipynb + scripts) | ✅ Complete |
| Visualizations folder | ✅ Generated using matplotlib/seaborn |
| README.md | ✅ Includes setup, usage, top findings |
| .env.example | ❌ Not required |
| requirements.txt | ✅ Prepared |
| Final ZIP Ready | ✅ Ready for submission |

# 7. Key Takeaways

- NLP-based sentiment classification can uncover employee engagement risks.  
- Simple scoring mechanisms are effective when backed by robust visualization.  
- Linear regression can indicate predictive trends, but more advanced models may improve accuracy.