1. A description of my approach and methodology

I implement the SpaCy to classify the emotional tone of email bodies from the provided CSV file. The pipeline uses the spacytextblob component, which enables sentiment analysis through the TextBlob library. This approach combines text preprocessing with machine learning-based sentiment classification.

**Methodology Steps**

**1.Data loading and spaCy installation** :

* Before loading data, install spaCy using the command: conda install -c conda-forge spacy. spaCy is particularly useful in large database environments
* Used pandas to read the testin.csv file containing email data (subject, body, date, sender)

**2.Text Preprocessing**:

* Converted all text to lowercase for consistency
* Removed punctuation using regular expressions
* Normalized whitespace by replacing multiple spaces with single spaces

**3.Sentiment Analysis Pipeline**:

* Utilized spaCy's NLP pipeline
* Integrated TextBlob's sentiment analysis capabilities via spacytextblob extension
* Configured a three-class sentiment classifier (Positive/Neutral/Negative)

-Positive sentiment: polarity score > 0.4

-Negative sentiment: polarity score < 0

-Neutral sentiment: scores between 0 and 0.4

**4.Implementation**:

* Applied the preprocessing and classification to each email body
* Added sentiment labels as a new column in the DataFrame

2. Key Findings From the EDA

BASIC DATA STRUCTURE

Total records: 2191

Missing values per column:

Emails: 145

body 0

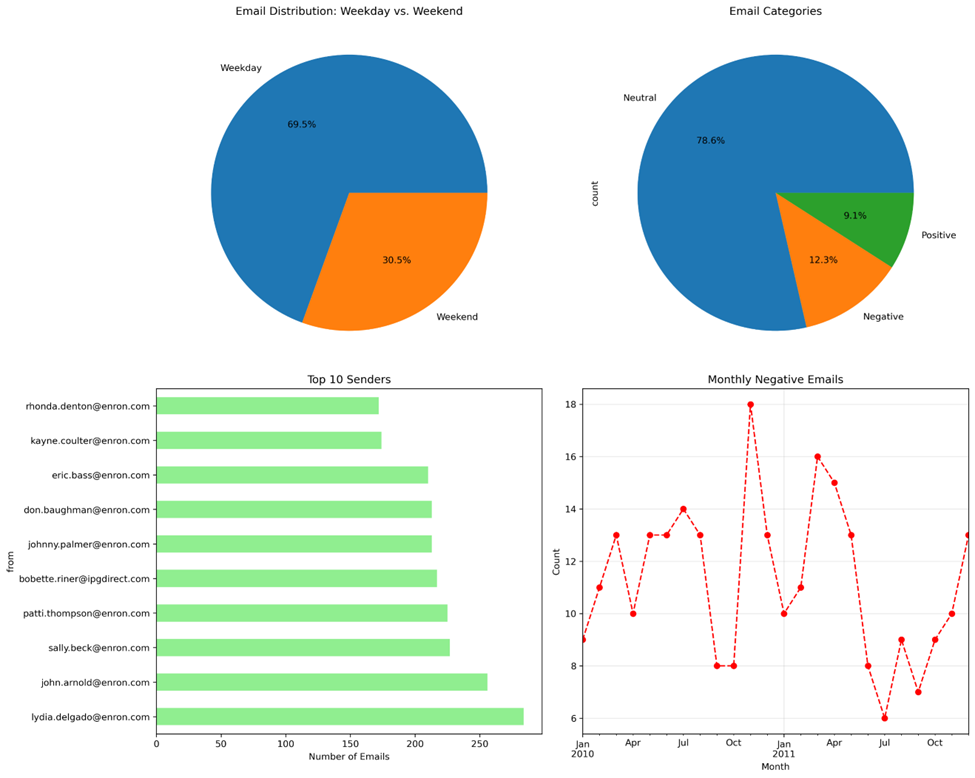
date 0

from 0

sentiment\_label 0

Here is some basic information about the employee sentiment dataset. The database contains 2,191 records. The number of missing values varies by column; for example, there are 145 emails without a subject in the subject column. Importantly, all other columns have no missing values, which simplifies the process for the upcoming sentiment analysis.

Here are a few key findings from the EDA:



In the chart of Email Distribution: Weekday vs. Weekend, some emails are sent on weekends, indicating that many employees still work during that time.

In the Email category pie chart, many employees maintain a formal tone in their emails, as evidenced by the 78% neutral category. Additionally, negative emails outnumber positive ones, but this is still acceptable.

We can observe a peak in the monthly negative email trend in October 2010, followed by another upward trend in October 2011. I believe that the number of negative tone emails will increase as the company its busy business cycle, which begins every October. This suggests that there is a strong relationship between workload and the negative tone in email.

A graph of a bar chart

AI-generated content may be incorrect.

I analyze the tone of emails over time. In the negative and neutral categories, there is not much change between 2010 and 2011. However, there is a slight increase in positive emails during this period. This suggests that some changes may have occurred.

3. Explanation of the employee scoring and ranking processes

Step 1: Preparing Employee Scores and time for Ranking

* Extracts month from the date column (format: YYYY-MM). Ensure that the data is consistent
* Converts sentiment labels (Positive, Neutral, Negative) into numerical scores (1, 0, -1). Following the requirement of Task 3
* Stores scores as integers for aggregation.

Step 2: Calculating Monthly Scores

* Groups data by month and sender (from).
* Sums sentiment scores for each employee per month.
* Resets index to convert the grouped result back into a DataFrame.

Step 3: Selecting Top 3 Employees per Month

* Sorts by month (ascending) and score (descending).
* Groups by month and takes the top 3 rows per group (head(3)).
* Resets index for clean output.

Step4: Formatting the Output

* Iterates through each row in top\_3.
* Detects new months and adds a header to classify each month
* Assigns ranks (1, 2, 3) for each month.

Step5: Bad 3 Employees per month

* Sorts by month (ascending) and score (ascending).
* Copy the steps like Top 3 process

4. Flight risk identification criteria and outcomes

First, I need to filter all the data to ensure that it only includes emails with a negative label. The requirement for Task 5 is to identify employees who have received four or more negative emails within any 30-day period. To achieve this, I will use a rolling window approach to analyze each email, starting from its date.

Within each identified 30-day window, I will count the total number of negative emails that fall within that range. Specifically, this involves counting emails where the date is greater than or equal to the start date and less than or equal to the start date plus 30 days. If any of these 30-day periods contain four or more negative emails, I will flag this occurrence appropriately. In that case, the result will indicate a return value of True, suggesting that there may be a potential issue that requires further review.

Outcome:

Negative employees in Row

0 bobette.riner@ipgdirect.com

1 don.baughman@enron.com

2 eric.bass@enron.com

3 john.arnold@enron.com

4 johnny.palmer@enron.com

5 kayne.coulter@enron.com

6 lydia.delgado@enron.com

7 patti.thompson@enron.com

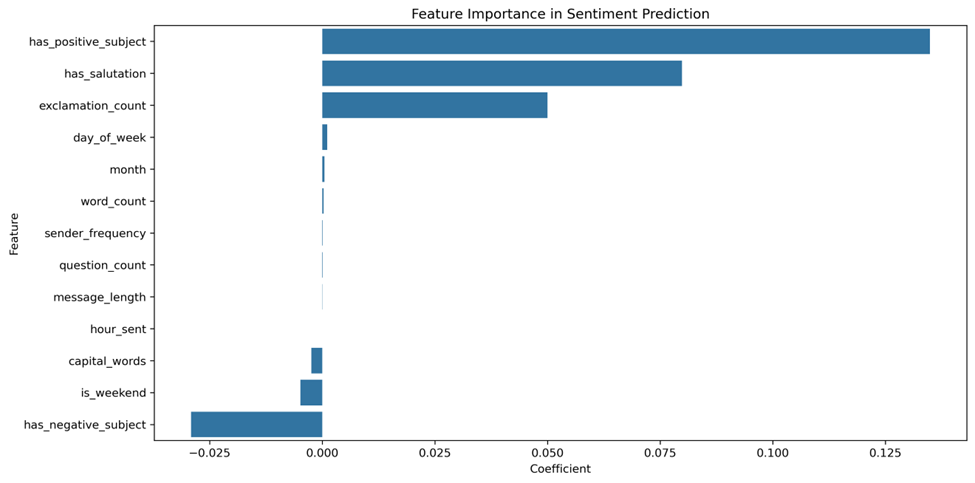
8 rhonda.denton@enron.com

9 [sally.beck@enron.com](mailto:sally.beck@enron.com)

5. Overview and evaluation of the predictive model.

The model is not useful for identifying general sentiment trends but may need to use other types of predictive model to precise enough for analysis of individual emails.

1. **R-squared (0.101)**:
   * Indicates that approximately 10% of the variance in sentiment scores is explained by our model
   * This is a low level of explanatory power for this model
   * Probably other factors influencing sentiment that I haven't captured
2. **Mean Absolute Error (0.154)**:
   * On average, our predictions are off by ±0.154 points on the sentiment scale
   * Since sentiment scores range from -1 (most negative) to 1 (most positive), this represents about 6% of the total scale range
   * For context:
     + 0.154 error on a -1 to 1 scale is reasonably good
     + Predictions for neutral emails (near 0) will be quite accurate
     + Extreme sentiments (near -1 or 1) may have larger errors
3. **Root Mean Squared Error (0.213)**:
   * Higher than MAE, indicating some larger errors are present
   * The square term in RMSE penalizes larger errors more severely
   * While most predictions are close, there are some emails where sentiment is harder to predict accurately



Clearly, the major factors affecting sentiment prediction are positive and negative words, along with certain punctuation that can express emotion. However, it is important to note that work time also has a enough effect on model predictions.