Introduction

As the linguistic dimensions are generated and two datasets are provided separately for the employment as well as unemployment forums, it tends to be curious to find out particular dimensions that bring extensive influences on the scores got toward specific comments. Furthermore, when considering the dates recorded by the unemployment forum are limited within nearly one-year from 2021 to 2022 while the employment forum has data containing almost 11 years from 2012 to 2022, the basic idea utilized here in this analysis should be regarding these two datasets as the train dataset and the test one. To be more specific, the data from the employment forum will be treated as the training base and the other one will be used as the test dataset practiced for the further prediction. Additionally, in order to apply this fundamental idea and go deeper to implement analysis such as clustering and predicting, few initial issues are required to be answered, including the reason why the dataset can be trained for testing the other and the query for which features and linguistic dimensions mainly effect the appraisal and the emotions transmitted. With the purpose of solving these questions, a fundamental analysis with data preprocessing will be implemented and three graphs will be shown as supplementary.

Fundamental Analysis

Data Comparison – the trend similarity

To answer the first question raised above, what has been done initially is comparing two datasets, and the target column is “score” representing the score received by a comment and it indicates the degree of the cognition from the public. For example, a higher score shows that more positive votes are computed than the negative thoughts. Along with the dates recorded, the trend of such scores received can be visualized in a time series and the average score for each year is utilized, and the counts for the scores during a year are also calculated as being shown by the size of the line.

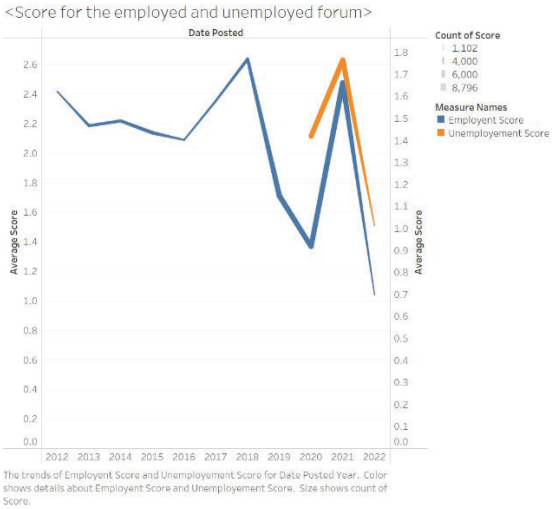


Figure 1.1. Score trend (Employment & Unemployment)

In Figure 1.1, it can be state that the trends of scores existing in the employment and unemployment forums are similar, although the data from the respondents in the unemployed forum is missing before 2020. Furthermore, the vertex detected in 2021 and

drop-offs happened since the later 2021 are synchronous for both datasets, and the numbers of comments left in two forums are significantly decrease in 2022. In this case, the trend similarity is found and it should be reasonable that test and predict unemployment data based on the training model made by the employment data. Simultaneously, the issue about the time-based influences may cause another debt for whether it appears to be rational if a prediction is made according to the model trained by the previous data, as the trends from 2019 express a significant fluctuation and uncertainty. To deal with this problem, the time-series analysis will be introduced in the next section concerning about one of the main topics which this report is going to focus and solve.

Exploratory Data Visualization

Before drawing the second graph, the data needs to be preprocessing to remove outliers and non-meaningful information for the classification; for instance, they are the contents of comments and posted date here. Next, set the feature columns using the linguistic dimensions and the target column as the scores, and they need to be scaled using a standard scalar to become variances used for the further test. In addition, the plotting histograms are offered to better understand the values of each variable due to the prerequisite asked for the feature identification and EDA process.

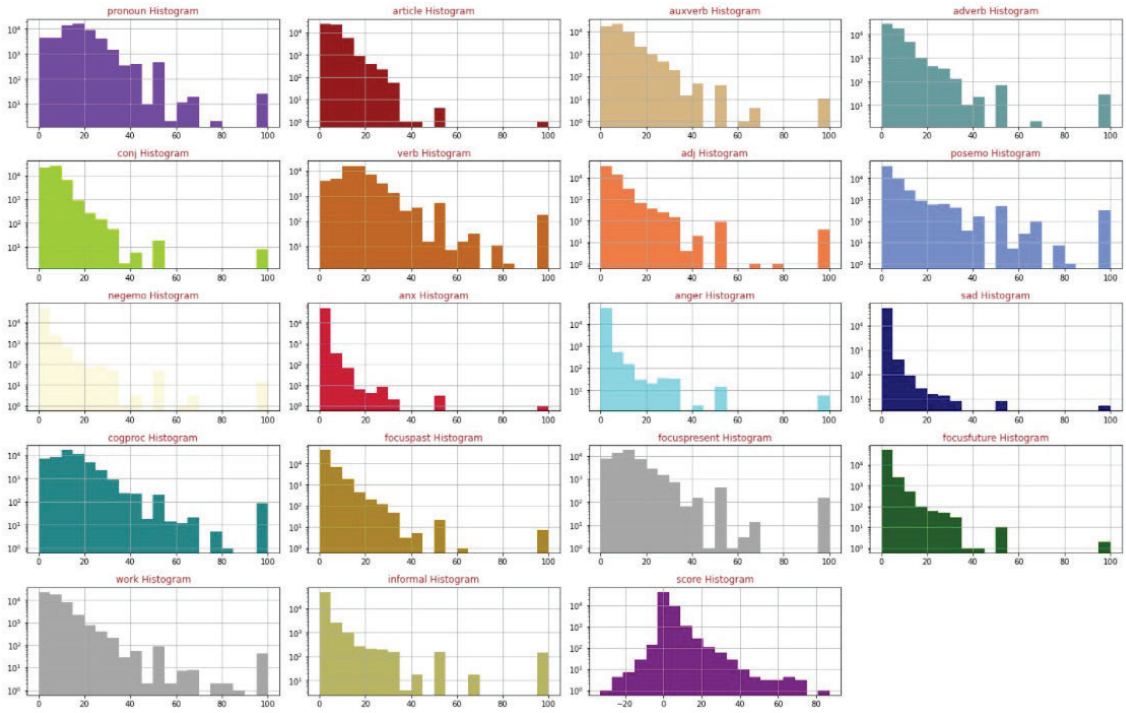


Figure 1.2.1. The single feature histograms

In Figure 1.2.1 where the single feature is tested for its specific influence made on the score's determination, it is a start for exploring the relationship between each categorical condition, and features such as "pronoun", "verb", "cogproc" and "work" dimensions can be generated as a group to be analyzed concurrently.

To tell more about this story, the heat map showing the correlations among different features can be added in this part as well.

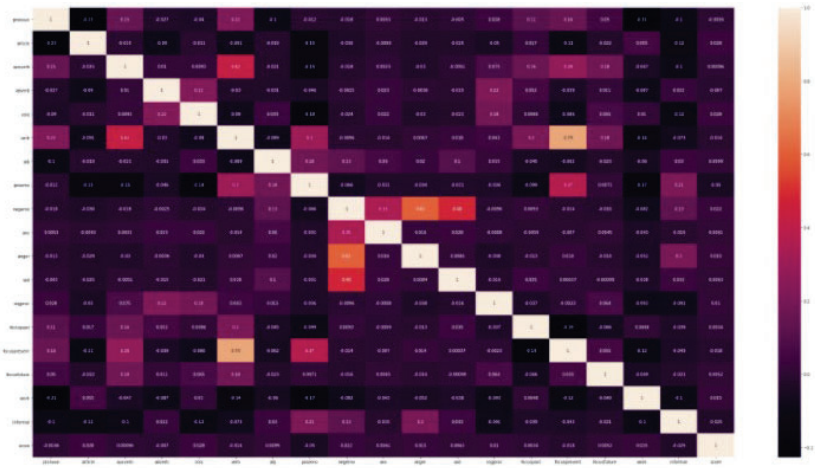


Figure 1.2.2. Heat map

Instead of only concentrating on the score got as the target, all features are tested to find the correlations. In Figure 1.2.2, the high correlations are found between the present focus and verb, and anger expressions, sadness, anxious words with the negative emotion; however, a prominent correlation is not found regarding the score. From this perspective, it may claim that the score got may not affected by the linguistic dimension used in the comments, but the emotions that are willing to be voiced from the authors do relate to the words they used.

Negative Emotions

As the finding discussed in EDA concerning about the high correlations between the upset words and negative emotions with the scores got, Figure 1.3 shows that the more negative the comment states, the lower score it will potentially obtain. What's more, comparing its impacts of gaining scores in the employment forum and unemployed forum, it indicates that the negative expression causes a bigger issue of gaining a lower score when more upset words used in the employment level, while it may lead to a bigger influence in the unemployment forum for reaching a higher score if those sad and anxious words are less used in the sentence. In this aspect, the prediction will be made in the further analysis should take the different effects detected in the different forums into account.

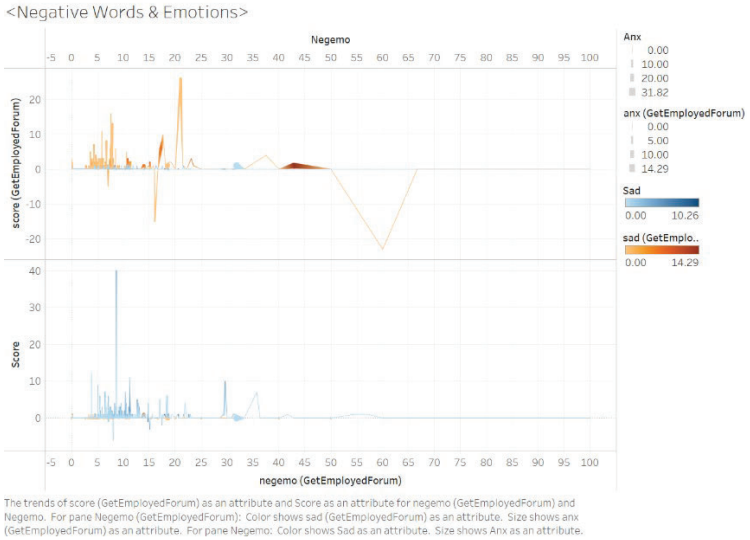


Figure 1.3. The impacts of negative words & emotions

Goal 1. Predicting Analysis

As being discussed above, a prediction will be applied for the unemployed forum based on the training model used the data from the employed forum, and the main task in this section is to predict the scores gained using linguistic features recorded, especially for the negative dimensions involved which appear to have a significant impact in the score gaining. Meanwhile, the time series forecasting will be practiced to adjust the model in order to try to modify the time influences shown in the section above.

Regression Modeling

In order to make a prediction based on the historical records, a linear regression model is applied initially to figure out the correlations between various attributes and scores gained on behalf of the supports from other readers. As being shown in the Figure 2.1.1 representing the target's correlation with different variables, it can indicate that the psychological processes expressed by a certain comment do hold a significantly essential impact on its score, which precisely fits the exploratory analysis practicing on the emotion conveyed by the comment that has been discussed above.

pronoun	article	auxverb	adverb	conj	verb	adj	posemo	negemo
0.001276	0.015008	-0.00633	-0.00583	0.012476	0.001326	0.010106	0.013478	-0.01252
anx	**anger**	**sad**	**cogproc**	**focuspast**	**focuspres**	**focusfutu**	**work**	**informal**
-0.00738	0.01057	-0.00062	0.003769	0.002635	0.001581	0.006857	0.003635	-0.00569

Figure 2.1.1. Linear regression correlated coefficients

Furthermore, when concentrating more on the psychological processes and the linguistic words used, a left-skewed distribution will be detected (Figure 2.1.2). In another words, although the emotional words used in comments can influence the scores received, it is hard to say that such effect works on those comments which contain few psychological patterns but fail to perform the emotions they want to express obviously, and the impacts brought from the psychological processes only lead to the difference when the distinct referring of a passion is felt by a reader. Additionally, not only for the psychological words, other features such as linguistic dimensions are also found to be languorous to control the scores, and the R square of this regression model is calculated as 0.0044, which is doubted to be too low to generate a proper prediction.

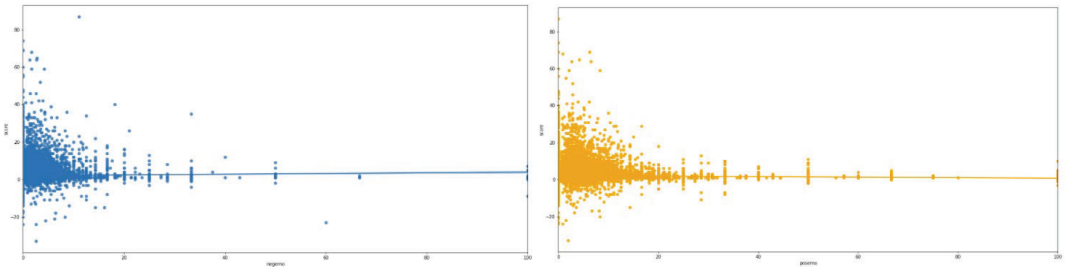


Figure 2.1.2. The negative emotion (blue) & the positive emotion (orange)

Time-series Forecasting

In the regression model, the date when a specific comment posted is not involved in the analysis, and neither did any explanatory analysis practicing before the model establishment concern about the influences caused by the time. Also, with the consideration of the low fitness of the regression model with a low value appeared for

the R square, one probability supposing for this fact could be the time impacts due to the fluctuation appears along the time-series, which may cause a doubt for whether it is adequate to build a model according to the historical data recorded. In this case, the ARIMA model is implemented to capture relationships existing in the data toward the target and aim to explain the autocorrelation between the data points using the past dataset. As the records in the dataset are not daily-based and they are severe fluctuating simultaneously, the time series decomposition should be introduced here by utilizing the weekly average scores received (Figure 2.2.1), where it may be queried that whether there is an accident happening in 2020 as the scores obtained during this time period appear to be significantly low compared with others. To discover the excuses behind this phenomenon, the auto correlation and time-series trend forecast have been applied, where the prediction will be made for the comments in the unemployment forum. However, both the results gained from the correlation with a coefficient rate shown as -0.0004 and the forecasting trend regarding the scores received in 2022 as shown in Figure 2.2.3 suggest that there are no significant influences led by the time. To be more specific, the stable average scores predicted for the year 2020-2022 are around 2.5 to 3.2, which are near the actual value as 2.9 in 2021, stating that there may not have accidents regarding the time impacts but the fluctuation is more likely to be caused by the extreme values received for specific comments.

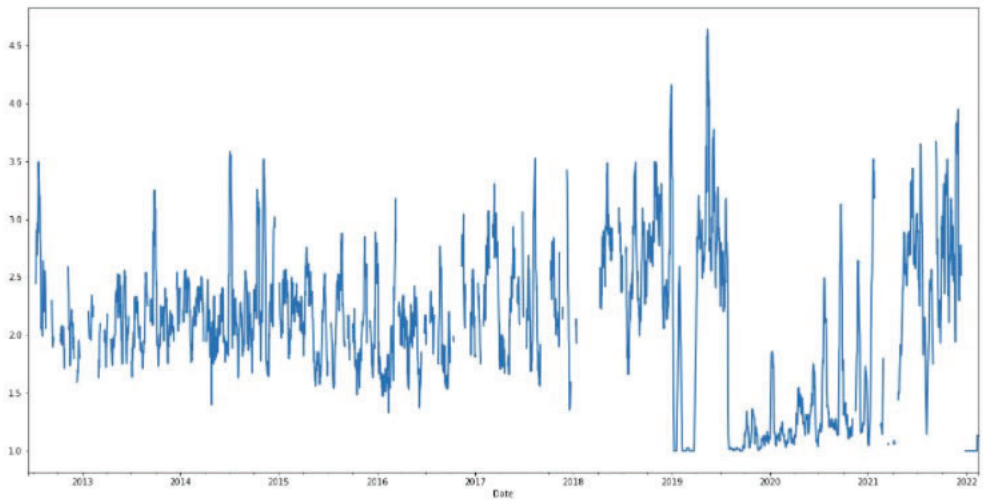
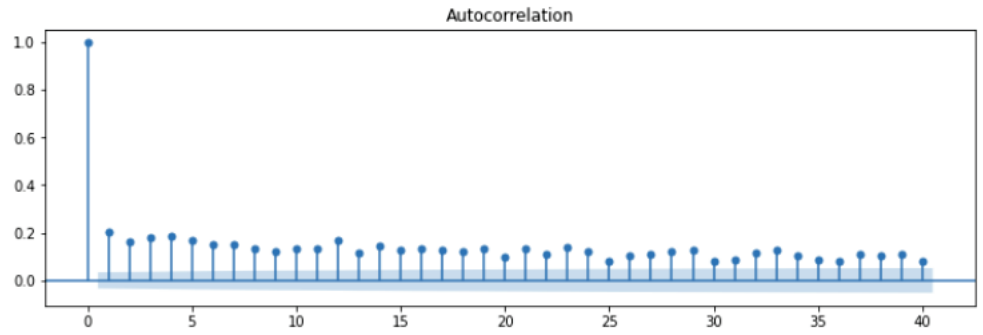


Figure 2.2.1. Score history with the change of time



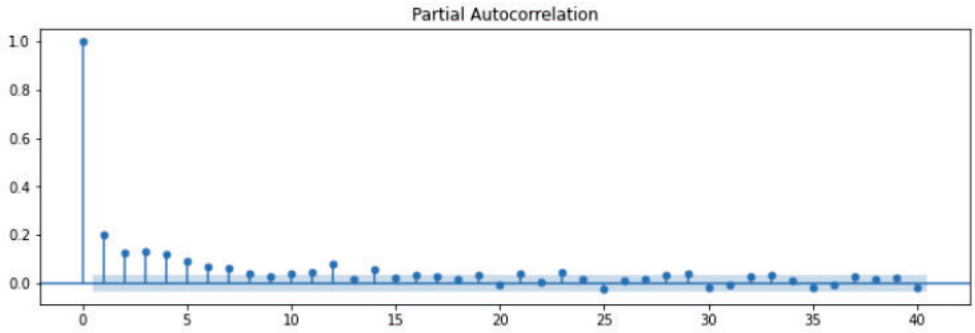


Figure 2.2.2. Auto correlation based on ARIMA

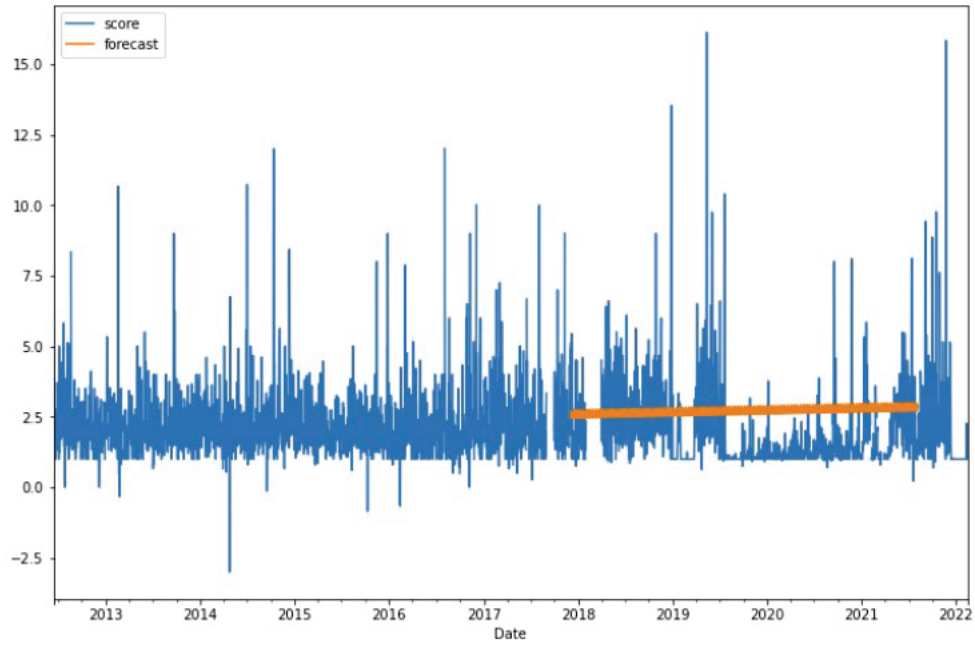


Figure 2.2.3. Time series forecast for the unemployment forum

Goal 2. Classification Analysis

The second goal of this analysis will be the classification for determining certain group of dimensions that can be tested together as a re-organized attribute towards the emotion expressed and score obtained. To deal with this, the SVM method is planned to be contained in this section.

Label Definition

Before moving directly to the classification, the first stuff need to be clarified is to categorize the scores into certain groups, where the distribution of the scores can be taken into account as a tool to assist this process (Figure 3.1). Additionally, to simplify this problem, only the employment dataset is used with 80% for training and the remaining 20% set as the test dataset; simultaneously, three categories of target are defined which represent the high and low scores by using the scores higher than 75% of total distributed scores and the ones lower than 25% of the total, and it assumes the scores are normal distributed. Moreover, to easier practice the model with the data frame processing requirements in Python, the label of each category is converted to the numerical value;

however, no exact meaning for these numbers, and they are created only to meet the demand of the simple operation. Based on these assumptions and data processing, the labels for the three categories are defined as being listed below.

Distribution.INV	<25%	[25%,75%]	>75%
Score value	0.037322	[0.037322,4.077865]	4.077865
Score category	low	normal	high
Numerical label	-1	2	5

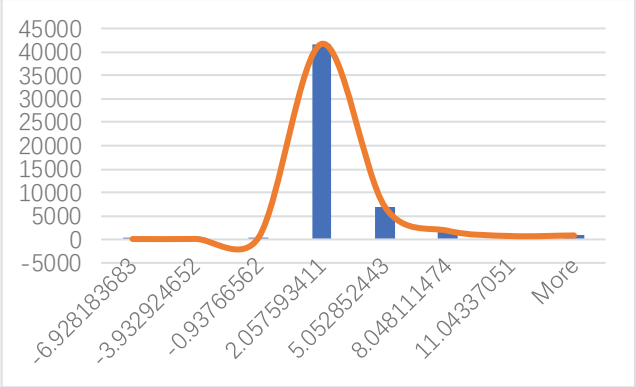


Figure 3.1. Distribution of the scores gained in the employment forum

SVM Modeling

Due to the time-consuming issue existing for running the whole dataset which owns an extremely large size, 5000 samples are randomly selected to allow the software to generate the solution. By applying the SVM model, few features are detected to be available to test together, which are the linguistic dimensions with the emotional expressions, and the time orientations with the personal concerns (Figure 3.2.1). A possible conjecture towards these two groups of variables is that people use related words and tones to write their comments which ultimately are used to express their actual feelings, and such mental suggestions upon the certain emotional expressions and concerns are the main factor for the decisions made by the readers about their acceptance and preferences of these comments.

Furthermore, to make a prediction based on the categories and classification did by the model, the average predicted score appear to be 2.09 for 2021 with a R square as 0.037, which is a bit low but at least higher than the one generated from the former regression model. In addition, when looking at the precision value for evaluating the predicted result with the rate of those being predicted to belong to a specific category are actually correctly assigned, 0.88, appearing as a high value, claims that the classification model works well (Figure 3.2.2). However, the average score is calculated with the single average score of each category and it may make bias as it only considers the distribution of the dataset, while the extreme values for the ones receiving the significantly high or low scores are ignored in this analysis.

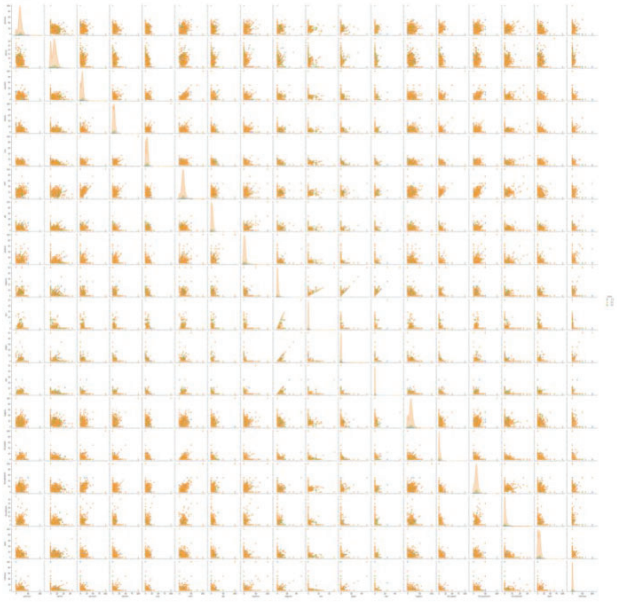


Figure 3.2.1. Multi-class distribution of data across different categories

	precision	recall	f1-score	support
-1.0	0.00	0.00	0.00	10
2.0	0.88	1.00	0.94	353
5.0	0.00	0.00	0.00	37
accuracy			0.88	400
macro avg	0.29	0.33	0.31	400
weighted avg	0.78	0.88	0.83	400

Figure 3.2.2. Model evaluation ratios regarding the tested samples

Future works

As for the future works, here are three main tasks can be done.

- The initial one should be the analysis utilizing the whole dataset for the classification. As the limited conditions prevent the large size data testing, a further study can be implemented on fixing this issue, either by spending more time to run the program or developing the code to make it more efficient which tends to be much more recommended.
- The improvement of the model accuracy. It still leaves some areas to try other predicting models which may be better to fit with this case, and a model search scaling can be included with the purpose of choosing the most proper method and algorithm to establish the model.
- Lastly, few categorical features are suggested to be involved to better describe the data. for example, several categorical features such as locations, the respondents' age and educational level can be involved to support the idea whether the emotional expressions can be accurately conveyed, which may be better to identify the potential influences on the score gaining.

Conclusion

To figure out which features influence the score gaining and make a prediction for the data recorded in the unemployed forum, the similarity of the trends existing in the both

datasets have been found with the following analysis applied to go deeper in calculating the correlations among variables through the regression, evaluating the time impacts based on the time-series forecasting, and making the classification with the regard of the scores' distributions. Although the score making seems to be a subjective stuff and no sufficient evidence provided to state the impacts caused by the change of the time, the analysis suggests that there still few parameters can contribute to the influences on it. Comparing to the other group of features, the emotional expressions are the most essential factors of determining the scores; nevertheless, the phenomenon for whether such emotions can be felt by the readers also play a significantly important role in the score making decisions.