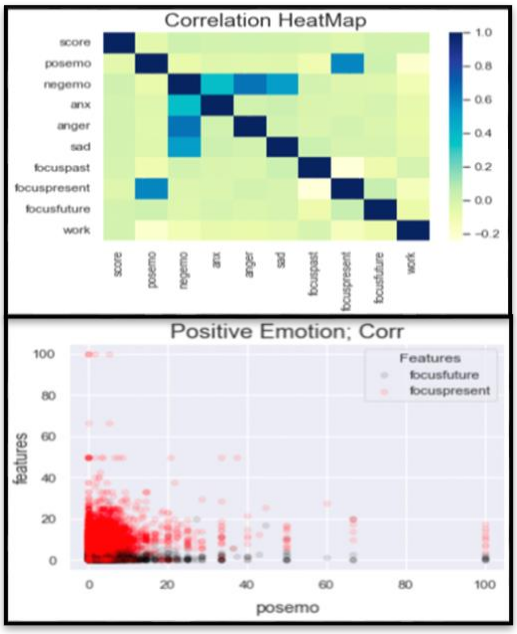
Majdi Mariano
661724154

Final Report

In this report we will evaluate the ‘GetEmployedForum’ dataset, to accurately describe trends and provide ample analytics to gain information. The first step in analyzing the data set is to reduce unnecessary fields and to find and delete all duplicate tuples. From there we will determine the number of important features in the dataset, as well as other features that could play a role in future visualization of the data set. Following the reduction of the dataset we will either normalize or standardize the data to correctly analyze the dataset.

Fields that were dropped include: [datePosted, pronoun, article, auxverb, adverb, conj, verb, adj, cogproc and informal]. These were dropped because they do not give any input into why the comments received the score that they did. In this case fields such as: [posemo, negemo, anx, anger, sad, focuspast, focuspresent, focusfuture and work], could give valuable input into the overall score of the comment. Tuples, in this case specific comments, that were dropped include all in which the score is equal to 1 and more than one of the remaining fields have values inputted.

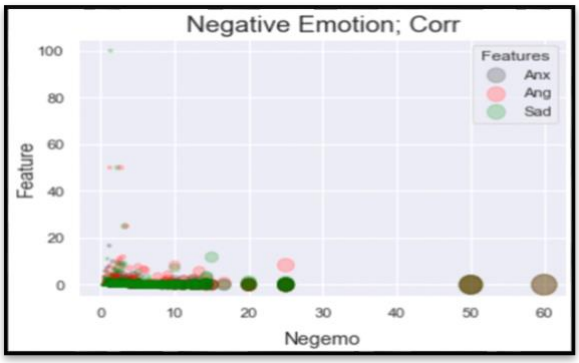


Part 1:

To get an early idea of all the correlations that could be found in the data set we have created a heat map. This will give us a visual representation as to what features correlate to each other and vice-versa.

The next form of visualization that we have chosen to deduce is a scatter plot of both negemo and posemo. These were chosen because they are two features with the highest number of correlations. Likewise, the scatterplot only includes two (posemo) to three (negemo) of the highest correlating features.

In both charts data is normalized and a sample of 5% of the overall population size is taken to increase the legibility of the charts. In the case of the negative emotion, we can see a lot of similarity between the features. In which we can assume, to some degree, that these three features have the same predictive relationship to the negemo value, though that level of predictiveness is to be determined. When looking at the positive emotion, we see more randomness between the two features. This is to be expected as only focus present has a (relatively) high correlating value to posemo.

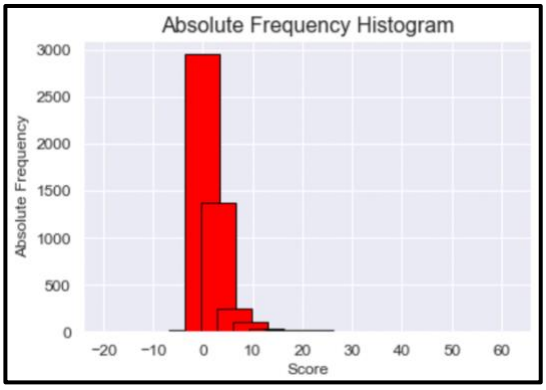


The final chart that was used to better visualize the data set was a histogram. In creating the histogram, we found the need to minimize the size of the bins to better understand the graphs movements, also, we had used a sample of 10% of the population. We can see in the graph below that the histogram is right skewed meaning the mean value is greater than the max value of the data set. It also shows the range in which most scores were placed: between a score of -1.75 and 1.5.

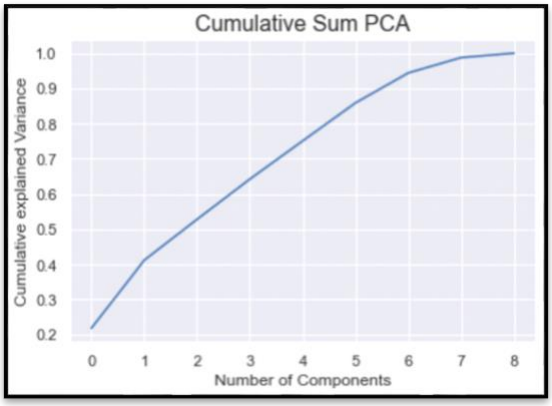
	start	end	class_mark	frec_abs
1.0	-18.00	-14.75	-16.375	3
2.0	-14.75	-11.50	-13.125	0
3.0	-11.50	-8.25	-9.875	1
4.0	-8.25	-5.00	-6.625	1
5.0	-5.00	-1.75	-3.375	15
6.0	-1.75	1.50	-0.125	2950
7.0	1.50	4.75	3.125	1366
8.0	4.75	8.00	6.375	240
9.0	8.00	11.25	9.625	101
10.0	11.25	14.50	12.875	29
11.0	14.50	17.75	16.125	16
12.0	17.75	21.00	19.375	11
13.0	21.00	24.25	22.625	14
14.0	24.25	27.50	25.875	2
15.0	27.50	30.75	29.125	2
16.0	30.75	34.00	32.375	0
17.0	34.00	37.25	35.625	1
18.0	37.25	40.50	38.875	3
19.0	40.50	43.75	42.125	1
20.0	43.75	47.00	45.375	1
21.0	47.00	50.25	48.625	0
22.0	50.25	53.50	51.875	0
23.0	53.50	56.75	55.125	0
24.0	56.75	60.00	58.375	1

Part 2:

In this part we will analyze whether there are any features that are able to predict the score of certain comments. We will use a PCA test to evaluate whether this can be answered or not. What the graph to the right shows is the cumulative explained variance by the total number of components. What is interesting is that there is only one component in which variance can be explained to a (relatively) high degree. Components following this grow at a decreasing linear rate.



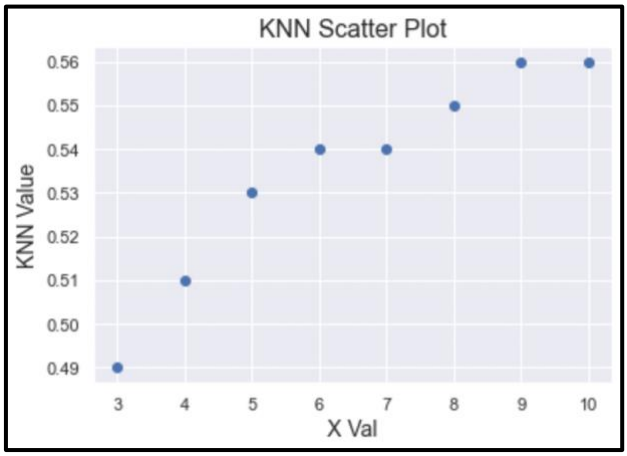
Using train_test_split we utilized the PCA data – to show all 9 components – to run a Linear Regression Analysis to determine the ranking of all feature importance. The Linear Regression Analysis ran with a test size of 80% enabling high predictability. Even with this measure taken, all the features show a low degree of predictability. From this analysis, and looking at the correlation chart shown earlier, it is easy to determine that many if not all features cannot be used to predict to an accurate degree the score that a specific comment would likely receive. What we had found very interesting is the high correlation between futurepresent and posemo. Negemo goes hand in hand with emotions like sad, anxious, and angry, though we personally see no logical explanation between the correlation of futurepresent and posemo.



Feature	Value
work	: 0.18
anx	: 0.11
negemo	: 0.07
posemo	: 0.06
focuspast	: 0.03
sad	: 0.0
anger	: -0.01
focusfuture	: -0.08
focuspresent	: -0.08

Part 3:

The final goal of this analysis is to use kNN to determine the nearest neighbors at a specified value. With an accuracy score of 53% the predictiveness of determining the score from a comment is barely more than that of predicting a coin flip. We believe this is due to the features that were given. Most of which being features that don't necessarily have anything to do with what the comment. It is interesting to see that the personal concern, the work feature, has the highest predictability, though the kNN analysis does not give any input as to why that is, nor does it support its predictability. Based off this analysis and others done previously, we believe it is safe to assume that the kNN analysis is not at fault, but the features itself are.



Part 4:

```
Correlation Sorted

Column Name   Value
_____________________
conj          : 0.02793
article       : 0.02411
negemo        : 0.02241
adj           : 0.01537
anger         : 0.01394
cogproc       : 0.0114
sad           : 0.00461
anx           : 0.00399
work          : 0.00267
focusfuture   : 0.00071
```

The problem what we will be solving in this part of the analysis is evaluate the predictability of each comments score. Within this analysis we will show the process in retrieving the highest correlating features, from there we created a new data frame using only those features in order to run various predictive analysis. The goal is to understand whether this new dataset will produce higher predictability than that of the original dataset, we will also attempt to understand why this could be.

The target feature in this analysis is the score that the comment received. What we had assumed early on is that the targets with higher values would have some sort of correlation to psychological processes, though this was quickly proven otherwise. To the right is a chart of all the correlating features with a value greater than 0. We might be able to assume that these values have the greatest effect on explained variance in the data set. To cross check this assumption we had ran a simple PCA test, evaluating the number of features that make up 90% of the explained variance in the dataset. That array is shown below. From this we can see that there are 10 features that make up 90% of the variance in the dataset. It is interesting to see that the correlation values of the linguistic dimensions are highest, this is most likely due to the frequency in which these words ('I', 'and', etc.…) are used. To clarify, when dropping these linguistic dimensions, the accuracy scores and cross valuation checks do not differ enough to change the overall output.

```
[0.33866471 0.46503849 0.56355455 0.65514085 0.71132883 0.75582118
 0.79690733 0.83664399 0.87335894 0.90312576 0.92778678 0.94951971
 0.96654222 0.98125067 0.98923948 0.99576938 0.99903101 1.        ]
```

To move forward in this analysis, we had started by performing an accuracy test using train_test_split with a test size of 50% and training size of 50%. In this we found an accuracy score of 53% when predicting with K Neighbors Classifier. We had also plotted the KNN cross value scores to visualize whether this test followed the normal sign wave. In this case it did to a small extent.

Moving forward from the KNN analysis, we had run two cross valuations against the Random Forest Classifier. In the first case we had ran the cross val score against a regular standardized data set using Standard Scaler. The results of this analysis were better than that of its predecessor, producing a cross val score of 58.4%. The final analysis was run with a PCA test using 10 components – this was found in the array shown above. This test was also cross valued with the Random Forest Classifier. The results of this were identical to that of the Standard Scaler result, coming in at 58.4%.

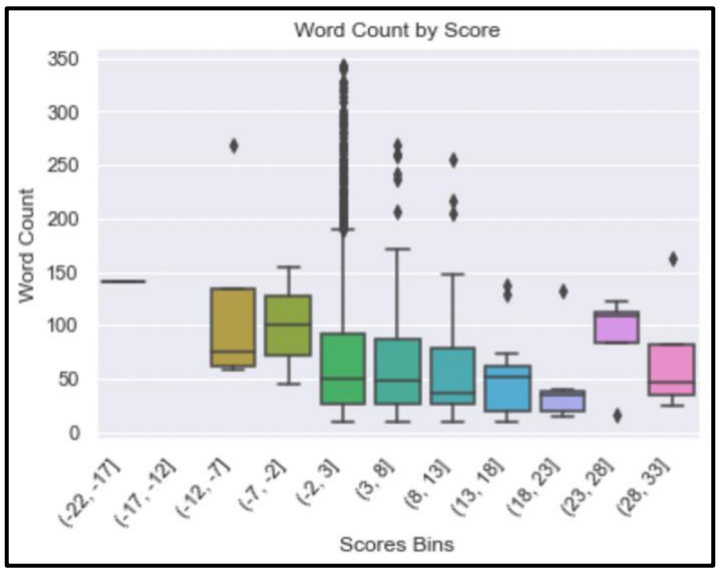
From these two cross val scores and the KNN analysis done earlier we have determined that the features in the data set have little predictability to the target variable. In all cases of prediction/analysis the overall value held little to no weight in the value of the score. This could be due to a large variety of factors that cannot be explained through data features that were provided. What we are presuming is that psychology plays a significant role in the way that users respond to comments posted on the site. In some cases, good

comments will not receive a high score because, and only because, users who appreciate the comment do not have an account or do not think that leaving a upvote or downvote is beneficial in anyway. Another example of why this could be a psychological phenomenon, is due to some users upvoting (or downvoting) comments based off of popularity, that is if comments already have a high number of votes a user may potentially feel more inclined to upvote that comment and vise-versa. These presumptions as well as many more, could be the root to the unpredictability found when evaluating the dataset.

Part 5:

Something that we had found intriguing was to see whether there was any such correlation between the number of words in the comment and the score that it had received. Though, the assumption that there would be any such correlation is very far-fetched, it is interesting to see if any types of patterns will occur in this analysis. In order to start the data set had to be altered once again. Using a sample of 5% the original population size we had created a new row in the dataset called 'Word count' to count the words in the 'Comment Content' column. From there, through simply scrolling through the new data frame, we saw there were a significant number of comments that had little to no meaning. Comments like 'Ahhhhhhhhhhhhhhhhh' received scores of 1 or more. In order to get rid of these comments we ran a loop to look through the length of each word. If the length was unreasonable, we had dropped the entire row from the data frame. After going through such alterations, we had found several other cases in which comments such as 'Thank you' received scores as well. In order to get rid of these we had set the new data frame equal values such that the 'Word Count' column was greater than 10. We had also gotten rid of outliers in which word count exceeded 350 words. We found that, in most cases if not all cases, comments that exceeded a word count had a score in the range of (-2, 2). These steps were important in our analysis because it allows us to get rid of outliers in the data, making it easier to model, read and visualize. Ensuring that our data makes sense, and portrays the right information is critical while evaluating any dataset.

In the graph to the left we see a box plot showing the number of words in the comment content and the associated score that it had received. Since this is only a small fraction of the original data set, we must remember that it is likely that this is not entirely accurate, though there are some patterns that are evident here. We can see that, in scores less than -2 and greater than 13, the average word count differs drastically to scores between that range. One exception remains and that is for scores with a range of (18, 23). Though we have no legitimate explanation to why this is, it is an interesting pattern to be realized. We can also see the number of outliers in the score range of (-2, 3)



showing that the range is the most common value for the scores of each comment. This does not necessarily entail that there is absolutely no correlation between the word count and the score given though it does suggest that it is a small correlating value. After running the correlation that value comes to 0.032201, which is the second highest correlating feature behind negemo.

Part 6:

To conclude the reports entirety, we had evaluated whether there was any predictability between the features in the data set and the target feature (score). In this evaluation we looked at both non-predictive tasks and predictive tasks to see whether or not any predictability was present in the data. In order to come up with the most accurate and time sensitive analysis, we first manipulated the data to exclude certain features and tuples that did not live up to expectations. Features that had did not have to do with psychology, i.e., linguistic dimensions, grammatical dimensions and date posted were removed. We had removed rows that did not hold enough data to be deemed reliable. As mentioned, several times in this paper, these steps are critical in preprocessing in order to reduce run time – though with a data set this small that isn't the main concern – improve readability and accuracy, as well as help find patterns in the dataset. Though, the latter is not a defined perk of preprocessing, it is very much easier to see patterns while manipulating data. The second goal in this analysis was to find whether the number of words in the comment content had any correlation to the score received. In this case more preprocessing was done in order to obtain the correct variable to run this quick analysis. As expected, and stated very early on in this paper, there was no clear feature that had significant enough correlation to dictate the score.

Psychological dimensions played the most significant role, after looking at the visualization it was easy to assume how human behavior could sway the score of a post. Word count also had some sort of correlation, though this was not as easy to understand. All other features had miniscule correlation to the score. Behind all of these correlating values, we had concluded that the predictability was almost equivalent to a coin toss. Showing either fault in the features themselves or fault in the methods of preprocessing and analysis.

In the future we would like to see whether the day of the week or the month of the post has any say in what score a comment received. We predict that, we would be able to see patterns in days of the week or months of posts due to the economy of job markets, being that many students graduate in May, we could assume a spike in posts around that months, though the response to those posts are what we would be looking to find.