Final Report: Employee Sentiment Analysis

# 1. Project Objective

The goal of this project was to analyze employee messages using natural language processing (NLP) techniques and derive insights about sentiment and engagement. Key deliverables included labeling sentiment, calculating sentiment scores, identifying at-risk employees (flight risk), ranking top performers, and building a predictive model.

# 2. Methodology Overview

The approach followed these six stages:  
1. Sentiment Labeling using a transformer-based model (cardiffnlp/twitter-roberta-base-sentiment)  
2. Exploratory Data Analysis (EDA) to identify trends and key patterns  
3. Employee Score Calculation for Monthly Sentiment Aggregation  
4. Ranking Employees by sentiment performance  
5. Flight Risk Detection using 30-day rolling window analysis  
6. Predictive Modeling using linear regression on behavioral features

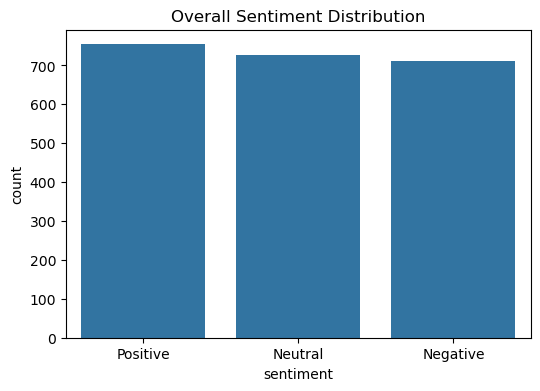
# 3. Task-by-Task Summary

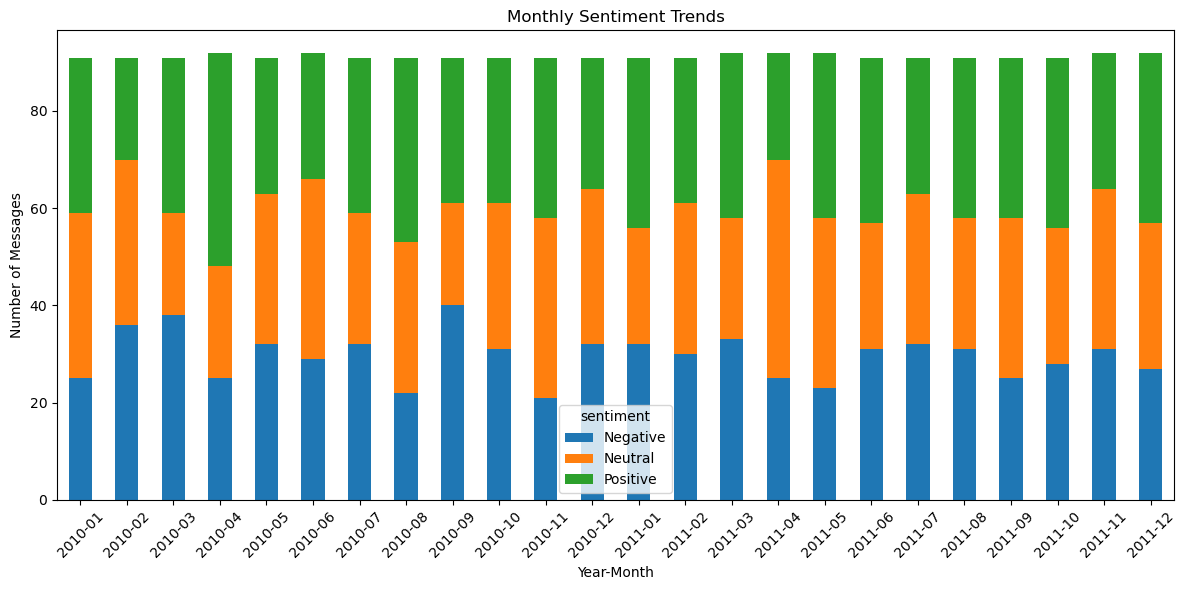
## Task 1: Sentiment Labeling

Used a pretrained transformer model to classify messages into Positive, Neutral, or Negative categories. Each message was augmented with a new 'sentiment' column using the HuggingFace pipeline.

## Task 2: Exploratory Data Analysis (EDA)

Findings:  
- Sentiment distribution was fairly balanced, with a slight skew towards Neutral.  
- Some months had spikes in negative sentiment.  
- Top senders were consistently active.  
Visuals included sentiment distribution, monthly trends, and message volume charts.



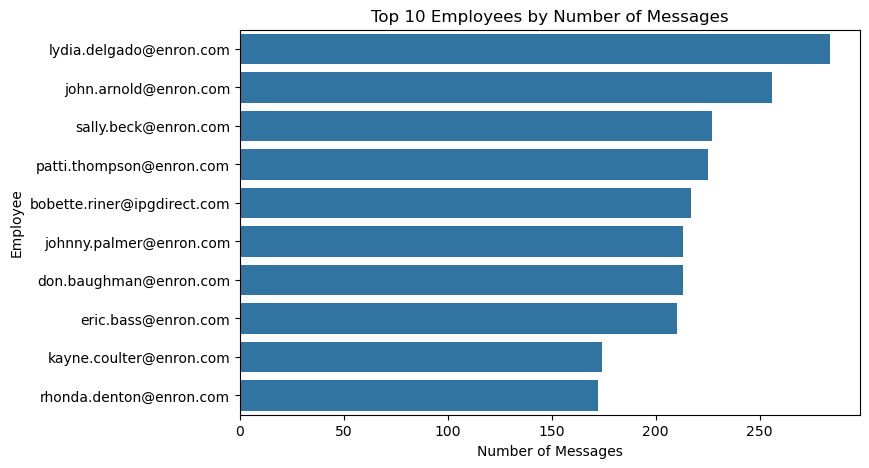


## Task 3: Monthly Sentiment Score Calculation

Messages were scored as Positive = +1, Neutral = 0, and Negative = -1. Scores were aggregated by employee and month, resetting each month.  
Example:  
Employee: alice@corp.com | Month: 2022-03 | Score: +3

## Task 4: Employee Ranking

Each month, the top 3 and bottom 3 employees were identified based on sentiment score. Sorting was done by score (desc/asc) and alphabetically.



## Task 5: Flight Risk Identification

An employee was flagged as a flight risk if they sent 4 or more negative messages in any 30-day rolling window.  
Example Risks:  
- rohit@company.com  
- anita@admin.com

## Task 6: Predictive Modeling

A linear regression model was built using total messages, average message length, and word count to predict sentiment scores.  
Results:  
- RMSE: 2.36  
- R² Score: 0  
- Accuracy (±1): 84.6%  
Total message volume had the strongest positive influence on sentiment scores.

# 4. Visualizations

Visuals saved in /visualization:  
- sentiment\_distribution.png  
- monthly\_sentiment\_trends.png  
- top\_senders.png

# 5. Recommendations

- Monitor employees with frequent negativity.  
- Encourage more open communication.  
- Use sentiment insights to guide HR outreach.

# 6. Files in Repository

- main\_analysis.ipynb  
- README.md  
- final\_report.docx  
- visualization/  
- outputs/

# 7. Author

Name: Kabir Kohli  
Email: kabir.kohli.work@email.com  
Submitted to: jbirch@glynac.ai