**Task 1: Sentiment Labeling**

**Objective**

Label each employee message with one of three sentiment categories: Positive, Negative, or Neutral.

**Chosen Approach:**

For labeling sentiment in employee messages, we used the cardiffnlp/twitter-roberta-base-sentiment model from Hugging Face. This is a RoBERTa-based transformer model pre-trained on Twitter data and fine-tuned for sentiment classification. It is designed to classify short, informal text into one of three categories: Positive, Neutral, or Negative.

We chose this model because employee emails—like tweets—are often brief, casually written, and sentiment-rich. The model’s training on Twitter data made it well-suited for understanding such content, including informal language, emojis, abbreviations, and implied sentiment.

**How the Model Labels Sentiment:**

The model takes the input text (in our case, the email body) and outputs one of the following labels:

* LABEL\_0 → Negative
* LABEL\_1 → Neutral
* LABEL\_2 → Positive

We mapped these labels to meaningful names and added them as a new column (Sentiment) in our dataset.

**Why This Approach Works for Our Project:**

* The model gives a clean three-way classification that directly aligns with our project’s goals.
* It does not require additional training or labeled data.
* It is lightweight and easy to integrate via Hugging Face's pipeline, making it reproducible and efficient.
* It handled our dataset well and produced a reasonable sentiment distribution.

**Why We Did Not Use Other Models:**

We tested rule-based models like VADER and TextBlob, but they performed poorly in our scenario. Specifically:

* They labeled over 95% of messages as Neutral, missing important sentiment cues.
* They lack contextual understanding and are not suited for domain-specific or nuanced messages.
* They rely on predefined lexicons and simple scoring methods, which are insufficient for short, ambiguous, or technical communication.

In contrast, the RoBERTa-based model provided more accurate and balanced sentiment labeling.

**Reproducibility:**

The entire process—from loading the model to generating sentiment labels—is done using a few lines of code via Hugging Face's pipeline, ensuring that the approach is easily reproducible by any team member.

**Task 2: Approach and Methodology – Exploratory Data Analysis (EDA)**

**Objective**

Understand the structure, distribution, and trends in the dataset through thorough exploration.

**Approach & Methodology**

The goal of this task was to gain a comprehensive understanding of the sentiment-labeled email dataset by examining distribution patterns, time-based trends, and employee-level sentiment behavior.

**The following steps were carried out:**

1. **Data Preparation:**
   * Loaded the dataset test\_with\_roberta\_sentiment.csv.
   * Converted the date column into datetime format to facilitate temporal analysis.
   * Extracted a month column from the date for monthly aggregation.
2. **Sentiment Distribution Analysis:**
   * Visualized the proportions of Positive, Neutral, and Negative emails using bar plots and pie charts.
   * Analyzed message length across sentiment classes with a boxplot, to investigate whether emotional messages are typically longer or shorter.
3. **Time-Based Trend Analysis:**
   * Aggregated messages by month and visualized the number of emails per sentiment using line plots and stacked bar charts.
   * Identified months with unusual sentiment patterns (e.g., spikes in positivity or negativity).
4. **Employee-Level Sentiment Patterns:**
   * Grouped emails by sender (from) and sentiment category.
   * Created bar plots to visualize total sentiment counts per employee.
   * Tracked individual employees' sentiment trends over time using line and bar charts.
5. **Linguistic Analysis:**
   * Generated word clouds for each sentiment class (Positive, Neutral, Negative) to identify commonly used words and tone variations in email content.

This staged EDA approach—starting from high-level distribution to detailed sender-specific analysis—allowed for a deep dive into both content and behavioral patterns.

**Key Findings**

1. **Sentiment Distribution**
   * Neutral messages dominated the dataset (~68.7%), followed by Positive (24.6%) and Negative (6.8%).
   * This suggests that most emails were factual or informational in tone.
2. **Sentiment Trend Over Time**
   * December 2010 and June 2010 recorded the highest number of Negative messages.
   * December 2010 also showed a spike in Positive messages, while Neutral messages dropped, indicating emotionally charged communication during that month.
3. **Employee-Level Patterns**
   * Bobette Riner sent the most Negative messages.
   * Johnny Palmer had the most Positive messages, implying a more optimistic tone.
   * Lydia Delgado sent the highest number of Neutral messages, reflecting a neutral/informational style.
4. **Top 5 Employees by Negative Messages**

The following individuals had the highest number of Negative-labeled emails:

* + Bobette Riner
  + John Arnold
  + Sally Beck
  + Patti Thompson
  + Lydia Delgado

**Task 3: Approach and Methodology – Employee-Level Sentiment Analysis**

**Objective:**

Compute a monthly sentiment score for each employee based on their messages.

**Approach:**

To understand how individual employees express sentiment in their internal communications, we transformed qualitative sentiment labels into quantitative scores. By aggregating these scores monthly and visualizing them using a heatmap, we could detect sentiment consistency, fluctuations, and outliers across employees.

**Step-by-Step Methodology:**

1. **Sentiment Mapping:**
   * Converted sentiment labels to numeric values for aggregation:
     + Positive → +1
     + Neutral → 0
     + Negative → –1
   * This numerical approach enabled meaningful calculations such as monthly averages and totals.
2. **Employee Name Extraction:**
   * Extracted employee names from email addresses by splitting the string before the @ symbol.
   * Created a clean, human-readable identifier (Employee\_name) for grouping and visualization.
3. **Monthly Sentiment Aggregation:**
   * Grouped the data by both employee and month.
   * Aggregated sentiment scores to calculate the net sentiment per employee per month.
   * This showed whether an employee was generally positive, neutral, or negative in a given month.
4. **Heatmap Matrix Preparation:**
   * Reshaped the grouped data using a pivot table:
     + Rows = Employees
     + Columns = Months
     + Values = Aggregated sentiment scores
   * Missing values (i.e., months without communication) were filled with zeros to maintain visual consistency.
5. **Heatmap Visualization:**
   * Used Seaborn’s heatmap to visualize the data:
     + Blue for positive sentiment
     + Red for negative sentiment
     + White/light for neutral or no messages
   * This allowed for quick detection of:
     + Consistent communication tone per employee
     + Sudden sentiment shifts
     + Behavioral trends and anomalies over time

**Conclusion**

* Johnny Palmer demonstrated consistently high positive sentiment, including:
  + 8 positive messages in June 2011
  + 7 in February 2011
  + 6 in January 2011
  + His sentiment remained positive in all observed months — a rare trend.
* Kayne Coulter had the most negative months, with at least five months showing a net negative sentiment score — possibly indicating disengagement or dissatisfaction.
* Most employees had moderate sentiment levels, with positive messages generally fewer than four per month.
* The heatmap effectively highlighted emotional consistency (or lack thereof) among employees, helping to identify:
  + Positive communicators
  + Employees with fluctuating or concerning sentiment patterns
  + Months with major sentiment shifts

**Task 4: Ranking of Positive and Negative Employees**

**Objective:**

Generate ranked lists of employees based on their monthly sentiment scores.

**Approach**

We utilized monthly aggregated sentiment scores (calculated in Task 3) and applied a ranking strategy to determine standout employee behavior in both positive and negative sentiment directions. This enabled identification of:

* Consistently positive communicators (possible cultural leaders)
* Repeatedly negative communicators (possible disengagement or conflict risks)

**Methodology**

1. **Ranking Employees by Monthly Sentiment:**
   * **For each month:**
     + Selected the Top 3 employees with the highest sentiment scores (most positive).
     + Selected the Bottom 3 employees with the lowest sentiment scores (most negative).
   * **Sorting was performed on the monthly sentiment score using:**
     + descending order for positive ranking.
     + ascending order for negative ranking.
2. **Data Grouping & Selection:**
   * The sentiment data was grouped by month, and top/bottom 3 employees were extracted per group.
   * This ensured that rankings were isolated to each month without cross-period interference.
3. **Visualization:**
   * **Two bar charts were created:**
     + One for Top 3 Positive Employees per Month
     + One for Top 3 Negative Employees per Month
   * **Distinct colors were assigned to different employees for easy comparison and trend spotting.**

**Conclusion**

The analysis showed that Johnny Palmer consistently demonstrated strong positive sentiment across all months, making him a clear outlier in positive communication. In contrast, Kayne Coulter frequently appeared among the most negative employees, suggesting recurring issues. While positive sentiment was concentrated among a few individuals, negative sentiment was more dispersed. These patterns can help identify both high-performing communicators and employees who may need further support or attention.

**Task 5: Identifying At-Risk Employees**

**Objective:**

Identify employees who are at risk of leaving based on their monthly sentiment scores.

**Approach:**

To identify at-risk employees, we first filtered the dataset to extract only the rows with a "Negative" sentiment. We then grouped this filtered data by employee name and email date to track negative communication patterns over time. Using a rolling 30-day window for each employee, we checked whether they had sent four or more negative messages within that period. If an employee met this condition, they were added to a set of potentially at-risk individuals. This approach helped flag employees showing repeated negativity in a short timeframe, which could indicate stress, disengagement, or dissatisfaction.

**Methodology:**

1. **Filtering Negative Messages:**
   * Filtered the dataset to include only emails labeled with the “Negative” sentiment.
   * This step isolated potentially concerning messages that may indicate dissatisfaction or frustration.
2. **Sorting by Employee and Date:**
   * Sorted the negative messages chronologically for each employee.
   * This made it possible to analyze sentiment patterns over time using a rolling window approach.
3. **30-Day Rolling Window Analysis:**
   * For each employee, applied a 30-day rolling window to check if they sent four or more negative messages within any such window.
   * This threshold was chosen to flag persistent negativity over a short time span, which may be a signal of distress, burnout, or disengagement.
4. **Flagging At-Risk Employees:**
   * If an employee met this condition, they were **flagged** as potentially **at risk**.
   * This check was repeated across the entire dataset for all employees.
5. **Result Compilation:**
   * Compiled a list of employees who triggered the rolling-window condition.
   * These individuals are considered for **further attention or HR intervention**.

**Key Findings:**

A total of **5 employees** were flagged as potentially at risk due to clustering of negative messages:

* **bobette.riner**
* **don.baughman**
* **john.arnold**
* **rhonda.denton**
* **sally.beck**

Each of these employees sent **four or more negative messages** within at least one **30-day window**.

**Task 6: Predictive Modeling – Approach, Methodology & Key Insights**

**Objective:**

Develop a linear regression model to analyze sentiment trends and predict sentiment scores using a variety of independent variables that may influence sentiment scores.

**Approach and Methodology:**

**1. Data Preparation and Feature Engineering**

* The raw communication dataset was enhanced with additional **textual features**:
  + word\_count: Number of words per message.
  + msg\_length: Character length of the message..
* Features were **aggregated monthly per employee**, including:
  + **Total / Average Message Length**
  + **Total / Average Word Count**
  + **Message Count per Month**
  + **Derived Metrics:**
    - sentiment\_per\_message = total sentiment / message count
    - sentiment\_per\_word = total sentiment / total words
    - interaction\_1 = total message length × average word count

These features were designed to capture not just the **volume** of communication, but also its **density** and **emotional intensity**.

**2. Target Variable:**

* The target was the **monthly sentiment score** (numeric), computed by summing individual sentiment values (+1 for positive, 0 for neutral, –1 for negative) per employee per month.

**3. Model Selection:**

* A **Linear Regression** model was selected for its simplicity and interpretability, suitable for understanding the linear relationship between communication patterns and sentiment scores.

**4. Data Splitting & Feature Scaling:**

* The dataset was split into **training (70%)** and **testing (30%)** sets.
* Input features were standardized using **StandardScaler** to ensure equal weighting and improve model performance.

**5. Model Training & Evaluation:**

* The model was trained on the training set and evaluated on the test set using:
  + **RMSE**: 1.2981 – shows average error magnitude
  + **MAE**: 0.9383 – interpretable average absolute error
  + **R² Score**: 0.5797-57% of the variance in sentiment scores is explained by the model
  + **Cross-Validated R²**: 0.5638 ± 0.1778

**Model Interpretation – Key Feature Impacts:**

| **Feature** | **Coefficient** |
| --- | --- |
| sentiment\_per\_message | **+1.5503** |
| message\_count | +1.4114 |
| total\_word\_count | +0.7792 |
| interaction\_1 | +0.4884 |
| total\_message\_length | –1.6305 |
| avg\_message\_length | –0.0273 |
| avg\_word\_count | –0.0363 |
| sentiment\_per\_word | **–0.6067** |

**Insights:**

* **sentiment\_per\_message** had the **strongest positive influence**, suggesting that a high emotional tone per message leads to higher overall sentiment.
* **sentiment\_per\_word** had a **strong negative effect**, indicating that longer or more verbose messages may dilute sentiment intensity.
* Features like **message count**, **word count**, and **message length** had smaller but positive effects.

**Conclusion:**

The predictive model offered valuable insight into how communication behaviors influence sentiment trends. With a respectable **R² score (~57%)**, it provides a solid baseline for forecasting sentiment.