LLM FINAL ASSESSMENT

Submitted by – Kushagra Jain

Data Analytics and LLM Internship

PROJECT OVERVIEW

The goal of the project is to evaluate employee sentiment and engagement using data analysis techniques such as sentiment analysis, exploratory data analysis, sentiment labelling, employee ranking and predictive modeling. I need to analyze employee messages to understand how people feel and how engaged they are at work. I'll be working with raw data that doesn't have labels, so I will be figuring out the patterns myself using text analysis and statistics. The project has several different parts where I will analyze the data and build models to get insights.

PROJECT OBJECTIVE

The main goal is to evaluate employee sentiment and engagement by performing the following:

* **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.

SENTIMENT LABELING

For sentiment labelling, I first prepared the text to be analyzed:

* Removing special characters, HTML tags, and extra whitespace.
* Normalizing text (lowercase everything).
* Removing stopwords (common but unimportant words like “the”, “and”, etc.).
* Lemmatization (reducing words to root forms like “running” → “run”).
* Combined subject and body text for full context of the email into a new column called full\_text

For the tool to be used for sentiment analysis, I used Vader as it is a reliable tool commonly used for analyzing sentiment from emails. It is rule-based, fast and reliable, and is great for social texts, emails et. As an alternative, I also tried transformer model DeBERTa by Microsoft, however it was not providing accurate results and was interpreting every mail as neutral.

For the sentiment analysis, I used the VADER tool. The labelling criteria in VADER is as follows:

* Compound Score Range >= 0.2: 'Positive'
* Compound Score Range <= -0.2: 'Negative'
* Compound Score Range between -0.2 and 0.2: 'Neutral'

These cutoff values are chosen for the sentiment analysis as; when doing sentiment analysis on corporate emails from employees, you're working with formal, nuanced, and often emotionally subtle language. Stronger thresholds (±0.2) reduce misclassification of slightly positive/negative emails as overly emotional.

The compound score is a normalized score given to text, calculated using a lexicon of words with known sentiment values and some heuristic rules

Here is a sample of the sentiment labelling:

A screenshot of a computer

AI-generated content may be incorrect.

EXPLORATORY DATA ANALYSIS

I performed exploratory data analysis on the data to understand the trends and patterns in the data regarding the sentiment scores. I examined the overall structure (eg. the number of records, data types, missing values, null values etc.).

A screenshot of a computer

AI-generated content may be incorrect.

I also investigated the distribution of the sentiment labels across the dataset to understand the sentiment trends among the employees.

A bar graph with different colored squares

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

As it can be clearly observed from the data, most of the emails are of a positive sentiment. After positive, the next most prevalent sentiment is neutral. This is good news for the company as this implies that most of the employees display a positive attitude in their work culture, and are generally happy (or atleast not upset) about their work and work related situations.

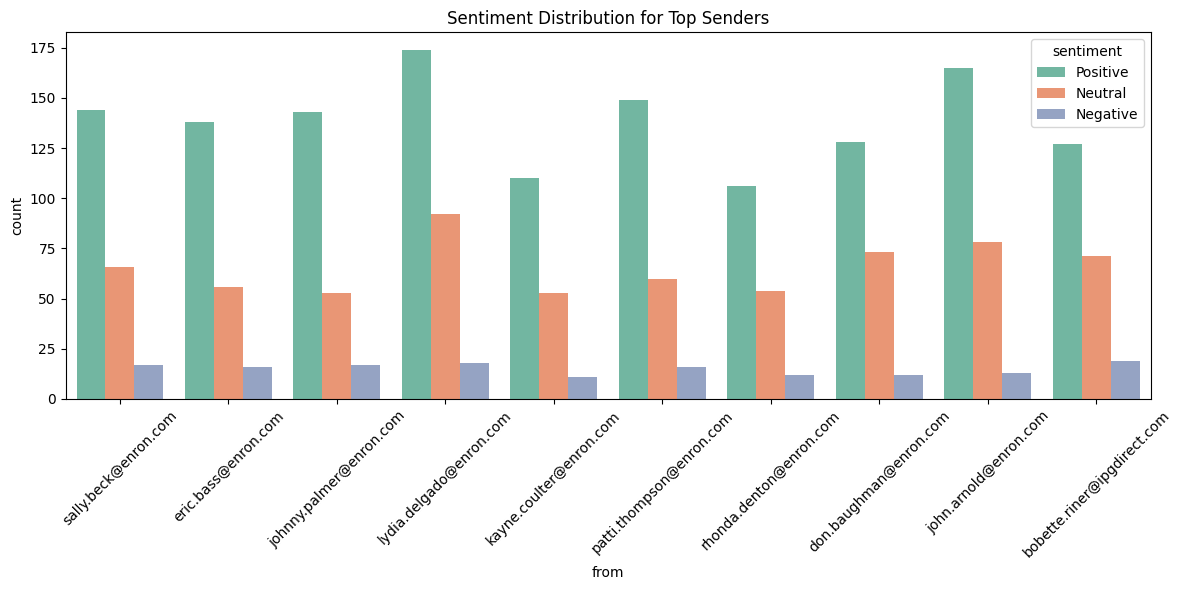
I analyzed the trends of sentiment scores over time to understand if there are any seasonal effects in the sentiments of the employees.

A graph of different colored lines

AI-generated content may be incorrect.

From the chart, I’ve learned that positive emails consistently dominate over time, while negative sentiments remain low and stable. Neutral emails show more fluctuation but trend slightly downward. This suggests overall communication tone is healthy. Based on this, I can predict a continued positive trend unless a major event disrupts it. This insight is important because it helps assess employee or customer morale over time, allowing me to act early if sentiment shifts negatively.

I also investigated the major sentiment for each individual employee, and the respective distributions of sentiments over time.



From this chart, I’ve learned that most top email senders tend to use a positive tone, with negative messages being very few across the board. Lydia Delgado and John Arnold stand out with especially high positive counts. This could mean they’re influential in shaping a constructive communication culture. I can use this to identify role models or potential internal influencers. It’s important because understanding who spreads positivity helps foster better workplace morale and communication strategies.

A summary of the sentiments among the employees and their nature can be found below:

A screen shot of a computer

AI-generated content may be incorrect.

TAKEAWAYS

Additional Patterns and Insights into Employee Engagement:

* Positive Sentiment Dominates: Suggests a generally optimistic culture or model bias (e.g., polite closings).
* Stable Positive Trend: Positive sentiment remains steady across months.
* Fluctuating Neutral/Negative Sentiments: Possible reaction to internal events or pressures.
* Top Senders = Mostly Positive: Likely leaders or HR roles maintaining positive tone.
* Email Length Patterns:

Positive emails → longest subject and body.

Neutral emails → shortest, likely factual or procedural.

* Anomalies:
* Spikes in negative/neutral sentiment may indicate specific stress events.
* Certain senders show higher negative sentiment—worth deeper review.

EMPLOYEE SCORE CALCULATION

I computed a monthly score for each employee based on their messages. The scores were aggregated on a monthly basis for each employee. These scores were reset at the beginning of every new month. The scoring metric is as follows:

* Positive Message: +1
* Negative Message: -1
* Neutral Message: 0 (No Effect)

The result of the task is the creation of the column monthly\_sentiment\_score.

EMPLOYEE RANKING

I generated a list of employees based on their monthly sentiment scores. I then ranked them to understand who the employees were who were most likely to send a positive or a negative message.

I created 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.

I sorted them first in descending order and then in alphabetical order. I ensured that the ranking is clearly derived from the sentiment score calculation as previously executed.

A screenshot of a computer

AI-generated content may be incorrect.

FLIGHT RISK IDENTIFICATION

This stage involved identifying the employees that are at a risk of leaving the company based on their monthly sentiment scores. A flight risk was established as any employee who had sent 4 or more negative emails in a span of 30 days (irrespective of the score). The 30 day period is a rolling count of days, irrespective of months.

A grey screen with white text

AI-generated content may be incorrect.

PREDICTIVE MODELING

I developed a linear regression model to analyze sentiment trends and predict sentiment scores using a variety of independent variables that may influence sentiment scores. The idea was to predict the sentiment scores of an employee based on their engagement and sentiment on mails. This would help us to identify potential flight risks.

FEATURE ENGINEERING

I first created the following features for the purpose of prediction in the dataset:

* Subject Length: The total length of the subject of the mail.
* Body Length: The total length of the body text of the mail.
* Word Count: The total word count of the combined text (subject and body text).
* Average Monthly Messages: The average number of messages sent by a particular user in a month.

I also checked for the distributions of the numerical variables in the model to check whether they are normally distributed or not.

A group of graphs showing different sizes of data

AI-generated content may be incorrect.

It can be observed that the variables subject\_len, body\_len and word\_count are skewed to the right, and therefore we need to transform these before running the model. I decided to perform log transformation on the said numerical variables to transform them into normally distributed variables.

MODEL DEVELOPMENT

I then proceeded to split the data into training and testing dataset (80%-20%). I then trained the logistic regression model on the training dataset and cross validated across the testing dataset. The model output and evaluation metrics are as follows:

A screenshot of a computer screen

AI-generated content may be incorrect.

Takeaways:

* The logistic regression model predicts positive sentiments well (F1-score: 0.80, recall: 0.90).
* It struggles with negative sentiments—those are barely detected (F1-score: 0.00).
* Neutral sentiments are moderately predicted (F1-score: 0.53).
* Overall accuracy is 70%, driven mostly by strong positive sentiment detection.
* The model may be biased due to class imbalance, with positives dominating.
* I need to improve recall for minority classes, especially negatives.

A screenshot of a computer

AI-generated content may be incorrect.

INTERPRETATION

* I built an ordinal logistic regression model to predict sentiment scores using features from emails like message count, log of subject length, log of body length, and log of word count.
* Among these, only log\_word\_count showed a statistically significant effect (p < 0.001), with a positive coefficient of 0.86.
* This means that as the word count of an email increases, the likelihood of the sentiment being more positive also increases.
* The odds ratio of 2.37 suggests that a unit increase in log word count more than doubles the odds of moving to a more positive sentiment category.
* Other variables didn’t show significant influence (p > 0.4), indicating they may not be strong predictors of sentiment in this context.

CONCLUSION

* The sentiment analysis model performs well overall, with 70% accuracy, largely driven by its strong ability to detect positive sentiments (F1-score: 0.80, recall: 0.90).
* However, it significantly underperforms in identifying negative sentiments (F1-score: 0.00), suggesting a potential issue with class imbalance.
* Neutral sentiments are predicted with moderate success.
* On the modeling side, only log\_word\_count was found to be a meaningful predictor—emails with higher word count were much more likely to carry positive sentiment.
* Other features such as message count, subject length, and body length didn’t significantly affect sentiment prediction.
* Overall, while the model effectively captures positivity through wordiness, it needs improvement in recognizing less frequent negative and neutral sentiments for balanced and fair performance.

The model can be improved further by adding more features. As we can see the model has a weak fit on our data, this can be improved by adding more features to the data to make the model more reliable. We can also develop advanced feature engineering techniques such as squaring and interaction terms to further develop the model fit on our data.