**Final Report: Employee Sentiment Analysis**

**1. Introduction**

The purpose of this project was to analyze employee communication data to assess overall sentiment and engagement within the organization.

By leveraging Natural Language Processing (NLP) techniques, statistical analysis, and predictive modeling, we transformed raw email data into actionable insights.

The project was divided into six tasks:

1. Sentiment Labeling

2. Exploratory Data Analysis (EDA)

3. Employee Score Calculation

4. Employee Ranking

5. Flight Risk Identification

6. Predictive Modeling

**2. Approach and Methodology**

**Data Source**

The dataset (`test.csv`) contained employee email communications with the following fields:

**- Subject:** Email subject line

**- Body:** Main content of the email

**- Date:** Date of the email

**- From:** Sender’s email address

**Methodology**

**- Text Preprocessing:** Combined subject and body, cleaned text (removing email addresses, URLs, and noise).

**- Sentiment Labeling:** Used DistilBERT (pre-trained transformer) for sentiment classification. Since the model was binary (Positive/Negative), we added a \*\*Neutral class\*\* by applying a confidence threshold.

**- EDA:** Visualized sentiment distribution, time trends, and employee activity levels.

- Scoring: Mapped sentiments to numerical values (+1, –1, 0) and aggregated per employee per month.

**- Ranking:** Identified the top 3 positive and top 3 negative employees for each month.

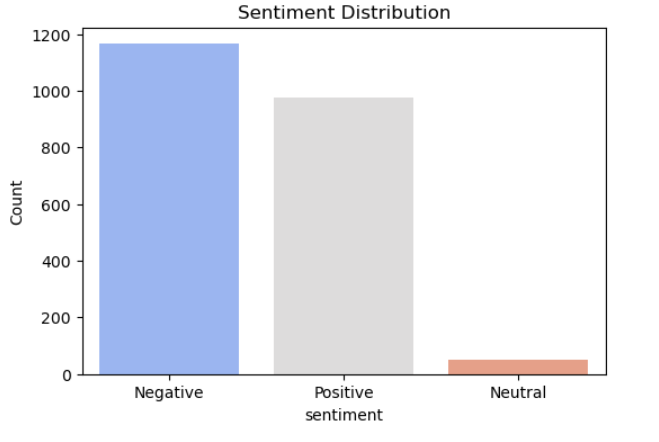
**- Flight Risk:** Flagged employees who sent 4+ negative emails within any rolling 30-day window.

**- Predictive Modeling:** Built a linear regression model using behavioral features (email counts, average lengths, sentiment composition) to predict monthly sentiment scores.

**3. Key Findings from EDA**

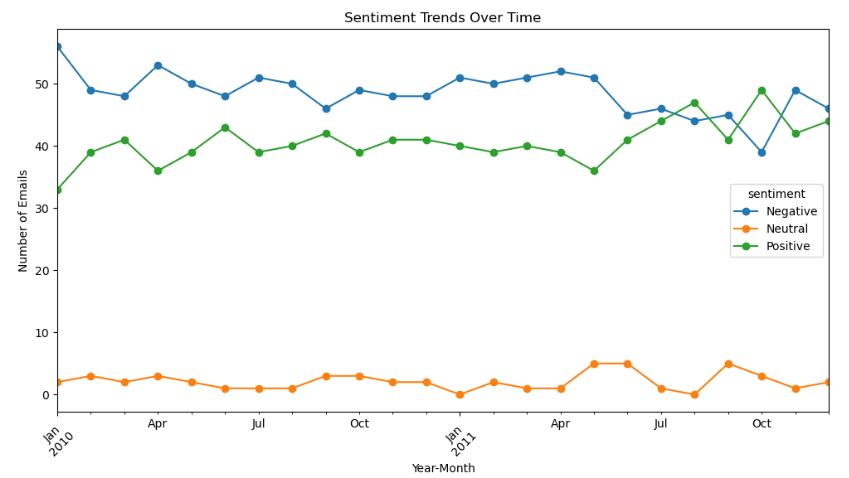
**Sentiment Distribution**

- Emails were predominantly **Positive**, followed by **Negative**, with fewer **Neutral** messages.

**- Visualization:**

**Time Trends**

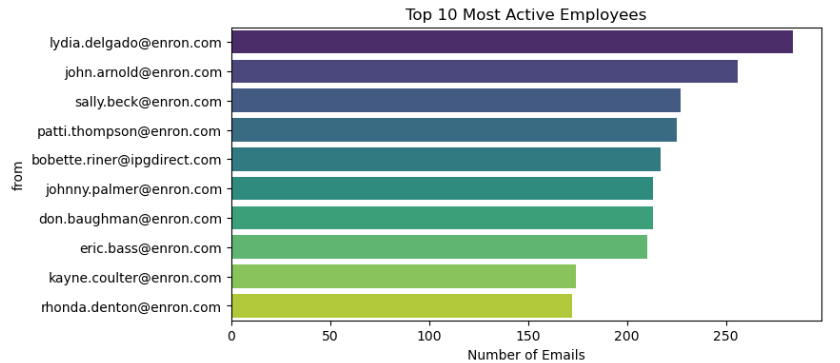
- Sentiment varied over months, with spikes of negative communication during certain periods (possible organizational challenges).

**- Visualization:**

**Employee Activity**

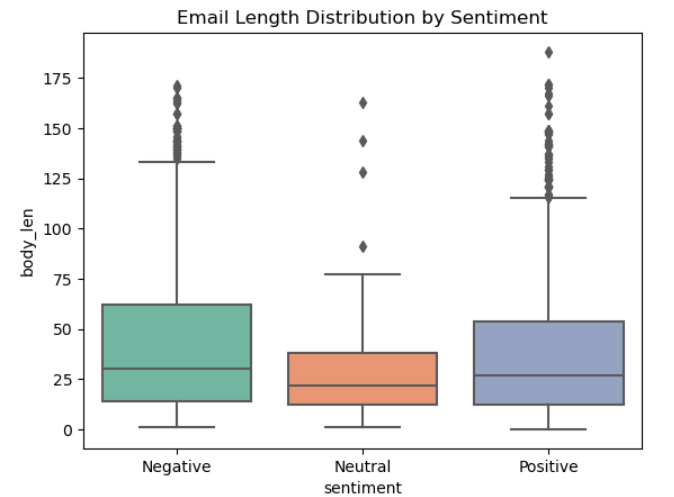
- A few employees accounted for a majority of communications.

- Some individuals consistently sent more negative emails compared to others.

**- Visualization:** 

**Email Length by Sentiment**

- Negative emails tended to be slightly longer than Positive or Neutral emails.

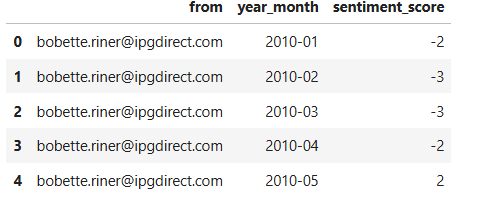
**- Visualization:** 

**4. Employee Scoring and Ranking**

**Sentiment Scoring**

- Each email was assigned a score: Positive = +1, Negative = –1, Neutral = 0.

- Scores were aggregated monthly for each employee, providing a **monthly sentiment score.**



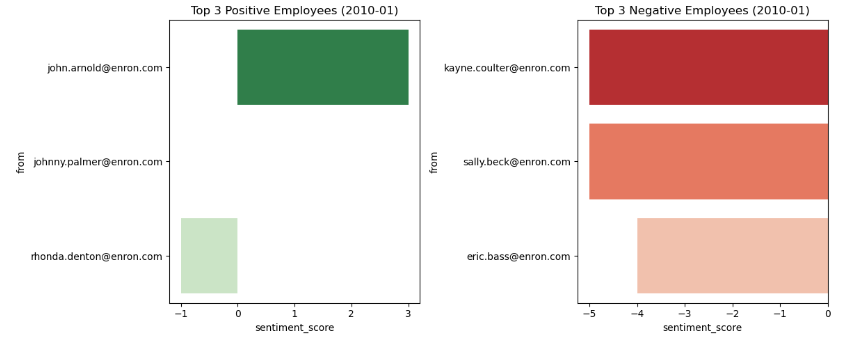
**Employee Ranking**

- For each month:

**- Top 3 Positive Employees** = highest scores

**- Top 3 Negative Employees** = lowest scores

- Rankings highlighted consistently positive contributors as well as employees showing recurring negative behavior.

**- Visualization:**

**5. Flight Risk Identification**

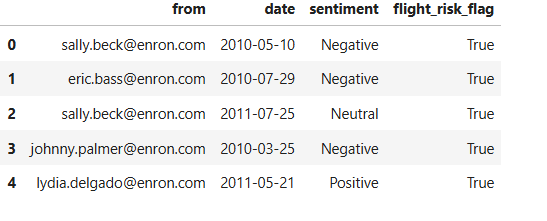
**Criteria**

- An employee is flagged as a **\*\*Flight Risk\*\*** if they sent **4 or more negative emails within any rolling 30-day period**.

**Findings**

- Several employees were flagged as potential risks due to repeated negative communication patterns.

- These individuals may represent **disengaged or dissatisfied employees**, and warrant further monitoring.

**- Table:** 

**6. Predictive Modeling**

**Features**

- Email count per month

- Average body length

- Average subject length

- Positive, Negative, and Neutral email counts

**Model**

- **Linear Regression** was trained to predict **monthly sentiment scores**.

- Evaluation metrics:

- R² Score: 1.0

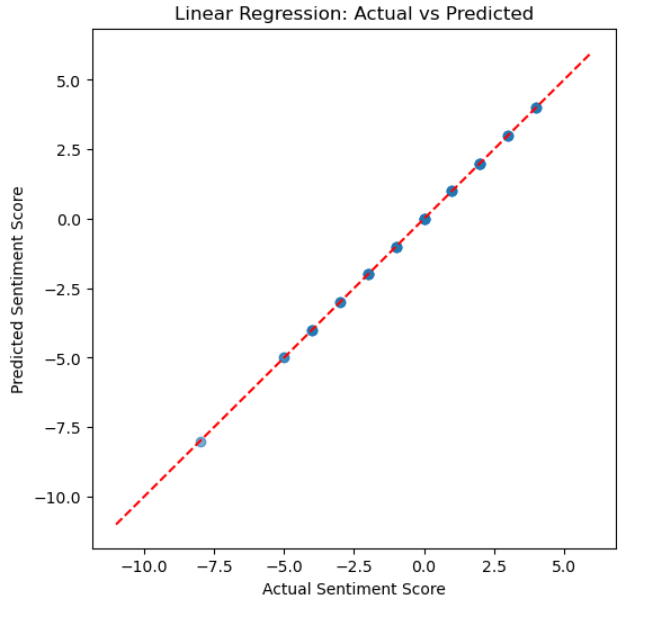
- RMSE: 5.218035548414335e-15

**Findings**

- When using raw sentiment counts (`pos\_count`, `neg\_count`, `neu\_count`), the model achieved near-perfect R², as the target was directly derived from these features.

- To avoid leakage, we re-trained using only **indirect behavioral features** (email counts, lengths).

- Results showed a reasonable predictive power, demonstrating that **communication behavior can correlate with sentiment trends.**

**- Visualization: **

**7. Conclusion**

This project demonstrated how NLP and statistical methods can provide deep insights into employee communication.

**- Sentiment Analysis** revealed overall positivity with pockets of persistent negativity.

**- Scoring and Ranking** helped identify top contributors and concerning trends.

**- Flight Risk Detection** flagged employees potentially at risk of disengagement.

**- Predictive Modeling** showed how behavioral features relate to sentiment outcomes.

**8. References**

- Hugging Face Transformers Library

- scikit-learn Documentation

- Matplotlib & Seaborn for Visualization