**Employee Sentiment Analysis – Final Report**

**Task 1: Sentiment Labeling**

**Approach and Methodology**

To assess employee sentiment from messages in the test dataset, we used a transformer-based language model to classify the emotional tone of each message. Our approach involved the following steps:

1. **Model Selection:**
   * We selected the cardiffnlp/twitter-roberta-base-sentiment model from Hugging Face's Transformers library.
   * This model was trained on a large corpus of Twitter data and is capable of classifying messages into three sentiment categories: **positive**, **neutral**, and **negative**.
   * It is well-suited for informal and casual text — making it ideal for analyzing internal employee communications, which may resemble social media text in tone.
2. **Data Loading and Preprocessing:**
   * Loaded the dataset test.csv into a Pandas DataFrame.
   * Combined the Subject and body columns into a new column named message to create a unified text field for analysis.
   * Preprocessing steps included:
     + Removing empty or NaN messages.
     + Ensuring text fields were proper strings.
     + Trimming long messages to 512 characters (maximum input for most transformer models).
     + Cleaning URLs and redundant spaces using regular expressions.
3. **Sentiment Analysis with Batched Inference:**
   * To optimize performance, sentiment predictions were processed in batches of 64 messages at a time.
   * The pipeline() method from Hugging Face was used to load and apply the sentiment analyzer.
   * Sentiment labels (LABEL\_0, LABEL\_1, LABEL\_2) were mapped to human-readable strings: **negative**, **neutral**, and **positive**.
   * Each message's sentiment was stored in a new column sentiment in the DataFrame.
4. **Output Storage:**
   * The augmented dataset, including the sentiment labels, was saved as a new CSV file sentiment\_df.csv.

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

**Approach and Methodology**

The goal of this task was to understand the structure and patterns in the sentiment-augmented employee message dataset (sentiment\_df.csv). Our methodology included the following steps:

1. **Data Structure Examination:**
   * Checked for missing values across all columns.
   * Inspected overall dataset shape and column data types.
   * Reviewed summary statistics to identify high-level characteristics (e.g., most frequent senders, subject lines, dates).
2. **Sentiment Label Distribution:**
   * Analyzed the proportion of each sentiment class (positive, negative, neutral).
   * Identified sentiment imbalance which may affect downstream modeling.
3. **Temporal Trends:**
   * Converted the date column to datetime format.
   * Grouped messages by month and plotted sentiment trends to observe changes over time.
4. **Behavioral Patterns and Anomalies:**
   * Explored sentiment proportions by individual senders.
   * Investigated subject line patterns across sentiment categories.
   * Identified repeated content or missing subject patterns.

**Key Findings from EDA**

**1. Dataset Structure and General Statistics**

* **Total Emails:** 2,191
* **Most Common Subject Line:** "(No Subject)" (141 messages)
* **Top Sender:** lydia.delgado@enron.com (284 messages)
* **Most Active Date:** July 1, 2011 (9 messages sent)
* **Column Types:** All columns were of object type, requiring conversion for date analysis.

**2. Sentiment Label Distribution**

* **Neutral:** 65.3% of messages (1,430 emails)
* **Positive:** 28.1% (616 emails)
* **Negative:** 6.7% (145 emails)

This indicates a significant class imbalance — most messages were neutral in tone, while negative sentiments were relatively rare.

**3. Sentiment Trends Over Time**

A monthly sentiment trend line plot revealed:

A graph of different colored lines

AI-generated content may be incorrect.

* **Neutral sentiment** consistently dominated across all months, typically ranging from **50 to 68 messages/month**.
* **Positive sentiment** showed an **upward trend**, increasing from ~20 to ~30 messages/month over the year.
* **Negative sentiment** remained consistently low, seldom exceeding 10 messages/month but with small **spikes in June and December**.

These patterns may reflect cyclical organizational factors (e.g., end-of-quarter reviews, holidays, or reporting periods).

**4. Sentiment by Sender**

We analyzed the top 10 most active senders and their sentiment distribution:

A graph with green and blue bars

AI-generated content may be incorrect.

* **Kayne Coulter** and **Bobette Riner** had the **highest proportion of negative messages (~9%)**, suggesting possible concern or dissatisfaction in tone.
* **Eric Bass** had the **lowest negative rate (3.8%)** and a relatively **high positive rate (30%)**, reflecting a more positive or optimistic communication style.
* Despite variation, **neutral sentiment remained the dominant tone** across all employees.

This may indicate **personality-driven or role-related communication patterns** among staff.

**5. Common Subject Lines by Sentiment**

An analysis of the most common subject lines by sentiment class revealed:

A graph with colorful squares and text

AI-generated content may be incorrect.

* **Neutral:** Procedural or administrative topics such as "Interview Schedule" and "Master Power Contracts".
* **Positive:** Congratulatory or celebratory themes including "Congratulations" and "Bearkadette Schedule".
* **Negative:** Often issue-related, such as "MISO proposed $1000 penalty" and "Out of the Office".

Subjects like "(No Subject)" and "Re:" were prevalent across all sentiment types, indicating that reply chains or undeclared topics are common in employee communications.

**Task 3: Employee Sentiment Score Calculation**

**Approach and Methodology**

The objective of this task was to quantify each employee's monthly communication sentiment by assigning numeric scores to their messages. This allowed us to track how an employee’s sentiment fluctuated over time. The methodology was as follows:

1. **Scoring Criteria:**
   * **Positive sentiment:** +1
   * **Negative sentiment:** -1
   * **Neutral sentiment:** 0
2. **Implementation Steps:**
   * Mapped each message’s sentiment label to a numerical score using a predefined dictionary.
   * Parsed the date column to extract the corresponding **year-month period** for each message.
   * Grouped the data by **employee email** and **month**, then summed the sentiment scores to get each employee's **monthly sentiment score**.
   * Renamed and sorted the resulting DataFrame for easier readability and analysis.

**Employee Sentiment Score Calculation Details**

* **Unique Employees Analyzed:** 10 (based on the from field)
* **Date Parsing:** Ensured all message dates were properly converted using pd.to\_datetime with error coercion for consistency.
* **Period Extraction:** Created a new month column by extracting the month and year from each message’s timestamp.
* **Grouping and Scoring:** Used groupby(['from', 'month']) followed by a sum() on the sentiment score column to compute each employee’s total sentiment per month.

**Key Insights**

* The **monthly sentiment score** offers a **quantifiable measure of employee tone over time**.
* Positive values suggest a predominantly encouraging or appreciative tone that month.
* Negative values may indicate dissatisfaction, stress, or interpersonal tension.
* Neutral scores may reflect procedural or emotionless communication.
* These scores form the basis for identifying **trends**, **communication shifts**, or **potential red flags** in employee behavior.

**Task 4: Employee Ranking**

**Objective**

The goal of this task was to identify and highlight top-performing and potentially at-risk employees based on their monthly sentiment scores. This insight can help HR or management identify trends in employee morale and communication tone over time.

Specifically, two ranked lists were created for each month:

* **Top Positive Employees**: 3 employees with the highest sentiment scores.
* **Top Negative Employees**: 3 employees with the lowest sentiment scores.

**Approach and Methodology**

1. **Input Data**: Used the monthly\_scores DataFrame from Task 3, which contains:
   * employee\_email
   * month
   * monthly\_sentiment\_score
2. **Ranking Logic**:
   * For each month, employees were **sorted first by sentiment score (descending)** and then **alphabetically by email** to break ties.
   * The **top 3 scores** formed the **positive list**; the **bottom 3 scores** formed the **negative list**.
   * A new column called rank\_type was added to distinguish between the two lists.
3. **Function Design**:
   * Created a function rank\_monthly\_employees(df) that extracts the top and bottom employees from a monthly slice of the data.
   * Applied this function across all unique months using groupby('month').apply(...) to construct a single consolidated rankings DataFrame.
4. **Presentation**:
   * Final rankings were displayed for each month in a clean tabular format showing:
     + rank\_type
     + employee\_email
     + monthly\_sentiment\_score

**Key Insights**

* This ranking system provides **early indicators of positive leadership or potential dissatisfaction**.
* Employees appearing frequently in the **Top Negative** list may warrant closer attention or outreach.
* **Consistently Top Positive** employees can be considered for recognition or leadership development.
* **Monthly granularity** allows for temporal trend monitoring—e.g., sudden shifts in sentiment ranking may correspond with organizational events or team changes.

**Task 5: Flight Risk Identification**

**Objective**

The aim of this task was to identify employees who may be at risk of leaving the company, based on the **frequency of negative communications** over time. Early identification of flight risks allows HR and management to proactively address concerns, intervene with support, or better understand underlying issues in team dynamics.

**Definition of Flight Risk**

An employee is flagged as a **flight risk** if they send **four or more negatively labeled messages** within **any rolling 30-day period**, regardless of the month or sentiment score magnitude.

**Methodology**

1. **Filter for Negative Sentiment**  
   The dataset was filtered to retain only emails labeled with a **‘negative’** sentiment.
2. **Sort for Temporal Analysis**  
   These records were sorted by sender (from) and date to prepare for rolling time window analysis.
3. **Apply Rolling 30-Day Check**  
   For each employee:
   * Extract the list of dates of their negative emails.
   * For each date, define a rolling 30-day window.
   * Check if at least 4 emails occurred within any such window.
   * If this condition is met, the employee is added to the **flight risk list**.
4. **Store and Present Results**
   * Flagged emails were compiled into a flight\_risk\_df DataFrame.
   * A flight\_risk = True column was added to mark these employees.
   * Results were displayed as a concise table.

**Key Findings**

The following employees were flagged as potential **flight risks**:

| **employee\_email** | **flight\_risk** |
| --- | --- |
| don.baughman@enron.com | True |
| bobette.riner@ipgdirect.com | True |
| patti.thompson@enron.com | True |
| john.arnold@enron.com | True |

* These employees had **4+ negative communications** within a **30-day window**, suggesting repeated expressions of dissatisfaction or stress.
* Some of these individuals also ranked highly in the **monthly negative sentiment rankings** (Task 4), reinforcing the concern.

An analysis of these individuals negative labeled messages over time:

A graph with different colored dots

AI-generated content may be incorrect.

* **Clusters of points** close together horizontally (same employee): This suggests the employee sent **multiple negative messages in a short period** — likely why they were flagged as flight risks.
* **Long gaps** between messages: Could indicate cooling off or recovery in sentiment.
* **High density in a short time**: May correlate with acute stress, conflict, or burnout period.
* **Consistent spread over time**: Indicates a more **chronic pattern** of negativity.

**Task 6: Predictive Modeling**

**Objective**

The goal of this task was to develop a predictive model to estimate an employee’s **monthly sentiment score** based on message-level communication patterns. This analysis enables early detection of mood shifts and supports proactive engagement by identifying the variables that most influence employee sentiment.

**Methodology**

**1. Feature Engineering**

To better understand employee message content and tone, several new features were created:

* word\_count: Number of words in each message.
* char\_count: Number of characters in each message.
* Sentiment flags:
  + is\_positive: True if the message is labeled as positive.
  + is\_negative: True if the message is labeled as negative.
  + is\_neutral: True if the message is labeled as neutral.

**2. Monthly Aggregation**

The dataset was grouped by employee and month. For each employee-month, we computed:

| **Feature** | **Description** |
| --- | --- |
| monthly\_sentiment\_score | Sum of sentiment scores for that month (**Target**) |
| message\_count | Total number of messages sent that month |
| avg\_word\_count | Average number of words per message |
| avg\_char\_count | Average number of characters per message |
| total\_words | Total number of words that month |
| total\_chars | Total number of characters that month |
| positive\_msg\_count | Number of positive messages |
| negative\_msg\_count | Number of negative messages |
| neutral\_msg\_count | Number of neutral messages |
| positive\_ratio | Share of messages labeled positive |
| negative\_ratio | Share of messages labeled negative |
| neutral\_ratio | Share of messages labeled neutral |

**3. Data Cleaning & Outlier Removal**

* The target variable (monthly\_sentiment\_score) was z-score standardized.
* Outliers beyond ±3 standard deviations were removed to improve model robustness.

**4. Model Preparation**

* **Target Variable**: monthly\_sentiment\_score
* **Independent Variables**:
  + message\_count
  + avg\_word\_count
  + avg\_char\_count
  + total\_words
  + total\_chars
  + positive\_ratio
  + negative\_ratio
  + neutral\_ratio
* The dataset was split using an 80/20 train-test split.
* Features were standardized using StandardScaler.

**5. Model Training & Evaluation**

* **Model**: Linear Regression
* **Evaluation Metrics**:
  + **Mean Squared Error (MSE)**: Measures average squared prediction error.
  + **R-squared (R²)**: Proportion of variance in sentiment explained by the model.

**Results**

* **MSE**: **1.54**
* **R²**: **0.61**

**Interpretation**

* The MSE of 1.54 indicates relatively low average prediction error.
* An R² of 0.61 suggests that **61% of the variance** in monthly sentiment scores can be explained by the engineered features.

**Model Insights: Feature Importance**

| **Feature** | **Coefficient** | **Interpretation** |
| --- | --- | --- |
| total\_words | +2.81 | Strongest positive driver of sentiment — more communication correlates with higher sentiment. |
| avg\_char\_count | +1.22 | Longer average messages (in characters) may reflect higher engagement. |
| positive\_ratio | +0.67 | More positivity in messages increases sentiment score. |
| message\_count | +0.51 | More frequent messaging correlates with positive sentiment. |
| neutral\_ratio | -0.31 | Slight negative association with neutral messaging. |
| negative\_ratio | -0.69 | Negative messaging lowers the sentiment score. |
| avg\_word\_count | -1.33 | Surprisingly negative — long messages (by word count) may sometimes reflect complaints or rants. |
| total\_chars | -2.28 | High total character volume may correlate with detailed negative content or stress. |

**Residual Plot Analysis:**

**1. Residuals vs. Predicted Plot:**

A graph with blue dots and red line

AI-generated content may be incorrect.

* Residuals are symmetrically scattered around zero with no visible pattern.
* Suggests good model fit and no obvious violations of linearity.

**2. Actual vs. Predicted Plot:**

A graph with green and red dots

AI-generated content may be incorrect.

* Data points follow the diagonal reference line fairly closely.
* Confirms strong correlation between actual and predicted sentiment scores.