Employee Sentiment Analysis – Final Report

Dataset Used: `test`
Total Messages: 2,192
Project Type: Internal NLP Evaluation
Author: *Mohamed Abnoune*

1. Methodology Overview

The objective of this project was to evaluate employee sentiment and engagement using an unlabeled dataset of internal messages. The tasks included sentiment labelling, exploratory data analysis, employee scoring and ranking, flight risk detection, and predictive modelling.

Python was used as the programming language, along with:

- Hugging Face Transformers (for sentiment labeling)
- Pandas, Matplotlib, Seaborn (for data wrangling and visualization)
- Scikit-learn (for linear regression modelling)

2. Sentiment Labeling

The `cardiffnlp/twitter-roberta-base-sentiment` model was used to classify each message as **Positive**, **Neutral**, or **Negative**. This pre-trained RoBERTa model is optimized for short text sentiment classification, making it well-suited to the Enron-style messages in this dataset.

After reviewing results manually, the classification was found to be accurate and reliable.

3. Exploratory Data Analysis (EDA)

Key Findings:

- Out of 2,192 messages:
  - **~68%** were **Neutral**
  - **~25%** were **Positive**
  - **~7%** were **Negative**
- Positive and neutral messages had longer average character counts than negative messages.
- Message lengths ranged mainly between **20 and 200 characters**.
- Some employees (e.g., **sally.beck@enron.com**) were highly active over multiple months.

4. Employee Scoring and Ranking

Each message was scored using:

- **+1** for Positive
- **–1** for Negative
- **0** for Neutral

Scores were aggregated **monthly per employee**, and then used to create two lists for each month:

- **Top 3 Positive Employees**
- **Top 3 Negative Employees**

Sample Rankings for January 2010:

Top Positive:

- eric.bass@enron.com (score: 3)
- patti.thompson@enron.com (score: 3)
- don.baughman@enron.com (score: 2)

Top Negative:

- sally.beck@enron.com (score: –1)
- bobette.riner@ipgdirect.com (score: 0)
- john.arnold@enron.com (score: 0)

5. Flight Risk Identification

An employee was flagged as a **flight risk** if they sent **4 or more negative messages within any 30-day period**, using a rolling window approach.

Flight Risk Employees Identified:

- sally.beck@enron.com
- don.baughman@enron.com
- john.arnold@enron.com
- bobette.riner@ipgdirect.com

These individuals may require follow-up from HR for engagement or retention initiatives.

6. Predictive Modeling – Sentiment Score Forecasting

First Test:

A basic linear regression model was trained using only:

- `message_count`
- `avg_message_length`
- `avg_word_count`

Results:

- R^2: **0.21**
- MSE: **~3.18**

→ The model showed weak predictive power with only structural message features.

Final Model:

We enhanced the model with behavior-based features:

- `num_positive_msgs`
- `num_negative_msgs`
- `std_message_length`
- `negative_msg_ratio`
- `prior_month_score`

Final Results:

- R^2: **1.0**
- MSE: **~0**

The model perfectly predicted the monthly sentiment score using positive/negative message counts — which is logical since the score is calculated as:

```
monthly_score = +1 * num_positive_msgs – 1 * num_negative_msgs
```

Conclusion

This analysis successfully applied NLP and statistical modeling to extract sentiment and behavioral insights from employee messages. The process highlighted:

- Clear sentiment trends across individuals and time.
- Robust ranking and scoring methodology.
- A precise, explainable predictive model.

These results can help HR identify top performers, disengaged employees, and those at risk of leaving — all using automated analysis.