**Employee Sentiment Analysis**

Lorraine 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, 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. After that, each message from the dataset is labeled and can be further applied to generate the monthly sentiment score for every employee, and then gain 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.

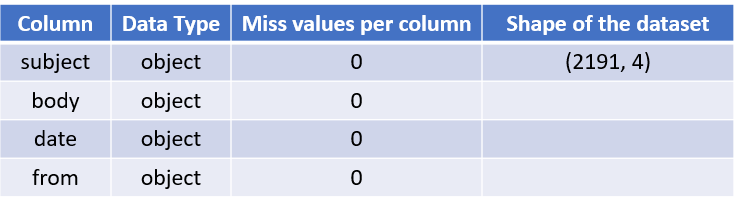


Table 1 Overall Data Structure

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 196 negative sentiment labels, 1019 neutral sentiment labels, and 976 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.

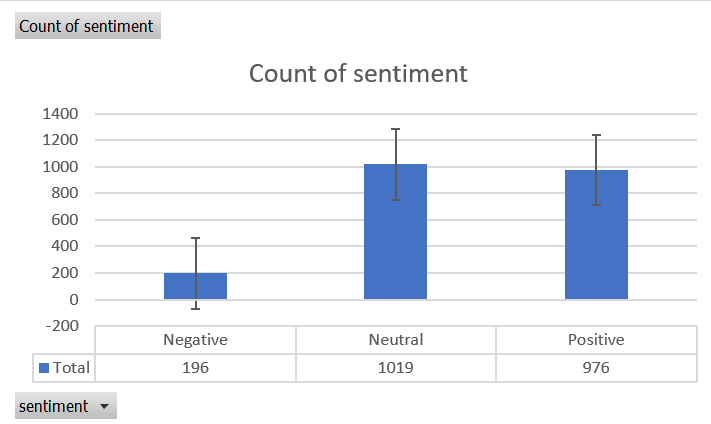


Figure 1 Count of all the Sentiment Labels from 2010 to 2011

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 on an overall scale, while the other two kinds of messages may remain fluctuating.

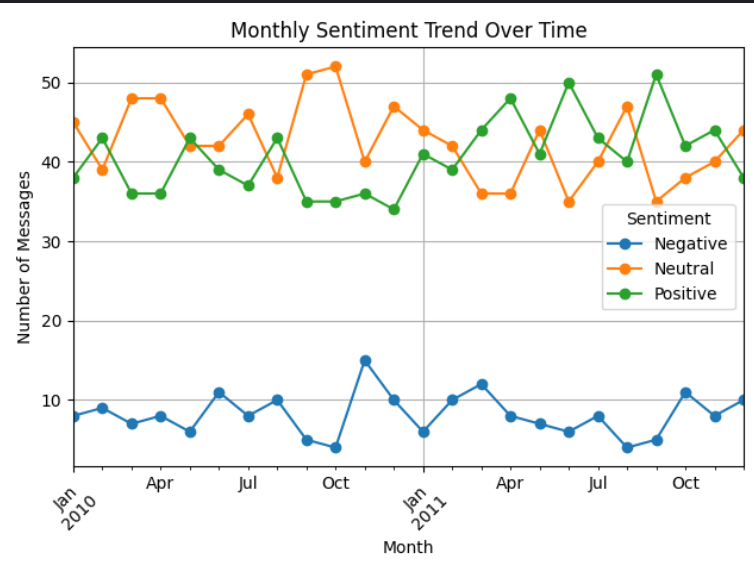


Figure Monthly Sentiment Trend Over Time (2010.1 to 2011.12)

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 from 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 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.

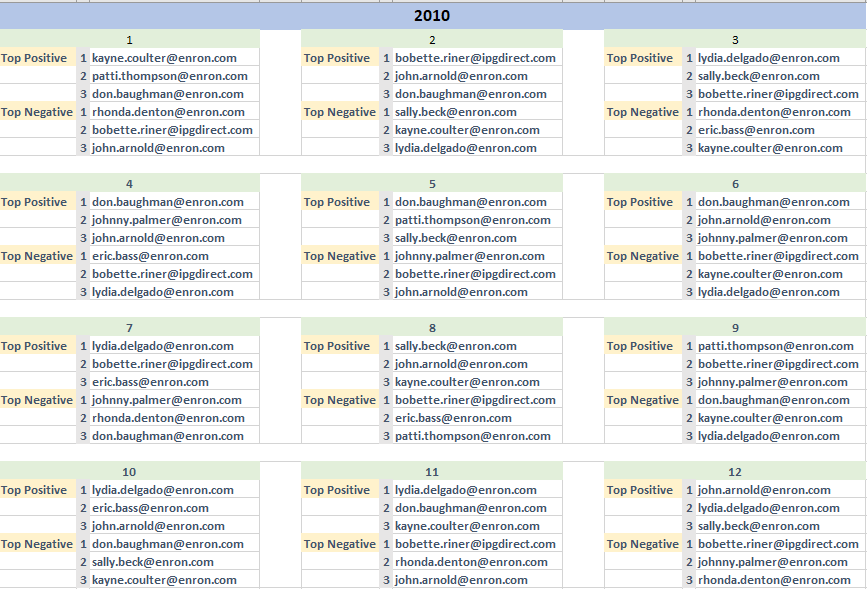


Table Top 3 Positive and Negative employees each month in 2010

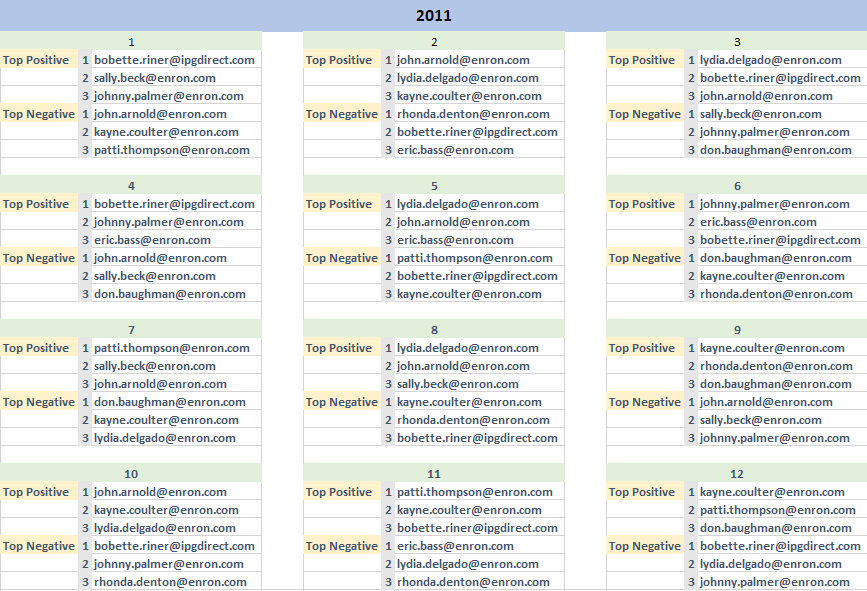


Table Top 3 Positive and Negative Employees each month in 2011

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, listed in the following table.

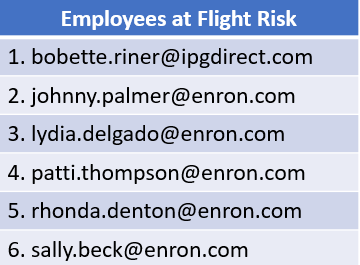


Table Employees who are at Flight Risk

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 that may influence sentiment scores are chosen, which are message count, average word count in one piece of message, the frequency of the three keywords “thank you”, “sorry”, and “best regards” occurring in the messages. 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.

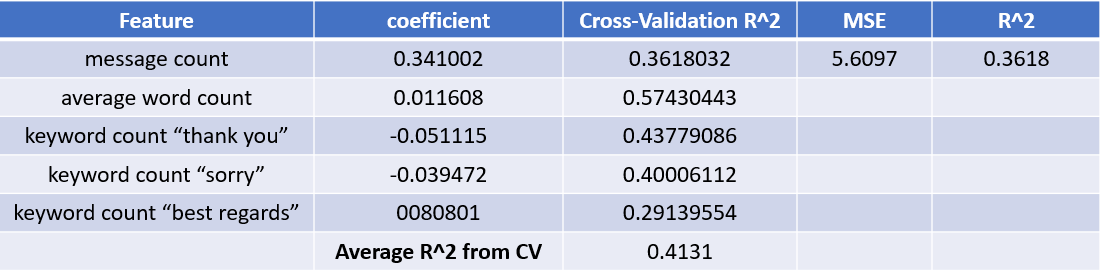


Table Coefficient of Each Feature and Various Metrics Showing Effectiveness

From this table, message count plays a significant role in positively affecting the monthly sentiment score of an employee, while the average word count, frequencies of keywords occurrence, such as “thank you”, “sorry”, and “best regards”, affect the dependent variable slightly compared with message count.

Furthermore, mean squared error and cross-validation R-squared value are used to evaluate the effectiveness of the model. The average R-squared value, which is about 0.41, shows that this multiple linear regression model explains about 41% of the variance in sentiment scores on average. Since the R-squared values are fairly consistent, ranging from about 0.29 to about 0.57, the model used here is stable. Besides, the difference between the single R-squared (about 0.36) and cross-validation R-squared (about 0.41) shows that the model is not overfitting.

To further evaluate the model, the residual plot is also provided.

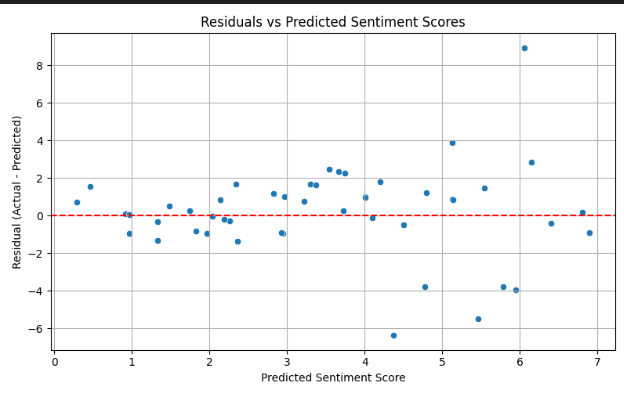


Figure Residual Plot---Residual vs. Predicted Sentiment Scores

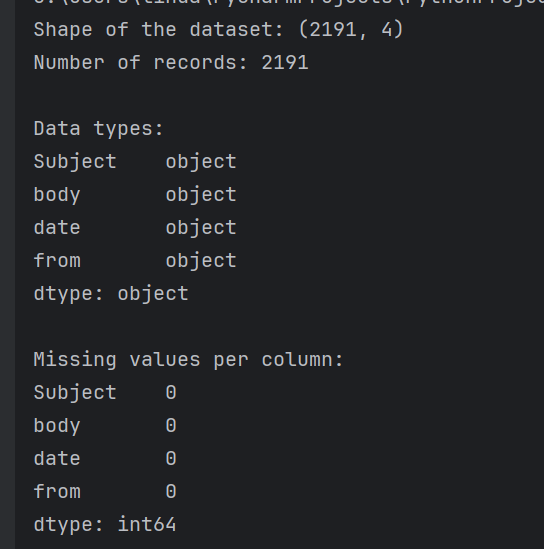
* **X-axis**: Predicted sentiment scores
* **Y-axis**: Residuals (Actual − Predicted)

Since most of the data points in this plot, when the predicted sentiment score is lower, are distributed around 0, linear regression is 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 more complex interactions not captured by this model. 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 2: EDA



Task 5:



Task 6:

