**Final Report: Employee Sentiment Analysis Project**

**Task 1 – Sentiment Labeling**

**Objective:**

To label employee message content with a sentiment category: Positive, Neutral, or Negative, using a transformer-based NLP model.

**Approach & Methodology:**

* Used the pretrained ‘cardiffnlp/twitter-roberta-base-sentiment’ model via Hugging Face Transformers.
* Cleaned messages by removing email signoffs (e.g., “Best regards”), removing usernames from email body, and extra white spaces.
* Dropped empty messages post-cleaning.
* Used a sentiment analysis pipeline to predict labels and confidence scores.
* Added results as two new columns in the dataset:
  + predicted\_label
  + sentiment\_score

**Justification of Approach:**

I selected the ‘cardiffnlp/twitter-roberta-base-sentiment’ model because it is one of the most effective pretrained models available for sentiment analysis on short-form, informal text. Unlike some versions of BERT, which lack predefined sentiment labels or require fine-tuning, this model outputs directly interpretable sentiment categories (Positive, Neutral, Negative), making it well-suited for our needs.

For sentiment analysis, I evaluated multiple models, including TextBlob, VADER, and the transformer-based cardiffnlp/twitter-roberta-base-sentiment. Although VADER and TextBlob are easy to use, they are rule-based models trained primarily on social media text and lexicons. Their effectiveness in interpreting formal business emails is limited. They also require manually setting sentiment thresholds (e.g., −0.1 to 0.1 for neutral), which can introduce bias and vary significantly across domains.

While Roberta is also trained on tweets, it outperformed the others in this project. I manually reviewed 10–12 email examples and found that Roberta's predictions aligned more closely with human interpretation. Unlike VADER and TextBlob, it does not require threshold tuning, as it outputs sentiment labels directly (Positive, Neutral, Negative), reducing subjectivity and streamlining workflow.

As employee messages are usually formal and to the point, RoBERTa turned out to be the best fit for this task. It does a good job understanding context and gives clear sentiment labels that match how people would interpret the messages. Another big plus is that we don’t need to spend extra time training or labeling data—it works well right out of the box on real workplace communication.

**Results Comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Message** | **TextBlob** | **VADER** | **Roberta** | **Comment** |
| we were thinking papasitos (we can meet somewhere closer to you this time)  again at around 8 | Neutral | Neutral | Neutral | All models agree |
| I know we've made good progress with respect to morning reports for EGM.  However, I continue to receive most reports after 9:00 a.m., which is too  late into the trading day. Please let me know if there are resources you  need, so this aspect of EGM can run as smoothly as ENA gas | Neutral | Neutral | Positive | Roberta captures emotional tone better |
| approved. pls. process. thx. | Neutral | Positive | Neutral | Comparing, Roberta works better |

A total of 2,152 messages were analyzed after cleaning. Sentiment labels were successfully assigned to all.

**Reproducibility**

* The notebook uses a clear preprocessing function.
* A public transformer model was used with deterministic label mapping.
* The pipeline truncates long messages and handles missing data safely.

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

**Objective**

The objective of this task is to explore and understand the structure, sentiment distribution, and patterns within the dataset of employee messages. This step is important because it helps us understand the data before we move on to scoring employees or identifying any flight risks.

**Approach & Methodology**

To explore the dataset meaningfully, we performed the following steps:

1. **Dataset Inspection**
   * Loaded the sentiment-labeled dataset generated from Task 1.
   * Verified that the body, predicted\_label, and date columns contained valid, non-null data.
   * Confirmed that the dates were usable for time-based analysis
2. **Sentiment Distribution**
   * Counted how many messages were labeled as Positive, Neutral, or Negative.
   * Created a bar chart to see how the sentiments were distributed.
   * Most of the messages are Neutral, which shows that employees usually keep their emails professional and stick to the facts. That’s pretty normal in a work setting where people are expected to be formal. But the low number of Negative messages might also mean that some employees aren’t comfortable sharing complaints or frustrations openly.
3. **Time Trend Analysis**
   * Extracted year and month from the date field to track sentiment evolution over time.
   * Created separate plots for 2010 and 2011 showing the volume of each sentiment per month.
   * There was a small increase in positive sentiment in 2011, which could mean that morale was getting better—possibly due to leadership changes or efforts by HR. At the same time, negative sentiment went down, suggesting there may have been less conflict or frustration. It’s worth looking at these changes alongside any major events that happened in the company during that time.
4. **Word Clouds by Sentiment**
   * Generated separate word clouds for each sentiment category to visualize frequently used terms.
   * Found that:
     + Positive messages included words like great, hope, appreciate, and congratulations.
     + Neutral messages focused on work-related terms like schedule, meeting, and report.
     + Negative messages included words like problem, sorry, and unavailable, which may point to issues or frustrations.
5. **Message Length Analysis**
   * Computed message length in characters.
   * Created boxplots to compare message length across sentiment labels.
   * Positive messages tend to be longer, often including expressions of gratitude, appreciation, or collaboration. Negative messages are shorter and more direct, which may reflect urgency, frustration, or disengagement. This distinction can help in flagging potential early warning signals from concise, negatively toned emails.

**Key Visualizations**

The following charts were created and stored in the visualizations/ folder:

* **Bar Plot**: Distribution of sentiment labels (sentiment\_distributions.png)
* **Monthly Sentiment Trend (2010 & 2011)**: Sentiment breakdown over time (sentiment\_trend\_2010.png, sentiment\_trend\_2011.png)
* **Word Clouds**: For Positive, Neutral, and Negative messages (word\_cloud\_for\_positive\_emails.png, word\_cloud\_for\_neutral\_emails.png, word\_cloud\_for\_negative\_emails.png)
* **Boxplot**: Message length by sentiment (email\_length\_distribution\_by\_sentiment.png)

Each visualization was interpreted with supporting commentary, including not just what the chart shows, but why it matters — such as linking short negative messages to potential employee disengagement or tracking morale changes over time.

**Findings & Interpretation**

* The dataset is well-structured with no missing or invalid values.
* Most messages are neutral, which likely reflects the formal nature of workplace communication.
* The increase in positive communication in 2011 suggests improving morale or a more supportive culture.
* The kinds of words people use vary a lot depending on the tone of the message, which will be helpful when we start building models.
* Message length could be used as an additional feature in predictive modeling.

**Conclusion**

This analysis gave us a solid understanding of the employee messages and how sentiment is expressed over time. These findings will be especially useful as we start building models to score employee sentiment and identify who might be feeling disengaged or thinking about leaving the organization.

**Task 3: Employee Score Calculation**

**Objective**

In this task, we calculated a monthly sentiment score for each employee based on the messages they sent. These scores give us a way to track how an employee’s tone or mood changes over time and will help with employee ranking and identifying potential risks later on.

**Approach & Methodology**

Each message was assigned a sentiment score using the following rules:

* **Positive message** → +1
* **Neutral message** → 0
* **Negative message** → –1

These values were then aggregated per employee per month.

**Grouping and Aggregation Process**

To calculate monthly sentiment scores accurately, we performed the following steps:

1. **Date Conversion**
   * Converted the date column to datetime format.
   * Extracted year\_month as a period (e.g., 2011-06) using .dt.to\_period("M").
2. **Grouping**
   * Grouped messages by employee (from) and year\_month using:  
     df.groupby(['from', 'year\_month']).
3. **Aggregation**
   * Used .sum() to compute the total sentiment score for each employee for that month.
   * Reset the grouping monthly so that sentiment score does not carry over between months.
4. **Output Structure**
   * Final table contains:
     + from (employee email)
     + year\_month
     + monthly\_sentiment\_score

**Example Output Table**

|  |  |  |
| --- | --- | --- |
| **from** | **year\_month** | **monthly\_sentiment\_score** |
| **bobette.riner@ipgdirect.com** | 2010-01 | 1 |
| **eric.bass@enron.com** | 2010-02 | 3 |

**Visualization**

We plotted a monthly sentiment score trend line for one sample employee. This helps visualize their emotional engagement over time.

* A rising curve suggests increasing positivity or morale.
* A declining or negative curve may suggest stress, frustration, or disengagement.

**Key Insights**

* This scoring method captures sentiment fluctuations over time.
* It enables comparative analysis between employees.
* Employees with **consistently negative monthly scores** will be reviewed in Task 5 as potential flight risks or low-engagement individuals.

**Deliverables**

* monthly\_sentiment\_scores.csv: Score data per employee per month
* Visualization: [monthly\_sentiment\_scored\_trend\_for\_sally.beck@enron.com.png](mailto:monthly_sentiment_scored_trend_for_sally.beck@enron.com.png)

**Task 4: Employee Ranking**

**Objective**

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

**Requirements**

* Create two distinct lists:
  + Top Three Positive Employees: The three employees with the highest positive scores in a given month.
  + Top Three Negative Employees: The three employees with the lowest (most negative) scores in each month.
* Sort them first in descending order of sentiment and then in alphabetical order by email.
* Ensure the ranking is clearly derived from the monthly scores calculated in Task 3
* Present the rankings in a clear and organized format.

**Methodology**

The sentiment score for each employee was calculated on a monthly basis, as outlined in Task 3. Using these monthly scores, employees were ranked within each month.  
For each month:

* The top three employees with the highest sentiment scores were identified and labeled as 'Top Positive'.
* The three employees with the lowest scores were labeled as 'Top Negative'.

Sorting was performed in two stages:  
1. By descending sentiment score (or ascending for negatives).  
2. By alphabetical order of employee email address for consistent presentation.

**Visualization Summary**Bar charts were created to visualize the Top Positive and Top Negative employees for each month. These visuals help easily identify sentiment trends and potential outliers in behavior across time.

**Chart 1: Top Positive and Negative Sentiment Scores per Month**

* **Interpretation**: There’s a steady flow of strongly positive messages each month, which suggests that appreciation or recognition is a regular part of internal communication. On the other hand, negative messages are not only rare but also not very intense, showing that strong negativity doesn’t happen often.
* **Why it matters:** Sudden spikes in negative sentiment—even if few—could indicate emerging conflicts, dissatisfaction, or miscommunication. Keeping an eye on these months is important for early intervention.

**Chart 2: Top Positive and Negative Email Senders per Month**

* **Interpretation**: The catplot shows that specific individuals repeatedly contribute to highly positive or negative emails across several months. This suggests that individual communication style plays a significant role in tone.
* **Why it matters**: Consistently negative top senders might require check-ins to ensure support or prevent potential disengagement. Positive contributors could be internal morale boosters or natural leaders worth recognizing.

**Deliverables**

* Visualizations:
  + top\_positive\_and\_negative\_scores\_per\_month.png
  + top\_positive\_and\_Negative\_Email\_Senders\_per\_Month.png

**Task 5: Flight Risk Identification**

**Objective:**  
Identify employees who are at risk of leaving the organization based on their recent negative sentiment communication patterns.

**Methodology:**  
To flag potential flight risk employees, we used the following approach:

1. **Data Selection:** We filtered the dataset to focus only on emails with a Negative sentiment.
2. **Grouping:** These emails were grouped by the sender (employee email address).
3. **Rolling Window Detection:**
   * For each employee, the dates of negative emails were collected.
   * We used a sliding window approach over these dates to check whether 4 or more emails were sent within any rolling 30-day window.
   * If an employee met this criterion, they were flagged as "At Flight Risk".
4. **List Extraction:** A unique list of such employees was compiled.

**Code Summary:**  
The logic was implemented using a loop across each employee's negative email dates. A nested loop checked the duration between the start and end of each window, adding employees to a set if the threshold was met.

# Identify employees at flight risk

employees\_at\_flight\_risk = set()

for sender, group in df\_neg.groupby('from'):

dates = group['date'].tolist()

n = len(dates)

start = 0

for end in range(n):

while dates[end] - dates[start] > pd.Timedelta(days=30):

start += 1

if (end - start + 1) >= 4:

employees\_at\_flight\_risk.add(sender)

break

**Output:**  
We displayed:

* A list of employees who met the flight risk threshold.
* A bar chart showing the Top 10 employees by number of negative emails.
* A focused bar chart of flight-risk employees with their negative email count.

**Interpretation of Visualizations:**

The first chart identifies employees with the highest number of negative emails, signaling possible dissatisfaction or work-related stress. The second plot focuses on employees already flagged as flight risks, showing that many of them also exhibit strong negative sentiment patterns. These visual insights support a targeted employee engagement strategy and affirm the value of sentiment analysis in HR risk management.

**Insights:**

* Employees such as 'sally.beck@enron.com’, ‘bobette.riner@ipgdirect.com’, ‘john.arnold@enron.com', 'don.baughman@enron.com' were flagged as at risk.
* Visualizations reinforced that high-frequency negative sentiment is a strong signal for disengagement.

**Conclusion:**  
This step helped us proactively identify employees whose negative email patterns over a short time window indicate possible disengagement or dissatisfaction. These findings can be useful for HR intervention strategies to improve retention and support.

**Task 6: Predictive Modeling - Sentiment Score**

**Objective:**  
To develop a predictive model that estimates sentiment scores based on a set of email characteristics, helping us understand the drivers behind employee sentiment trends.

**Methodology:**

We designed and implemented a Linear Regression model and Random Forest Regressor using several email-level features to predict the sentiment score for each message.

1. **Feature Selection:**  
   The following features were selected as predictors for the respective reasons:

* **message\_length**: Total characters in an email. Longer emails may show more emotion or detail.
* **word\_count**: Number of words. Longer emails might reflect more thought or stronger feelings.
* **monthly\_message\_count**: How many emails someone sends in a month. More emails could mean higher stress or involvement.
* **avg\_message\_length\_monthly**: Average email length per month. Short messages may show routine; longer ones may show more effort or concern.

These features were selected based on their likely association with emotional expression and engagement, both of which can influence sentiment.

1. **Data Preparation:**
   * Cleaned and filtered the data to ensure valid numeric values
   * Created additional features like monthly\_message\_count and avg\_message\_length\_monthly through grouping
   * Ensured proper alignment between message-level and employee-month level attributes
2. **Model Training:**
   * Applied an 80/20 train-test split
   * Trained a linear regression model using sklearn.linear\_model.LinearRegression()
   * Trained a RandomForestRegressor(random\_state=42) from sklearn.ensemble
3. **Model Evaluation:**  
   We evaluated model performance using the following metrics:
   * **R² Score:** Helps explain how much variance is captured by the model.
   * **Root Mean Squared Error (RMSE):** Penalizes large prediction errors and gives more weight to outliers.
   * **Mean Absolute Error (MAE):** Easier to interpret as it reflects the average error magnitude.

**Results:**

**Linear Regression:**

* + **Training Set:**
    - R²: 0.048
    - RMSE: 0.139
    - MAE: 0.116
  + **Test Set:**
    - R²: 0.071
    - RMSE: 0.131
    - MAE: 0.110

**Random Forest Regressor:**

**Training Set**

* R²: 0.878
* RMSE: 0.050
* MAE: 0.039

**Test Set**

* R²: 0.087
* RMSE: 0.130
* MAE: 0.103

The Linear Regression model served as a baseline. Compared to linear regression, Random Forest achieved slightly higher R² and lower error rates, indicating that the relationship between the selected features and sentiment may be nonlinear. Although both models struggled to explain large variance in sentiment, the Random Forest’s performance suggests room for improvement with more complex models or content-level features.

**Visual Diagnostics:**  
To better understand model performance, two key plots were generated from linear regression:

* **Residual Plot** **–** to examine prediction errors (residual\_plot.png)
* **Actual Vs Predicted Plot** – to assess prediction alignment (actual\_and\_predicted\_sentiment\_scores.png)

**Insights:**

* While the model captures some variation in sentiment score, it leaves most variance unexplained.
* Random Forest suggests nonlinearities in the data that the linear model misses.
* This suggests that additional or more sophisticated features (e.g., sentiment intensity, topic modeling) may improve performance.
* Current linear model provides a baseline, but more advanced models (like XGBoost) could offer better results.

**Conclusion:**  
This task demonstrated how message metadata can be used to model sentiment. While the Linear Regression provided a starting point, the Random Forest model offered modest improvement, showing the importance of exploring more advanced and content-aware features. These findings serve as a foundation for future modeling efforts aimed at improving employee sentiment forecasting.

**Project Assessment Summary**

|  |  |
| --- | --- |
| Evaluation Area | Comments |
| **Documentation** | Each task is clearly described with methods, justifications, and reproducible steps |
| **Clarity & Organization** | Structured by task, includes bullets, visuals, and tables for ease of understanding |
| **Testing & Validation** | Verified sentiment output, monthly groupings, and model evaluation |
| **Reproducibility** | All logic implemented using reproducible and transparent code |

**Overall Outcome:**  
The project meets all defined objectives. Future work could improve modeling accuracy, but the current workflow offers actionable insight for HR and organizational leadership.