**Final Report on Employee Sentiment Analysis**

**1. Approach and Methodology**

My project analyzes an unlabeled dataset of employee messages to assess sentiment and engagement. I initialized a sentiment analysis pipeline using a comprehensive pretrained Cardiff Natural Language Processing (NLP) Model, transformer based and finetuned for sentiment analysis to derive sentiment trends, behavioral patterns, and potential flight risks.

**Data Preparation:** A large dataset of emails was cleaned and augmented with derived features such as word count, punctuation patterns, and date information.

**Sentiment Analysis**: I used a transformer-based language model to classify each email as positive, negative, or neutral.

A screenshot of a computer

AI-generated content may be incorrect.

**Feature Engineering / Exploratory Data Analysis:** I examined the overall data structure with extracted features at the email and aggregated employee-month level, including average word count, message length, subject features, and engagement frequency. I analyzed trends over time using the distribution of sentiment labels across the dataset such as monthly sentiment volume as indicator for flight risk, length of email as indicator for employee engagement, and length of sentences per email vs sentiment as indicator for employee engagement.

**Scoring and Ranking**: Monthly sentiment scores were calculated per employee based on their messages with positive message as +1, negative message as -1, and neutral message as 0 and used for subsequent ranking and risk analysis.

**Risk Identification:** I used a rule-based approach to flag employees at risk of departure by generating ranked lists of employees based on their monthly sentiment scores with 2 distinct lists of top three positive employees and top three negative employees. I then filtered, grouped, and extracted a list of employees that had sent 4 or more negative emails in the span of 30 days.

**Predictive Modeling**: I used a linear regression model to analyze sentiment trends and predict sentiment scores using a variety of features that influenced sentiment scores such as email count, average word count, message length, subject length, number of exclamation marks / question marks, and number of unique subjects. The model’s accuracy and performance were evaluated by metrics such as R² score, Mean Squared Error, and the positive or negative weight of coefficients.

**2. Key Findings from the EDA**

For **Trend 1**, I investigated the monthly sentiment volume as an indicator for flight risk, which is how likely an employee will quit. Negative sentiment emails as frequently as 4 times a month reflect a higher risk of the employee quitting. After plotting the monthly sentiment volume over time and the sentiment ratio over time, my findings revealed that neutral sentiment dominated email traffic specifically in earlier months, which could possibly indicate healthy communication. Positive sentiment was more common in longer and more descriptive emails while negative sentiment tended to appear in shorter emails.

In the monthly sentiment over time graph, neutral sentiment emails are the highest with negative emails being consistently the lowest. This could possibly indicate healthy communication. If positive sentiment is healthy and increasing alongside neutral, this could reflect professional yet constructive engagement. However, low negativity alone doesn’t guarantee employee well-being as it may reflect disengagement rather than harmony.

A graph of a number of bars

AI-generated content may be incorrect.

In the sentiment ratio over time graph, it can be observed that for neutral emails, it is most frequent in Jan 2010 and lowest in December 2010. For negative emails, it is lowest in November 2010, and highest in December 2010. For positive emails, it is also highest in December 2010, and lowest in February 2010. December 2010 shows heightened emotional polarization as positive and negative emails peak in this month. This suggests stronger emotional employee engagement at the end of the year possibly due to year-end stress or celebration, performance reviews, project deadlines, and/or organizational changes (e.g., bonuses, restructuring, layoffs). The large drop in neutral emails from January to December may reflect increasing emotional transparency as the year progresses, and a shift from routine communication to more reactive and expressive messaging.

A graph of different colored lines

AI-generated content may be incorrect.

For **Trend 2**, I investigated the length of employee email as a strong indicator for employee engagement. Longer emails suggest more effort, though, or emotional investment which are potential signs of active engagement. Shorter emails can indicate disengagement, minimal communication, or burnout from the employees. Sudden drops in length suggest a decline in motivation while sudden spikes can reflect emotional stress especially paired with negative sentiment.

A graph of blue rectangular shapes

AI-generated content may be incorrect.

The graph of Average Email Length vs Sentiment shows positive emails with a higher word count while neutral and negative emails are around the same length. This demonstrates that positive emails tend to be more expressive and thoughtful with employees putting more effort into articulating appreciation, praise, or encouragement. These messages may include context, gratitude, and elaboration, which aligns with higher engagement. Neutral and negative emails are more concise, which indicates blunt or short communication, often associated with frustration or urgency. This can also be associated with employees not feeling safe or motivated to elaborate and could suggest growing dissatisfaction or risk of disengagement.

A graph of blue dots

AI-generated content may be incorrect.

For **Trend 3**, I investigated the length of sentences per email vs sentiment as an indicator for employee engagement. Positive emails will often have the longest sentences which can be interpreted as thoughtful, expressive, and articulate messages that indicate higher engagement and positive emotional energy. Employees may take time to explain, appreciate, or elaborate their ideas and opinions. Neutral emails may have medium sentence length and can be interpreted as objective and factual tone suggesting positive communication. Negative emails would have the shortest sentences that reflect abrupt communication.

**A graph of blue rectangular bars

AI-generated content may be incorrect.**

**A diagram of a graph

AI-generated content may be incorrect.**

**3. Employee Scoring and Ranking Processes**

For employee scores, I first parsed the date column to datetime and extracted the year-month for monthly aggregation. Then I mapped the sentiment labels to numeric scores with positive messages as +1, negative messages as -1, and neutral messages as 0. I then grouped by employee (sender) and month to aggregate these scores monthly for each employee. Finally, I summed these sentiment scores and printed out a list of monthly sentiment scores for each employee.

A screenshot of a computer code

AI-generated content may be incorrect.

For employee ranking, after converting date to date-time, extracting year-month, and mapping sentiment to numeric score, I aggregated monthly sentiment score per employee by sorting alphabetically and then by month. Based on the monthly scores I generated in the previous task, I generated rankings for the top 3 employees with the highest positive scores and the top 3 employees with the lowest negative scores and combined them into 2 distinct lists to print out.

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**4. Flight Risk Identification Criteria and Outcomes**

After loading the dataset and converting date to datetime format, I filtered only negative sentiment emails and sorted them by sender and date. Then I grouped by employee and used a sliding window check over the sorted list of email dates. For every set of 4 consecutive emails, it checks if those 4 emails were sent within 30 days and if yes, the employee is considered at risk. If at risk, it records the employee’s name and total count of negative emails in flight risk data. Finally, it extracts a list of these employees that have sent 4 or more negative mails in the span of 30 days.

The list of employees is:

**A group of black text

AI-generated content may be incorrect.**

**5. Overview and Evaluation of the Predictive Model**

For my linear regression model, I started with a large set of engineered features from the dataset that may correlate with the sentiment, such as {email\_count, avg\_word\_count, avg\_message\_length, std\_word\_count, std\_message\_length, subject\_excl\_count, avg\_subject\_length, avg\_question\_marks, num\_unique\_subjects, avg word count, message length, subject length, avg question marks}. These features are grouped by sender and year-month to align with monthly sentiment.

The linear regression model uses LinearRegression from scikit-learn which splits the data into the training and test set. It trains the model on the training set and validates it using metrics like mean squared error (MSE) and R² score on the test set to assess prediction accuracy and the proportion of variance explained by the model.

I tested the model with different combination of features. From my non-exhaustive experiment, I found that the best performing model uses the following four features: {***avg word count, message length, subject length, subject excl count, question mark count***}.

**A graph with blue dots and red line

AI-generated content may be incorrect.**

**R²: 0.149** - This shows that the model explains ~15% of the variance in monthly sentiment scores, which shows that it is a meaningful predictor of monthly sentiment score especially in noisy data like emails. This reflects that while these four features do influence sentiment, a lot of the variation is still unexplained due to unmeasured context, tone, or other factors that are external to email text.

**MSE: 3.411** - The average squared error between predicted and true sentiment scores is around 3.411.

**Intercept: -0.2453747082027118** - The coefficient is negative which indicates that the model predicts a slightly negative baseline.

Given the low **R²** score and high MSE, it is evident that linear regression model using the features that I have created may not be the most suitable to predict sentiment scores. I believe it is mainly due to the categorical nature of the labels and the noisy labeling by the Cardiff NLP model. A classification model such as a random forest (or decision tree) may be more suitable to utilize the large number of the features that can be derived from emails to classify sentiment. Logistic regression is another alternative if we formulate the problem as a per-employee-per-month classification problem.

**Evaluation of other Features:**

**email\_count:** A sharp drop in email count might reflect disengagement which is a possible negative signal. High volume may indicate high workload or engagement.

**num\_unique\_subjects**: Variety in subjects might suggest an employee is involved in diverse work (potentially correlating with engagement and positive sentiment. Low variety might mean repetitive tasks, which could relate to burnout or negative tone.

**avg\_hour\_of\_day:** Sending emails outside normal hours may indicate overwork or stress — possibly correlating with negative sentiment. Mid-day communications may be more neutral or positive.

**avg\_word\_count: +0.0030** - Very small positive effect, longer average email bodies are weakly associated with more positive sentiment.

**subject\_excl\_count: +0.3643** - Strongest positive coefficient: more exclamation marks in subject lines are linked to higher sentiment scores which likely captures positive enthusiasm more than frustration. It also suggests it is the most emotionally charged and predictive of positivity where employees who express enthusiasm are often more engaged.

**avg\_question\_marks: –0.0412** - Slightly negative: more question marks in emails correlate with more negative sentiment, possibly reflecting confusion, doubt, or dissatisfaction. The negative coefficient also aligns with the hypothesis that confusion can be a negative signal.

**num\_unique\_subjects: +0.1912** - Moderate positive: diverse subject lines suggest engagement with multiple tasks or areas is tied to more positive sentiment.