**Employee Sentiment Analysis**

Youyi Lei

6.25.2025

1. **Introduction**

This project uses an unlabeled dataset of employee messages to analyze and assess the sentiment and engagement of each employee. Generally, this project consists of six distinct tasks, including sentiment labeling, exploratory data analysis (EDA), employee score calculation and ranking, flight risk identification, and predictive modeling. Each task in this project represents an aspect of data analysis and model development.

1. **Approach and Methodology.**

In this project, Python is widely used as the coding language, and several libraries or packages from Python, such as Pandas, NumPy, TextBlob, VADER, and scikit-learn, are utilized to accomplish each distinct task.

The project begins with the usage of the TextBlob library to analyze the sentiment of the text input from the dataset and decide whether it is positive, negative, or neutral based on its sentiment polarity. Due to the low accuracy of setting default thresholders after justification, several groups of thresholders, including setting -0.15 to 0.15 as neutral, -0.2 to 0.2 as neutral, and, are checked. However, these groups of cutoff values turn out to contain errors that incorrectly label neutral messages as positive and fail to identify obvious negative tones. As a result, I apply another rule-based tool for sentiment analysis named VADER (Valence Aware Dictionary and sEntiment Reasoner) to improve the accuracy. After several manual checks, I set the cutoff value from -0.3 to 0.3 as neutral and added the customized analyzer to adjust the sentiment lexicon of certain words in the workplace, and several neutral phrases are also considered. Since the range for neutral messages is wider than the default situation, this approach avoids incorrectly detecting positives in business language. For further manual review, sample borderline cases near decision thresholds are printed out, which helps to analyze errors and tune the threshold.

However, both TextBlob and VADER rely on lexicon-based approaches, which lack contextual understanding in real life. Besides, even though several domain-specific words in business are added for customization, these approaches may still reduce the accuracy when facing formal English in business emails since they are widely applied for language from social media. In order to handle contextual nuances such as metaphor and irony, and work on workplace communication, pretrained sentiment analysis models such as cardiffnlp/twitter-roberta-base-sentiment and nlptown/bert-base-multilingual-uncased-sentiment could be further applied. However, I did not verify these models because it seems that my computer cannot currently support this kind of model.

After sentiment labeling, each message from the dataset is applied to generate the monthly sentiment score for every employee, and then the monthly ranking. At last, there is a multiple linear regression model used to explore the relationship between the monthly sentiment score and various features, including message length, word count, and the frequency of specific keywords in the messages. Machine learning algorithms are also included to split the dataset and then test the remaining data for future prediction.

1. **Key findings from the EDA.**

From the exploratory data analysis process, it is first clearly shown that the number of records of this dataset used here is 2191 and contains four columns, which are all object data type and are listed in the following table.

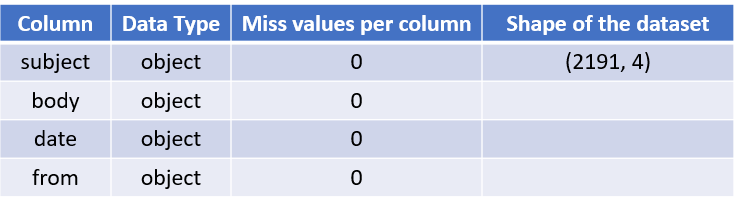


Table 1 Overall Data Structure

After that, the dataset is further analyzed to plot the figure below, showing the distribution of the positive, negative, and neutral messages. Among all the labelled 2191 records, there are 90 negative sentiment labels, 750 neutral sentiment labels, and 1351 positive sentiment labels. As a consequence of this, negative messages consist of the smallest proportion among all of the records, which could be a satisfying and acceptable sign for the company and represent the overall positive engagement of employees.

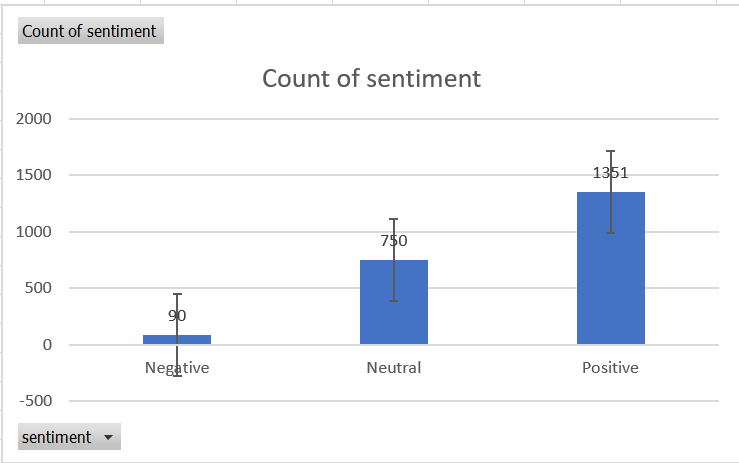


Figure 1 Count of all the Sentiment Labels from 2010 to 2011

In addition, this figure illustrates the monthly sentiment trend over time, with the positive sentiment represented by the green line, the negative sentiment represented by the blue line, and the neutral sentiment represented by the orange line. From the trend line, the number of positive messages seems to be increasing, and the number of negative messages seems to be decreasing on an overall scale. Besides, compared with 2010, there are more positive emails and fewer negative emails in 2011, which indicates improvements in the attitudes of employees in the workplace. After checking the dataset along with this plot in detail, April and September contain relatively the most positive messages and the least negative messages. Possible reasons may include changes related to human resources, such as new recruitments and promotions, completed transactions, corporations, or agreements. While, there are no obvious fluctuations in the condition of neutral messages. As a result, this monthly sentiment trendline plot explains the overall pattern of the number of each kind of message between 2010 and 2011, indicating employees’ attitude and engagement, and also predicting the future condition.

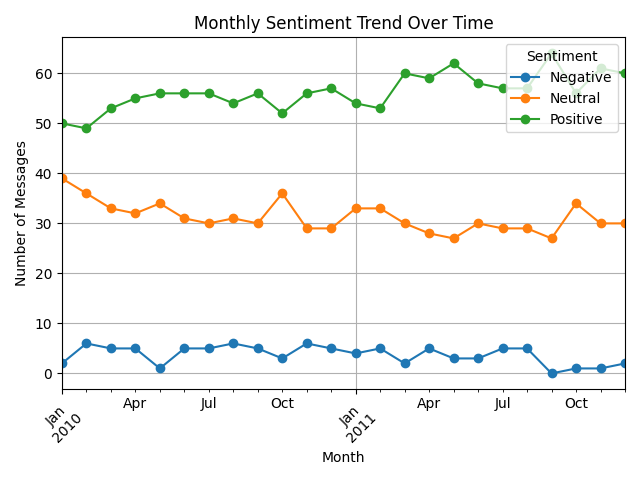


Figure Monthly Sentiment Trend Over Time (2010.1 to 2011.12)

1. **Explanation of the employee scoring and ranking processes.**

In this task, the sentiment score for each employee on a monthly basis is calculated first by assigning a score to each message they sent, according to the following standard.

* **Positive Message: +1**
* **Negative Message: –1**
* **Neutral Message: 0 (no effect)**

All the scores are saved in a single column and reset at the beginning of each new month. In order to group messages by month, the Groupby() method in Pandas in Python is applied. So that all rows such as ‘messages’ from the same employee in the same month are collected together. Detailed methods include using the ‘from’ column in the dataset as the email sources, creating another column named ‘month’, which is extracted from the ‘date’ column. After grouping, the score column is selected, and the scores for each group are summed.

After gaining the monthly sentiment scores of all the employees, a ranking is available. For each month in 2010 and 2011, the monthly sentiment score of each employee is compared with others, and then the three employees with the highest scores and the lowest scores are picked and saved as a singer dataset. The three employees with the highest positive scores (most positive) and the lowest scores (most negative) in a given month for 2010 and 2011 are listed in the following two tables. From these two tables, specific employees who sent the most positive and negative messages each month can be found, which may provide the human resources department with insights into the solicitude for employees and related regulations within company.



Table Top 3 Positive and Negative employees each month in 2010

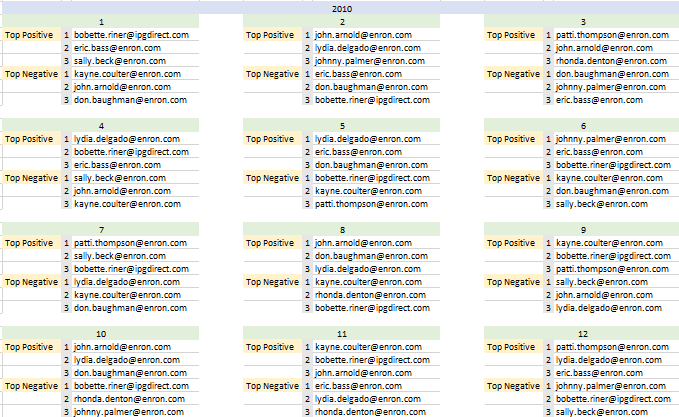


Table Top 3 Positive and Negative Employees each month in 2011

1. **Flight risk identification criteria and outcomes.**

Since the flight risk is an employee who has sent 4 or more negative emails in the span of 30 days, whether an employee has sent over 4 negative emails can be decided from the previously labeled dataset. Thus, the number of negative sentiments of each employee is calculated, and those that have a number of 4 or more are extracted to be at flight risk, listed in the following table.

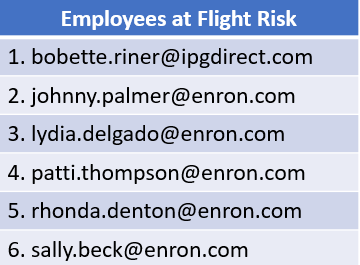


Table Employees who are at Flight Risk

1. **Overview and evaluation of the predictive model.**

Finally, in this project, there is a multiple linear regression model used to analyze sentiment trends and predict sentiment scores. A variety of independent variables, including message count, message length, word count, and the frequency of the three keywords “thank you”, “sorry”, and “best regards” occurring in the messages, are chosen, which may influence the monthly sentiment score. In this multiple linear regression model, 80% of the data are split into the training set and the remaining 20% is the testing set to evaluate model performance. After picking the independent variables and fitting them into the model, the table below shows the coefficient of each feature and the metrics chosen to evaluate the model.

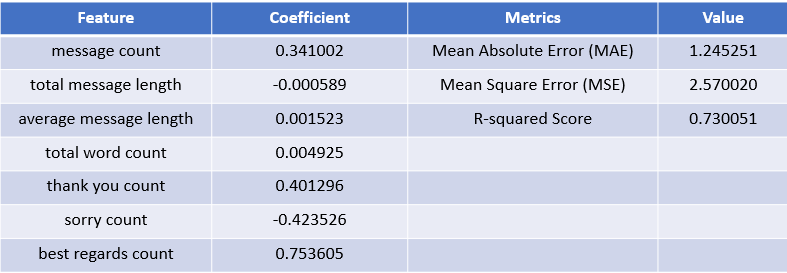


Table Coefficient of Each Feature and Various Metrics Showing Effectiveness

From the coefficient column of this table, except for total message length and sorry count, other features all positively affect the monthly sentiment score of an employee. Among them, message count, thank you count, and best regards count have more significant influences on sentiment score compared to other features. However, considering the various requirements of communication efficacy, clarity, and informativeness in the workplace with respect to different possible situations encountered, there may not be very direct and obvious positive or negative relationships between the total message length and the total word count of an email and the sentiment score.

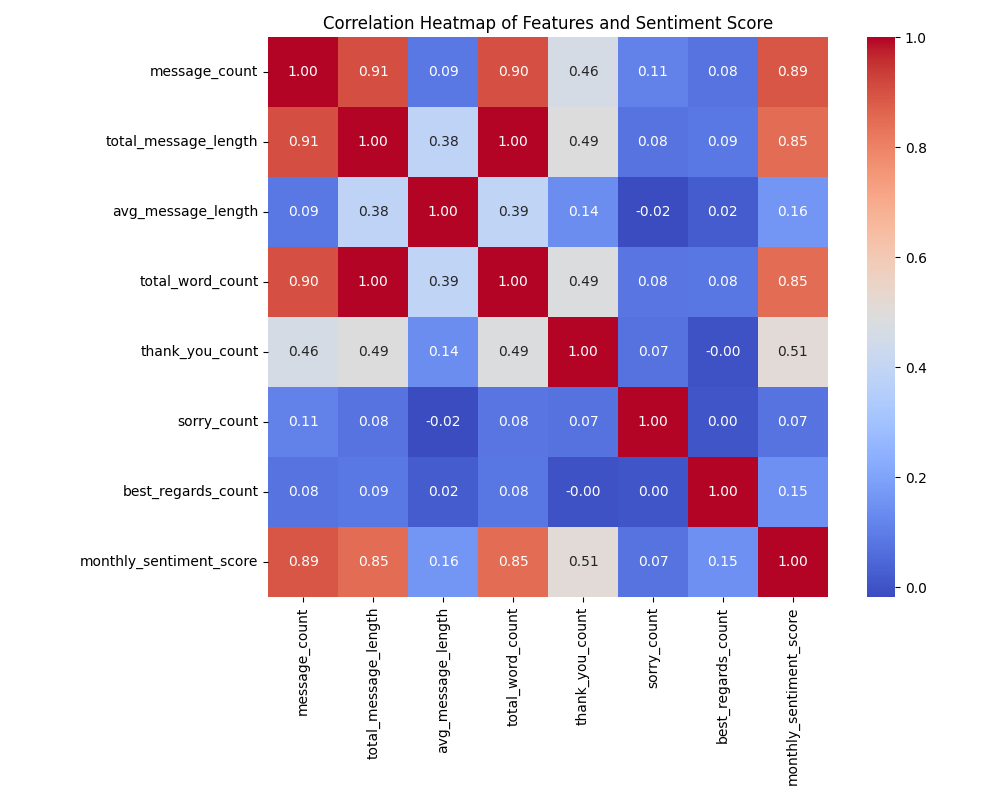


Figure Correlation Heatmap of Feature and Sentiment Score

Besides, the correlation heatmap above illustrates the more detailed correlation between multiple variables. It indicates that for business emails in the workplace, which contain a larger proportion of formal language and polite language, there are strong correlations between message length, message count, and the frequency of the usage of keywords like “thank you”. This heatmap also helps detect multicollinearity of those multiple variables and may further decide on more appropriate features to avoid multicollinearity.

In terms of the metrics to evaluate the model, the mean absolute error (MAE), mean squared error (MSE), and the R-squared score are printed out and listed in the table. From the R-squared score, this model explains about 73% of the variation, indicating the goodness of the model's fit with the data. With the average absolute error between predicted and the actual sentiment of about 1.245, and the average squared error of about 2.57, the model performs fairly well since they are both lower.

To visualize the relationship between the actual and predicted sentiment scores and justify the model performance, following plots are attached.

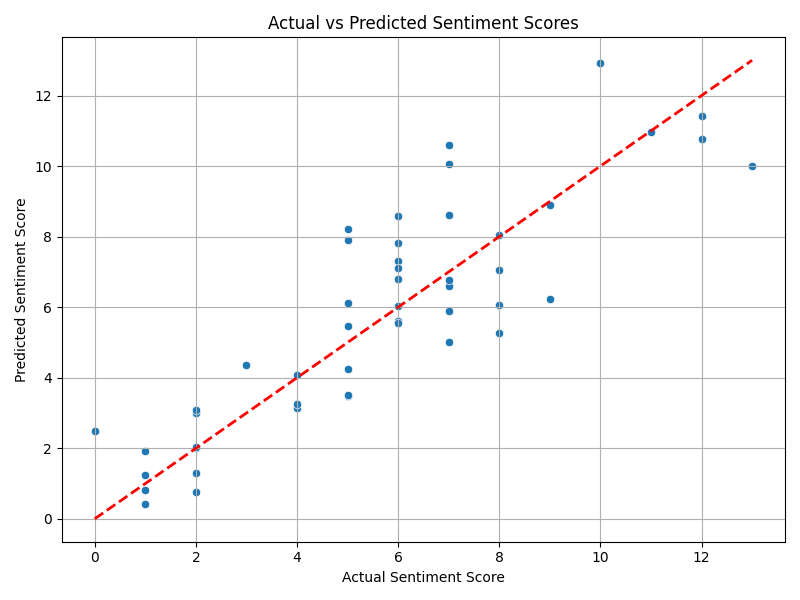


Figure Actual vs. Predicted Sentiment Scores

The above plot first assesses how close the predictions are to the actual sentiment values, with the red dashed line indicating the ideal situation. Since most of the data points, which are represented by blue dots, are close to the red dashed line, the prediction of this model is accurate and close to the real-life situation. However, from those points set far away from the red line, there exist worse predictions and even bias in the model, and the reason might be the large number of contextual nuances in human communication that fail to fit in the model.

In addition, the residual distribution plot below is applied to check the normal distribution of errors. According to the plot, the peak of the bell-shaped line is very close to 0, which means that the linear regression model is proper enough for sentiment score analysis and prediction. AS a result, a linear regression model can be applied to analyze features that affect sentiment scores of business emails in the workplace and predict the condition in the future.

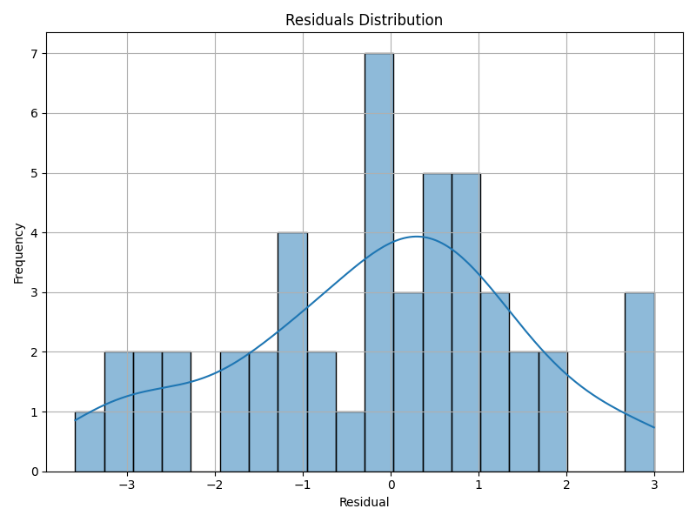


Figure Residual Distribution

Below is also the Residual vs. Predicted sentiment scores plot that indicates the residual pattern.

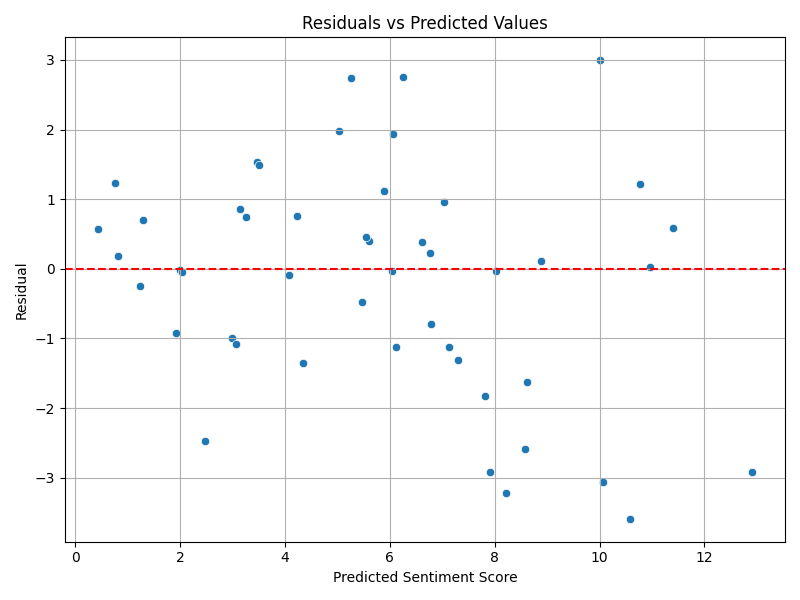


Figure Residuals vs. Predicted Sentiment Scores

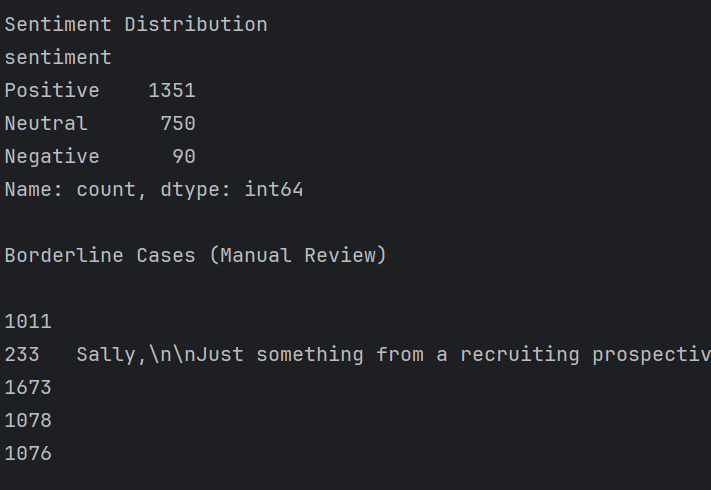
* **X-axis**: Predicted sentiment scores
* **Y-axis**: Residuals (Actual − Predicted)

Since most of the data points in this plot, when the predicted sentiment score is lower, are distributed around 0, linear regression is again proper for this dataset. Moreover, there is no clear pattern or funnel shape, which suggests that errors are spread fairly evenly across predicted values and reassures the appropriateness of linear regression. However, as the predicted sentiment score becomes bigger, data points are more distributed, which may indicate that there are outliers or non-linear relationships in the dataset. Even though it is the case, by applying a multiple linear regression model, specific insights can still be concluded, while with other models like the logistic regression model, the performance could be improved.

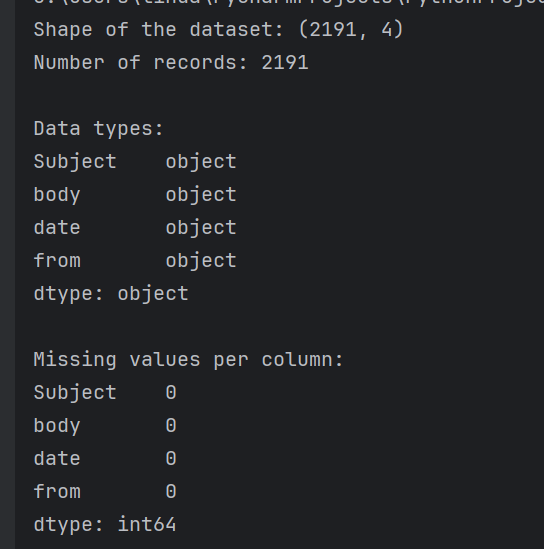
Appendix:

Running Result Screenshot:

Task 1:



Task 2: EDA



Task 5:



Task 6:

