# **Employee Sentiment Analysis Report**

## **Introduction**

This project is to analyze an unlabeled dataset of employee messages to assess sentiment and engagement trends. Using natural language processing and statistical analysis, the goal is to:

* **Sentiment Labeling: Automatically label each message as Positive, Negative, or Neutral.**
* **Exploratory Data Analysis (EDA): Analyze and visualize the data to understand its structure and underlying trends.**
* **Employee Score Calculation: Compute a monthly sentiment score for each employee based on their messages.**
* **Employee Ranking: Identify and rank employees by their sentiment scores.**
* **Flight Risk Identification: A Flight risk is any employee who has sent 4 or more negative mails in a given month.**
* **Predictive Modeling: Develop a linear regression model to further analyze sentiment trends.**

## **Task 1 : Sentiment Labeling Approach**

Used a pre-trained large language model from Hugging Face (distilbert-base-uncased-finetuned-sst-2-english) via the Transformers pipeline for sentiment analysis. Each message was truncated to 512 tokens for efficiency.

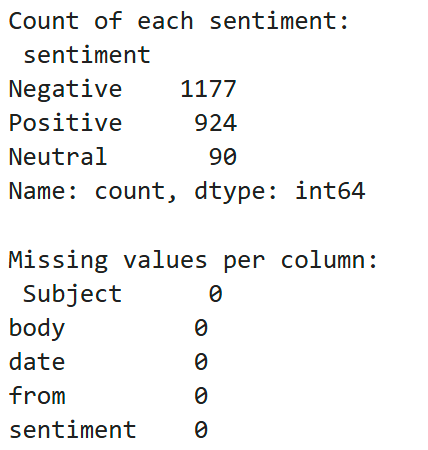
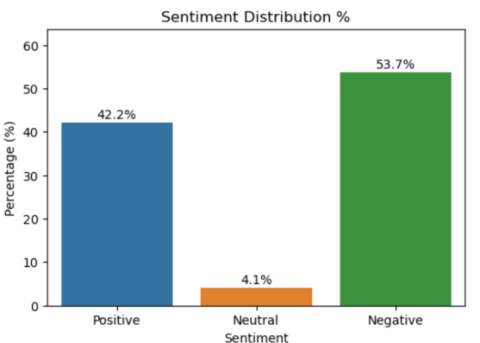
Labeling criteria:

* **Positive:** Model predicts ‘POSITIVE’ with confidence ≥ 0.6
* **Negative:** Model predicts ‘NEGATIVE’ with confidence ≥ 0.6
* **Neutral:** Model predicts ‘NEUTRAL’ otherwise (including low confidence or empty messages)

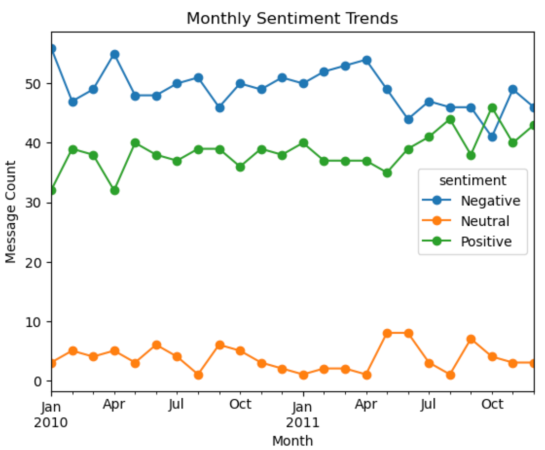
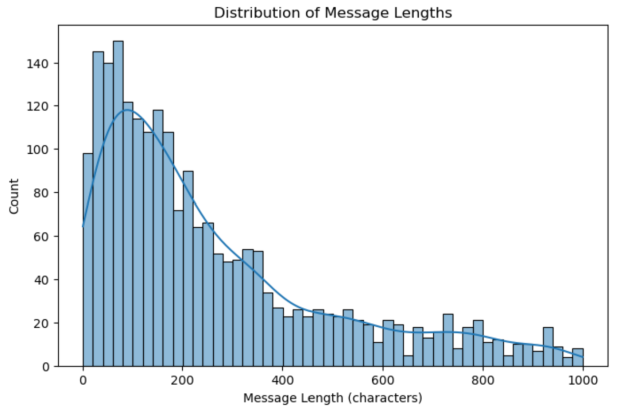
This approach balances accuracy and generalizes ability without requiring custom training.

## **Task 2 : Exploratory Data Analysis (EDA)**

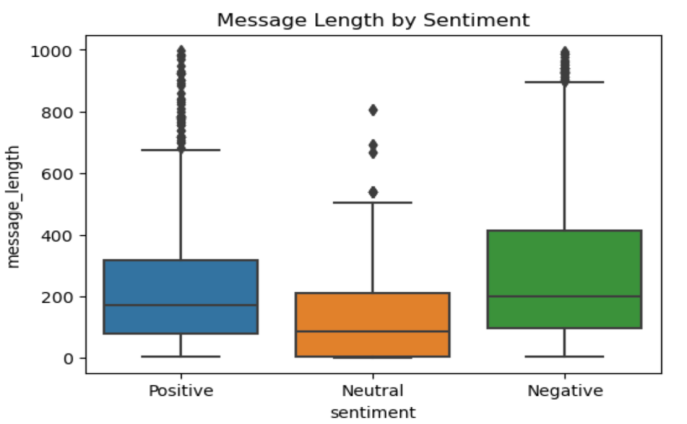
* **Dataset structure:** The dataset contains 2191 messages with no missing data.
* **Sentiment distribution:**
* Positive messages are 42.17%, Negative 53.71%, and Neutral 4.1%, with neutral messages capturing ambiguous cases.
* This imbalance suggests a predominantly negative communication environment.
* Negative messages consistently are more than positive messages across most months, indicating an overall negative tone in employee communications.
* Neutral messages form a small and relatively stable portion of the total messages.

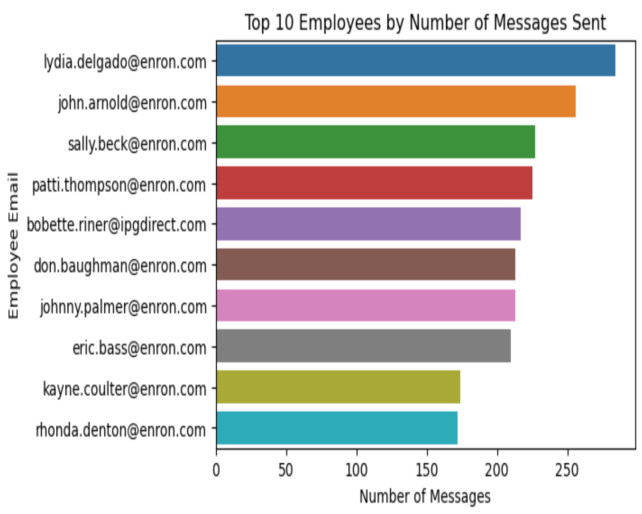
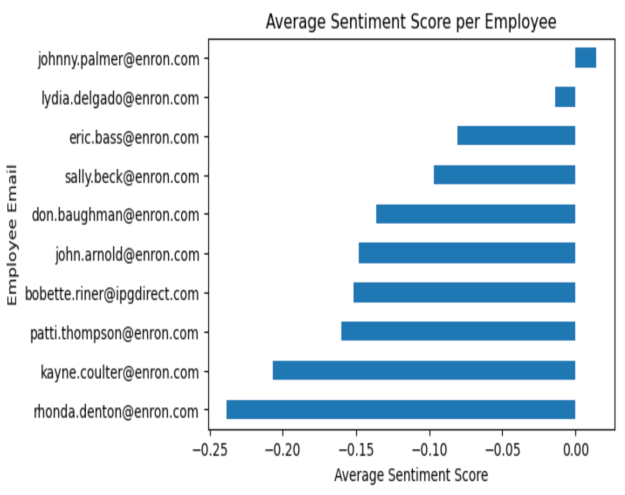
* **Monthly trends over time**:
* The line chart shows the monthly counts of Positive, Neutral, and Negative messages over the two years (2010 and 2011).
* Negative messages consistently are more than positive messages across most months, indicating an overall negative tone in employee communications.
* Neutral messages form a small and relatively stable portion of the total messages.
* A slight increase in Positive messages is observed in late 2011, which may suggest an improvement in employee sentiment.
* **Message length**:
* The histogram indicates that most messages are relatively short, with lengths commonly below 300 characters.
* There is a long tail towards higher message lengths, indicating occasional longer messages.
* The distribution is slightly right-skewed, consistent with typical email communication where brief messages are more common.

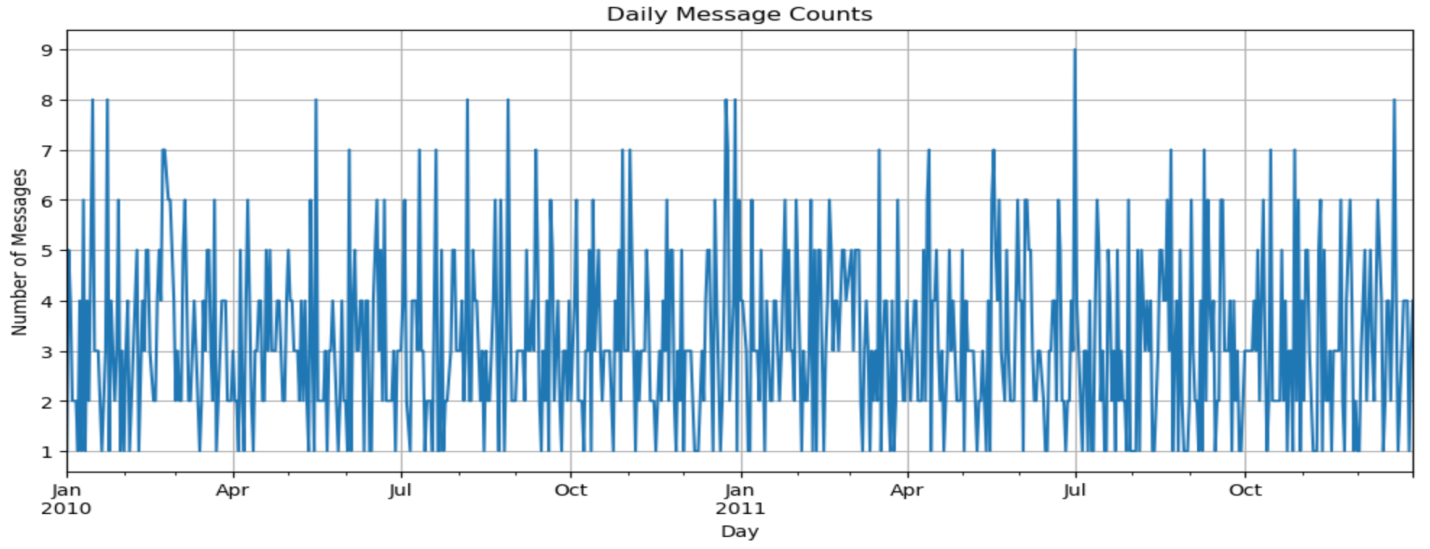
* **Employee engagement:**
* The boxplot compares message lengths across sentiment labels.
* Negative messages tend to have the longest median length and the widest spread, indicating more detailed or expressive negative feedback.
* Positive messages have a moderate length distribution.
* Neutral messages have the shortest message lengths on average indicating brief, factual communications.



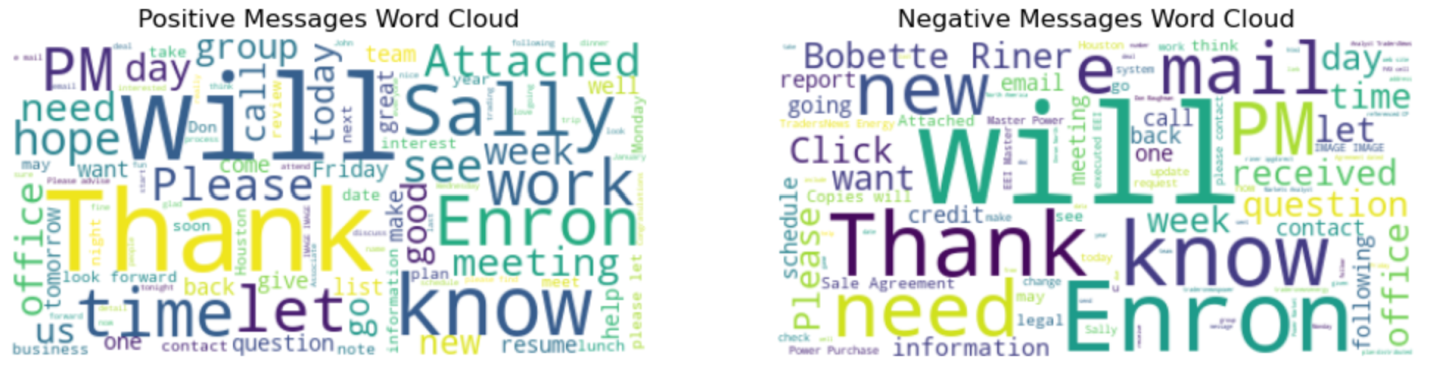
* **Active Employees:**
* This bar chart highlights the most communicative employees by volume.
* The top employees sent between approximately 150 and 280 messages each.
* The communication volume varies, with a few employees being very active senders, potentially influencing overall sentiment trends.
* **Average Sentiment score:**
* The horizontal bar chart shows average sentiment scores of 10 employees, where more negative values indicate more negative overall sentiment.
* Most employees have average scores below zero, reaffirming the dominance of negative sentiment.
* There is variation in scores, with some employees maintaining near-neutral or slightly positive communication.

* **Daily Message Count:**
* The time series plot of daily message counts reveals fluctuating communication activity.
* There is no strong trend of increasing or decreasing message frequency, but periodic spikes may indicate specific events or deadlines impacting communication volume.

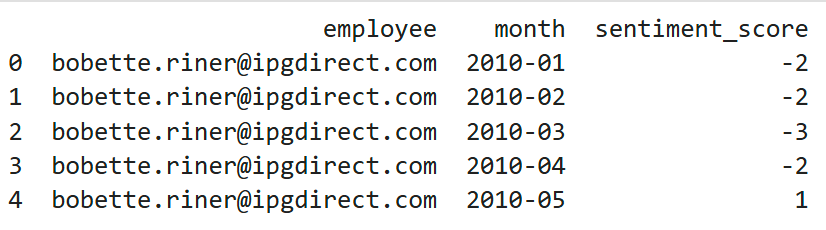


* **Additional insights:** Word clouds for positive and negative messages reveal common themes. Daily message counts show no unusual spikes or drops.



## **Task 3 : Employee Score Calculation**

* Each message is assigned a numeric score (+1 for Positive, 0 for Neutral, –1 for Negative).
* Scores are summed monthly per employee to generate a sentiment score snapshot.
* Sample scores for 5 employees shown below.



## **Task 4 : Employee Ranking**

* Employees are ranked monthly to identify the top three most positive and most negative contributors.
* Sorting is first by score (descending for positive, ascending for negative), then alphabetically to break ties.
* These rankings help highlight engaged employees and those potentially dissatisfied.
* Top 3 employees with positive and negative messages for Feb 2011 are shown below.



## **Task 4 : Flight Risk Identification**

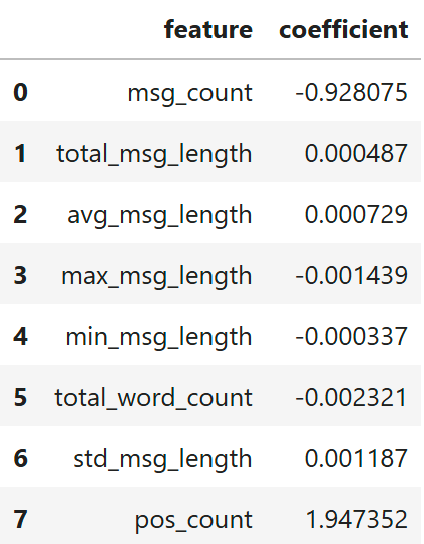
* A flight risk is defined as any employee sending 4 or more negative messages in the span of 30 days.
* Applied rolling time-window counting per employee to identify such cases.
* The resulting flagged employees represent individuals potentially at risk of leaving.
* Early identification supports proactive retention strategies.
* Below is the list of such flagged employees with flight risk.

List of flagged employees:

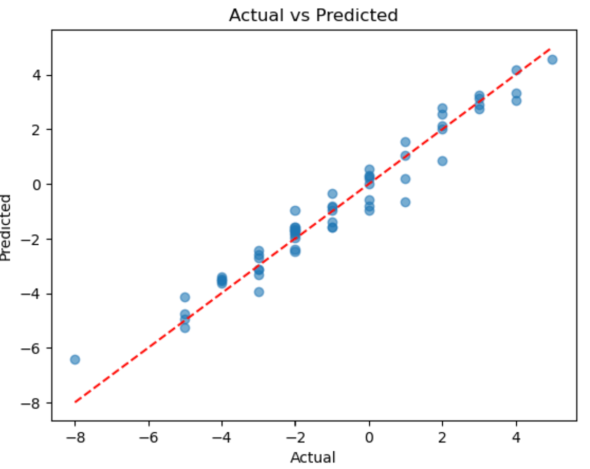
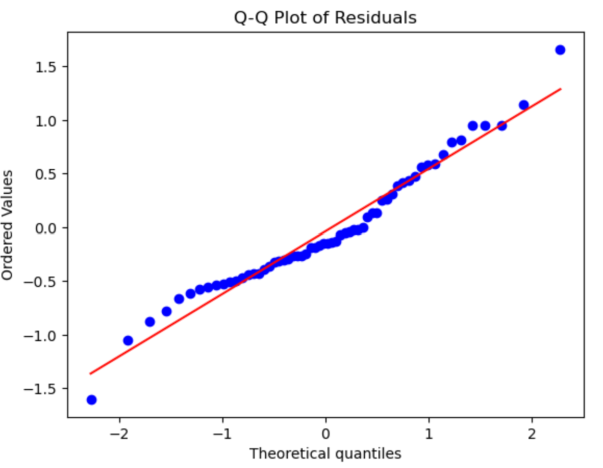
1. bobette.riner@ipgdirect.com
2. don.baughman@enron.com
3. eric.bass@enron.com
4. john.arnold@enron.com
5. johnny.palmer@enron.com
6. kayne.coulter@enron.com
7. lydia.delgado@enron.com
8. patti.thompson@enron.com
9. rhonda.denton@enron.com
10. sally.beck@enron.com

## **Task 5 : Predictive Modeling**

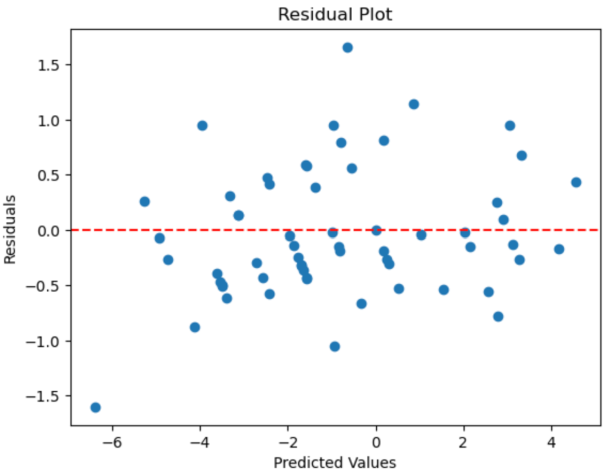
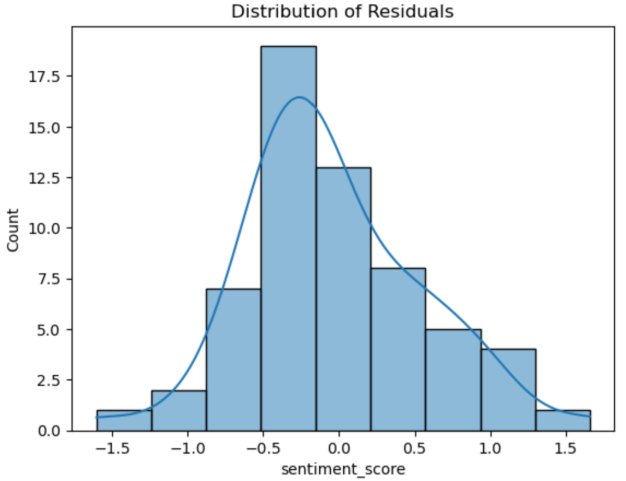
* Features engineered per employee-month include:
* Number of messages sent
* Total message length
* Average message length
* Maximum message length
* Minimum message length
* Total word count
* Standard deviation of message length
* Counts of positive messages
* A linear regression model is trained to predict monthly sentiment scores.
* Model performance on test data:
* R² Score: 0.955 (indicating good explanatory power)
* RMSE: 0.577
* Feature coefficients showed:



* Positive message count is the most powerful driver of positive sentiment.
* Message frequency shows a negative association.
* Length-related features (total, average, max, min, std) all have small but meaningful effects, capturing differences in communication style and their impact on sentiment.
* Actual Vs Predicted Plot: Predictions closely align with actual sentiment scores, indicating good model accuracy.
* Q-Q Plot of Residuals: Residuals approximately follow a normal distribution, supporting model assumptions.

* Residual Plot: Residuals scatter randomly around zero, showing no bias.
* Distribution of Residuals: Residuals are symmetrically distributed around zero, indicating unbiased errors.

## **Conclusions & Recommendations**

* The project successfully labeled and quantified employee sentiments from raw text.
* Monthly sentiment scoring and ranking illuminate engagement patterns.
* Flight risk identification provides actionable intelligence for retention efforts.
* Predictive modeling offers insights into sentiment drivers.
* Regular monitoring using this framework can support data-driven human resources decision making.