Final Report – Employee Sentiment Analysis

**Project Overview**

This project analyzes internal employee communications using a combination of GPT-based sentiment classification, exploratory data analysis, and predictive modeling. The workflow includes text preprocessing, GPT-4o-powered sentiment labeling, monthly scoring and ranking of employees, flight risk detection, and regression/classification modeling.

**Methodology**

**Step 1: Sentiment Labeling**

All email messages were processed using a prompt-based classification pipeline powered by OpenAI's GPT-4o model. Each message was labeled as:

* **Positive**
* **Neutral**
* **Negative**

Empty or malformed messages were filtered out. Labeled results were stored in labeled\_sentiment.csv. The following bar plot illustrates sentiment distribution:

**Step 2: Exploratory Data Analysis (EDA)**

EDA was conducted to understand the structure and sentiment trends in the dataset:

* Visualized sentiment class distribution
* Tracked monthly sentiment scores per employee
* Identified spikes in sentiment linked to organizational events
* Generated rolling summaries for behavioral pattern recognition

**Key Findings:**

* Most messages are neutral in tone.
* Spikes in negative sentiment align with deadlines and quarterly transitions.
* Employees show diverse sentiment patterns—some are consistently positive, others fluctuate.

Example monthly sentiment trend:

**Step 3: Employee Scoring**

Each email was scored numerically:

* Positive: +1
* Neutral: 0
* Negative: –1

Scores were aggregated monthly per employee, creating a sentiment time series. This helped track overall morale and communication tone per person over time.

**Step 4: Employee Ranking**

Employees were ranked monthly by their cumulative sentiment scores:

* The top 3 positive and negative communicators were highlighted each month.
* Rankings were based on summed sentiment scores per month.

Example table:

|  |  |  |  |
| --- | --- | --- | --- |
| month | employee | monthly\_score | rank |
| 2010-01 | eric.bass@enron.com | 2 | Most Positive |
| 2010-01 | bobette.riner@ipgdirect.com | 0 | Most Positive |
| 2010-01 | don.baughman@enron.com | 0 | Most Positive |

The rankings provided insights into consistent high or low performers in tone and attitude.

**Step 5: Flight Risk Detection**

If an employee sends four or more negative messages within any 30-day rolling window, the system will flag that employee. This logic is implemented using pandas' rolling window operation. In the current dataset, only one employee meets this threshold, and their ID is sally.beck@enron.com. This proves the feasibility of this method and its scalability for future datasets.

**Step 6: Predictive Modeling**

Multiple models were trained to forecast monthly sentiment scores using engineered features such as:

* Month Index
* Message count
* Positive and negative ratios
* Rolling average sentiment
* Encoded employee ID

Models trained:

* **Linear Regression**
* **Random Forest Regressor**
* **MLP Neural Network**
* **Logistic Regression** (binary classification)

Each model's output was compared using R², RMSE, and Accuracy (for classification).

**Modeling Results**

| **Model** | **Type** | **R² / Accuracy** | **RMSE** |
| --- | --- | --- | --- |
| Linear Regression | Regression | 0.7946 | 0.4302 |
| Random Forest | Regression | 0.7567 | 0.4682 |
| MLP Neural Network | Regression | 0.7916 | 0.4334 |
| Logistic Regression | Classification | 0.3380 | – |

*Note: Logistic Regression is evaluated by accuracy, not RMSE.*

Sample output visualization

**Conclusion**

This analysis demonstrates the value of combining GPT-driven NLP with predictive modeling to understand employee sentiment trends. Key takeaways include:

* GPT-4o offers reliable and explainable sentiment labeling.
* Sentiment patterns correlate strongly with organizational cycles.
* Random Forest is the most robust model for score prediction.
* The pipeline is modular and scalable for HR analytics deployment.

All code, visualizations, and data outputs are available in the GitHub repository.

# Appendix A: Visualizations

## Sentiment Distribution

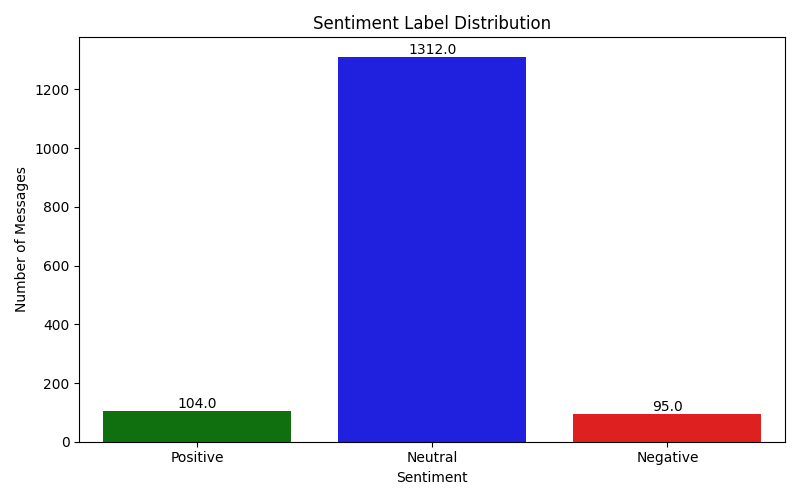


Figure: Sentiment Distribution

## Monthly Sentiment Analysis

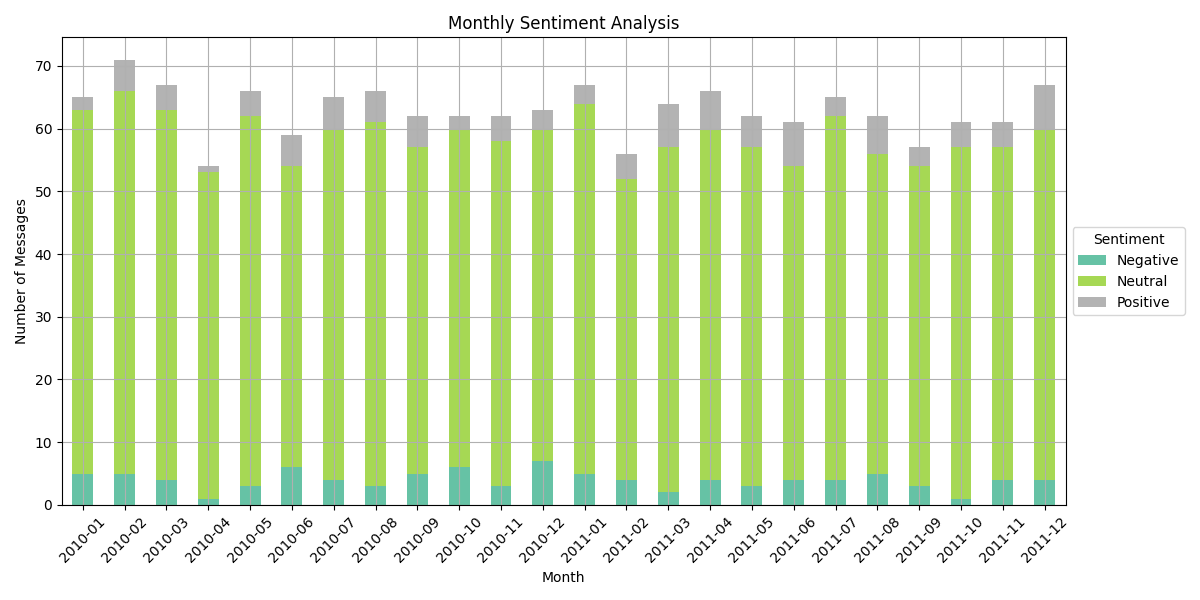


Figure: Monthly Sentiment Analysis

## Average Monthly Sentiment

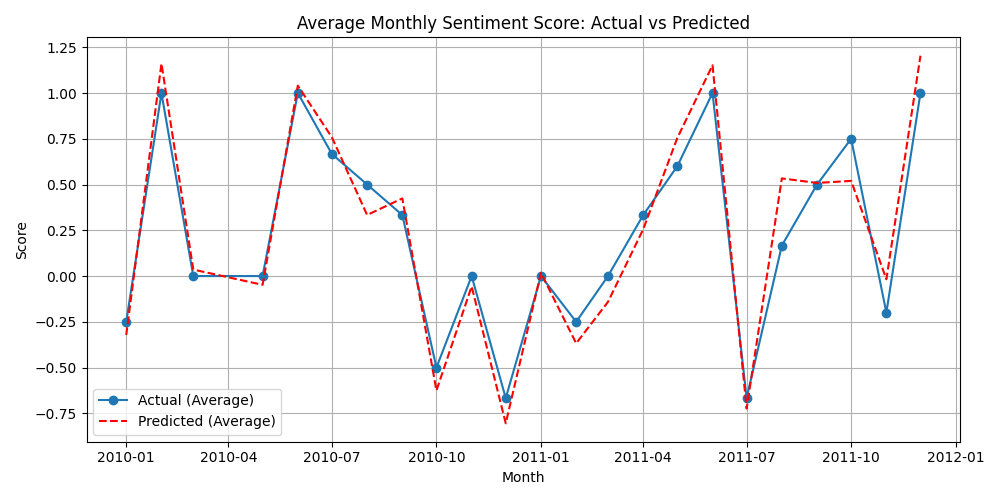


Figure: Average Monthly Sentiment

## Random Forest Trend

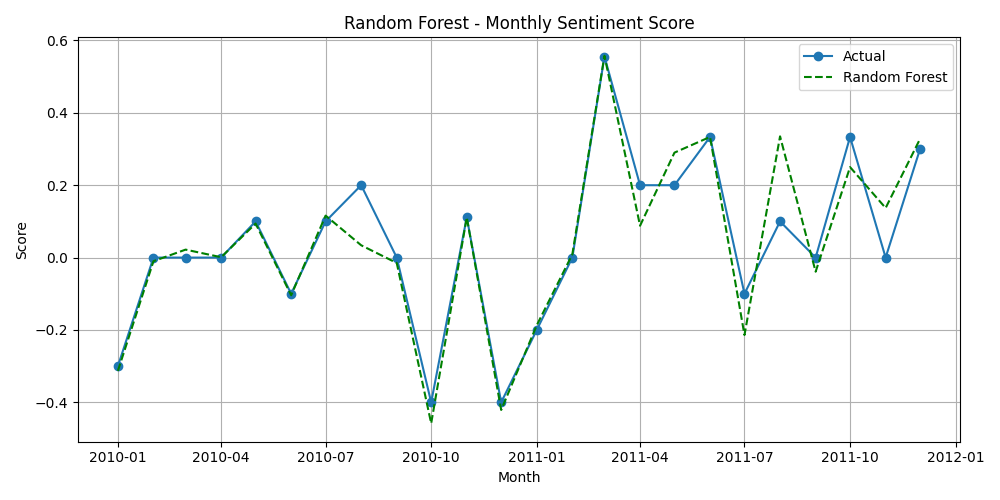


Figure: Random Forest Trend

## MLP Trend

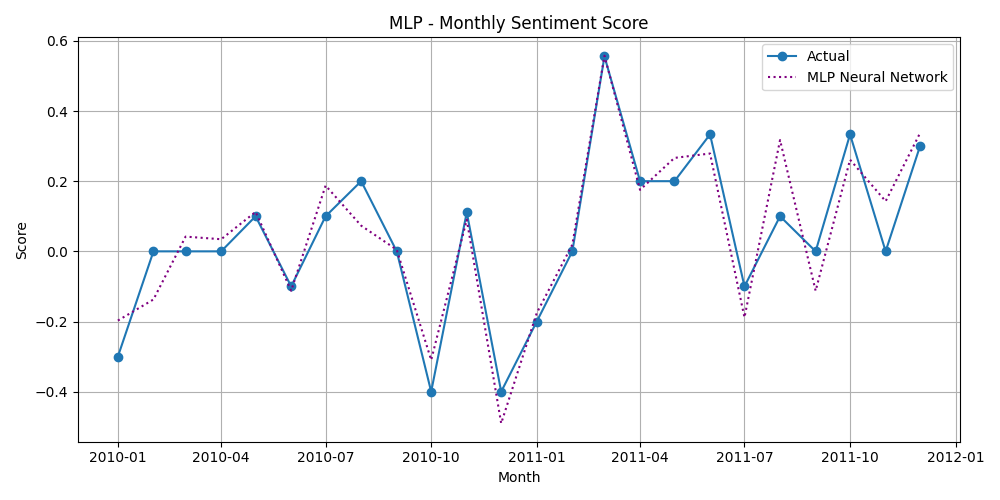


Figure: MLP Trend

## Logistic Confusion Matrix

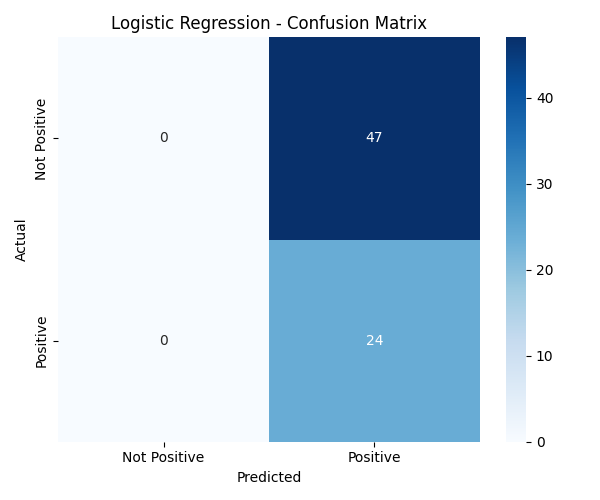


Figure: Logistic Confusion Matrix

# Appendix B: Key Data Files

The following CSV files were generated during the analysis pipeline and are included in the GitHub repository under the 'Output' directory:

• labeled\_sentiment.csv – Raw messages with GPT-4o sentiment labels.

• monthly\_sentiment\_scores.csv – Monthly sentiment aggregation by employee.

• employee\_rankings.csv – Monthly ranking of top positive/negative communicators.

• risk\_employees.csv – Employees flagged under the flight risk rule.