Final Report — Employee Sentiment Analysis

Author: Syed Mohsin  
Date: August 20, 2025

# Executive Summary

This project analyzes employee communication data to assess sentiment, engagement, and potential attrition risk. Using natural language processing (NLP) techniques with a pretrained transformer model (DistilBERT SST-2), messages were automatically labeled as Positive, Negative, or Neutral. Subsequent exploratory data analysis (EDA), sentiment scoring, employee ranking, flight risk detection, and predictive modeling provided actionable insights.  
  
Key Findings:  
- Neutral messages dominated overall sentiment distribution.  
- Negative spikes occurred in certain months.  
- Several employees consistently ranked in Top 3 Positive and Negative categories.  
- Some employees flagged as flight risks (≥4 negative mails in rolling 30 days).  
- Predictive modeling showed message frequency and negative count were strongest drivers of sentiment.

# Data Description

Columns: Subject, body, date, from.  
Preprocessing: Subject+body merged, missing values handled, date parsed.  
Dataset size: [Insert row count].

# Methodology

Sentiment Labeling: DistilBERT SST-2, thresholds for Neutral.  
EDA: Distribution plots, monthly trends, employee breakdown, message length analysis.  
Scoring: Positive=+1, Negative=-1, Neutral=0 aggregated monthly.  
Ranking: Top 3 Positive and Top 3 Negative per month.  
Flight Risk: ≥4 negative messages in 30-day rolling window.  
Predictive Modeling: Features (msg count, avg length, word count, pos/neg/neu counts), Linear Regression, evaluated with R², MAE, RMSE.

# Exploratory Data Analysis (Key Insights)

- Neutral dominates, Positive/Negative smaller share.  
- Negative spikes in certain months, possible org issues.  
- Few employees contribute most messages.  
- Negative messages tend to be longer.

# Employee Scoring & Ranking

Scores aggregated monthly. Rankings show top positive and negative employees per month.  
Example (April 2001):  
- Top Positive: [Employee Names & Scores]  
- Top Negative: [Employee Names & Scores]

# Flight Risk Detection

Employees flagged if ≥4 negative mails in rolling 30 days.  
Flight risks identified: [list employees].

# Predictive Modeling Results

Linear Regression Results:  
- R²: [value], MAE: [value], RMSE: [value].  
Feature importance shows Neg\_Count and Msg\_Count most influential.

# Limitations & Future Work

- Neutral classification could be improved with 3-class model training.  
- No metadata (department/role) used.  
- Linear Regression only; advanced models may improve accuracy.

# Conclusion

The pipeline successfully labeled messages, performed EDA, scored and ranked employees, detected flight risks, and predicted sentiment trends. Insights show potential for HR decision support with NLP-driven analysis.

# Appendix

Tools: Python, pandas, scikit-learn, transformers, matplotlib, seaborn.  
Reproducibility: Notebook runs end-to-end.  
Outputs: CSV + PNG visualizations saved in outputs/ and visualizations/ folders.