TABLE OF CONTENTS

N.B. THE TERM "FORUM" USED IN THE DOCUMENT IN ALL CASES BELOW REFER TO THE COMBINED AGGREGATION OF ALL THREAD DATA

DATA CLEANING

1. Removed all the rows which had a word count of less than 20. We did this because having a low word count does not give reliable values for the language attribute. A low word count will cause the values to be biased to some of the language attributes. We chose word count less than 20 because after researching on the internet, we found out that average number of words used in a sentence are about 20.

2. Removed all negative values (if any) from the language attributes as this would affect mean and other statistical calculations. Also, negative values for language attributes do not make sense.

INVESTIGATION ON LANGUAGE SIMILARITY BETWEEN MEMBERS WHO ARE COMMUNICATING WITH EACH OTHER VIA THREADS AND THE FORUM IN GENERAL

In this investigation, we are analyzing whether there is a language similarity between members who talk to each other directly (via threads) and in the forum in general.

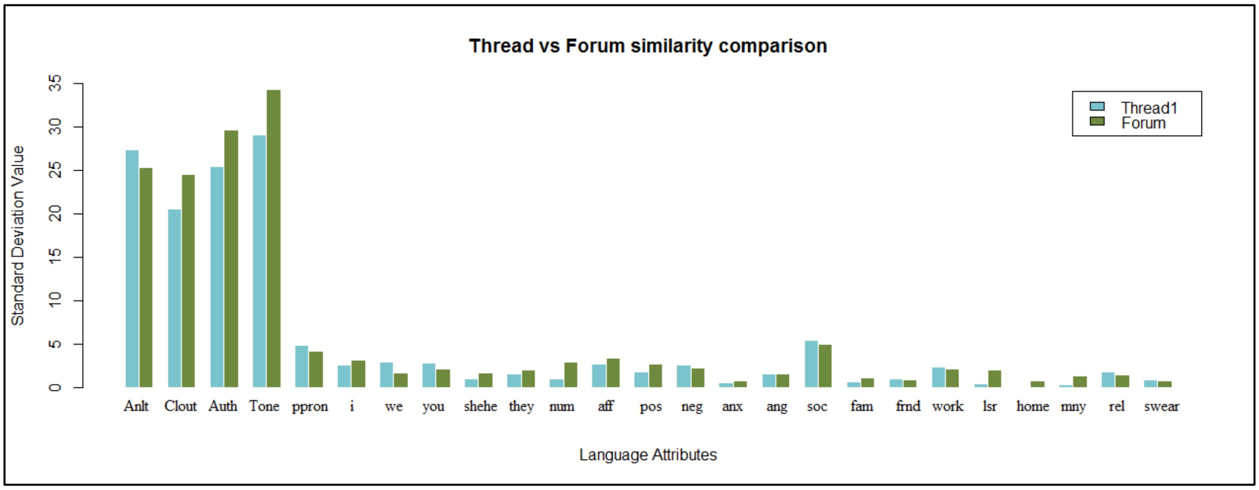
To find this, we will be using standard deviation. As we know, the standard deviation provides some idea about the distribution of scores around the mean (average). The smaller the standard deviation, the narrower the range between the lowest and highest scores or, more generally, that the scores cluster closely to the average score.
But in this analysis, it is easier to think of it as a measure of similar words used between members. If everyone uses the same words, then the standard deviation would be zero and the similarity would be the highest (or perfect). So, if members use similar words, then the standard deviation would be low and similarity would be high and vice versa.

Now to do this, **the procedure followed is shown below**

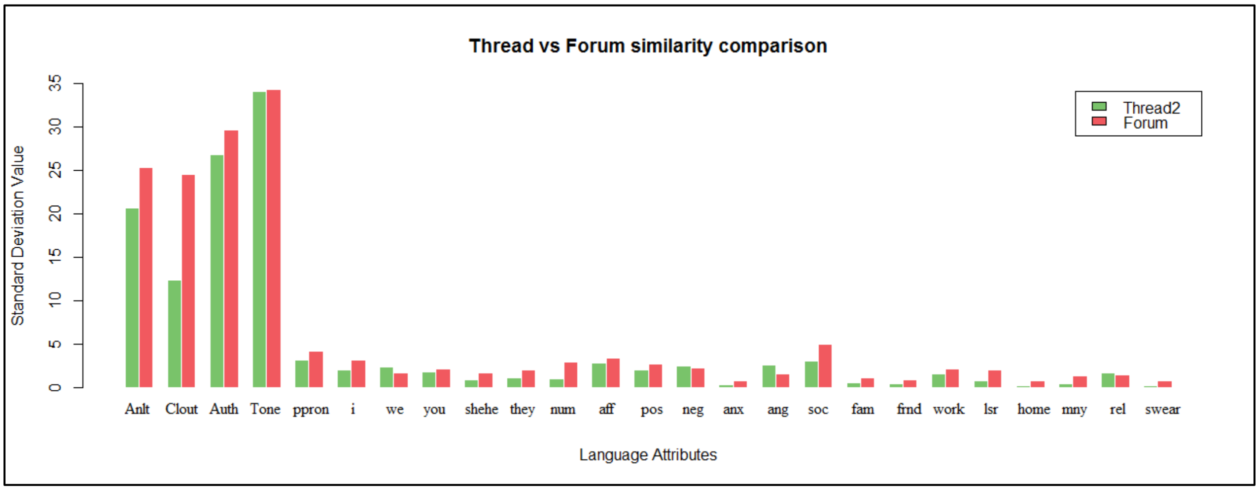
- The standard deviation for each of the language attributes from the forum was calculated.
- After that two different random threads from the forum were grouped and separated. The standard deviation for each of the language attributes according to the threads were calculated.
- Random threads were chosen to remove selection bias.
- The final part was displaying the standard deviation values side by side of each language attribute on a grouped bar plot in a thread vs forum manner.

The results are shown below.

Thread 1



Thread 2



<u>Insights gained from the visualization</u>

As seen above, in both the comparisons, the standard deviation value of thread was lower than the forum in general. This suggests that there is more language similarity between members of a same thread than members of the whole forum.

Although there might be exceptions in a few cases, like in the first image, where the standard deviation value of "Analytic" is greater in the thread than in the forum. This may be due to a lot of factors, mainly the topic of the thread that might have caused this.
But in general, the trend shows that there is more language similarity between members of the same thread.

<u>Reason of visualization choice</u>

A grouped bar plot was used because

- It is an easy representation of data.
- Versatile and Widely Used; making it more visually perceivable
- Allows direct comparison of multiple data series per category

Recommendation

As we know, one author can post multiple times. So obviously, the posts made by the same author will have language similarity.
Therefore if we can remove that ambiguity, we can achieve a better result.

INVESTIGATION ON WHETHER THE PROPORTION OF LANGUAGE EXPRESSING OPTIMISM IS DIFFERENT BETWEEN GROUPS, AND/OR DOES IT CHANGE OVER TIME?

In this investigation, we will be analyzing the proportion of optimism expressed by language in different groups.

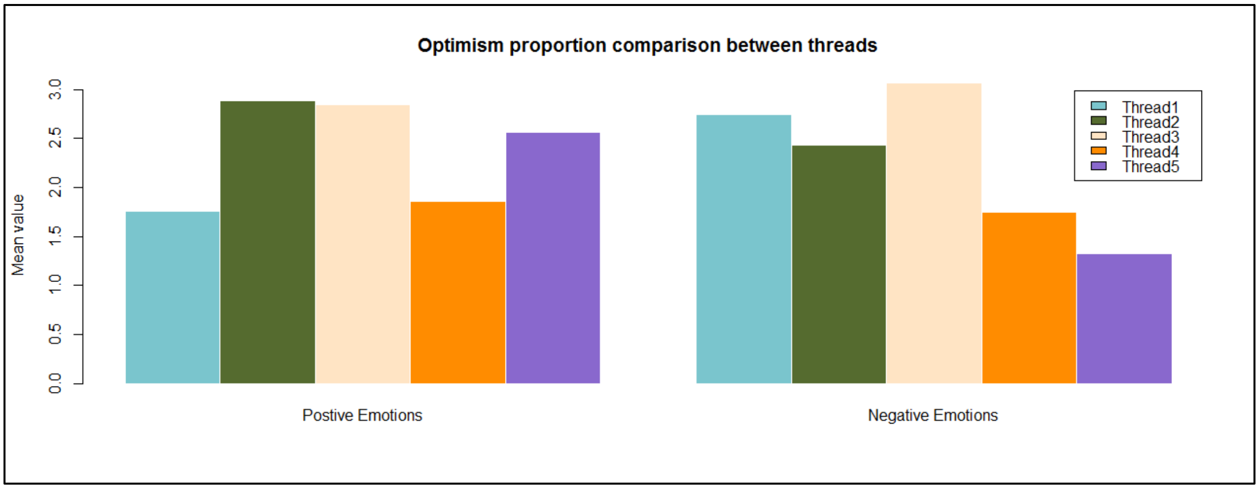
As we know, more optimistic people tend to show more positive emotions and less optimistic people show less positive emotions or even more negative emotions. So, to find this, we will be using “posemo” and “negemo” from the language attributes that are provided to us.

Firstly, we will be investigating whether the proportion of language expressing optimism is different between groups.

Procedure followed for this analysis

- 5 different and random threads were grouped and separated.
- Random threads were chosen to remove selection bias.
- The mean value of “posemo” and “negemo” for each of the threads were calculated.
- Values of posemo and negemo were displayed as side by side of each thread.

The results are shown below.



Insights gained from the visualization

As seen from the above graph, the positive and negative emotions for each group are quite different. The positive emotion for Thread2 and Thread3 are similar but the difference in their negative emotions is apparent. As stated earlier, negative and positive emotions are related to optimism.

This shows us that different groups have different proportion of expressing optimism.

Reason of visualization choice

A grouped bar plot was used because

- It is an easy representation of data.
- Versatile and Widely Used; making it more visually perceivable
- Allows direct comparison of multiple data series per category

--

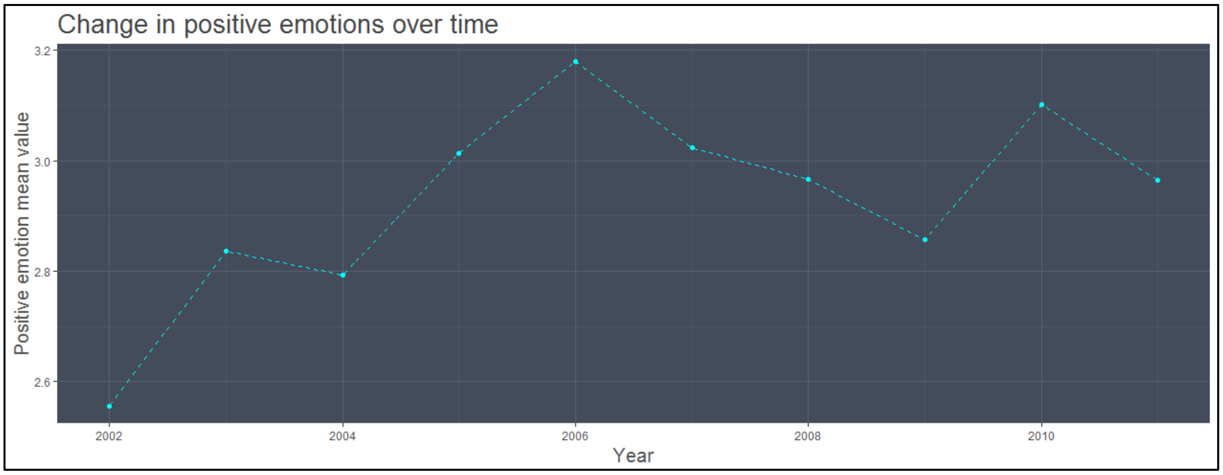
NOW WE WILL BE INVESTIGATING WHETHER THIS CHANGES OVER TIME.

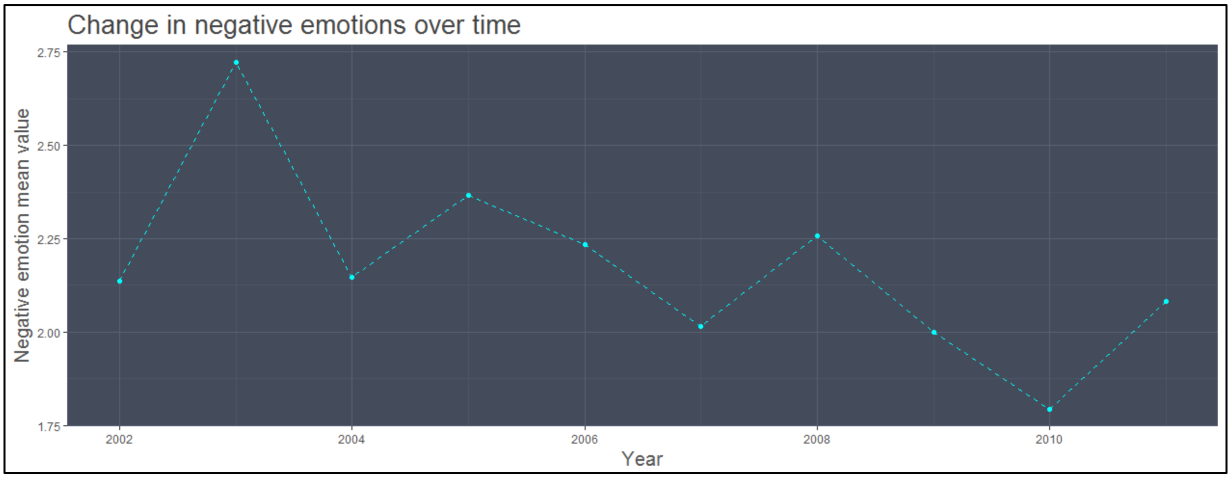
--

Procedure followed for this analysis

- Data was grouped according to Year (calculated from the “Date” attribute of the dataset)
- The mean value of “posemo” and “negemo” for each year was calculated.
- The values were plotted over a line graph in a “value vs year” format.

The results are shown below





Insights gained from the visualization

As seen from the above data, the positive and negative emotions do change over time and therefore so does expression of optimism. This can be caused due to many factors, like tragedies happening or something good happening etc.

Reason of choice of visualization

A line chart was used because

- It is effective when showing how things change over time
- Values between data points can be estimated and extrapolated into the future
- It is simple to construct and read

INVESTIGATION ON WHICH CATEGORY OF WORDS WERE USED MOSTLY

In this investigation, we will be analyzing which category of words have been used a lot using a word cloud. We will be doing this for both the forum and a specific thread.

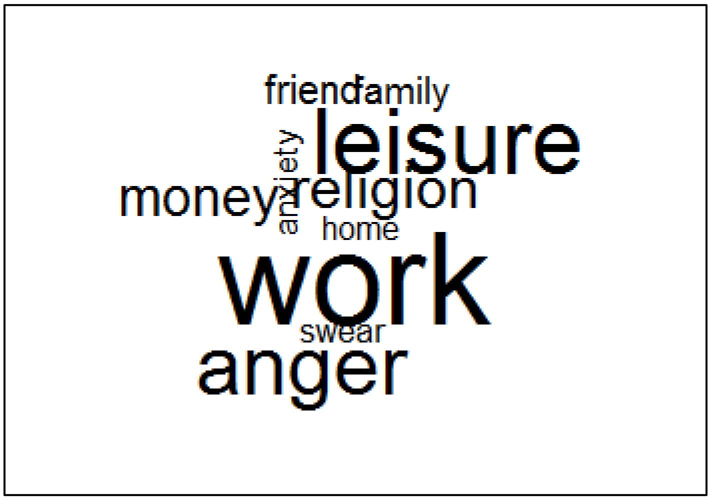
Procedure followed

- Mean of every language category was aggregated and calculated for the forum.
- A specific thread was grouped and separated and the mean of every language category of that thread was calculated.

Note: The word category "social" was removed and not used because it had a very higher value in proportion to the other categories, hence overshadowing them. So, for a fair comparison, the word cloud was generated without the "social" category.

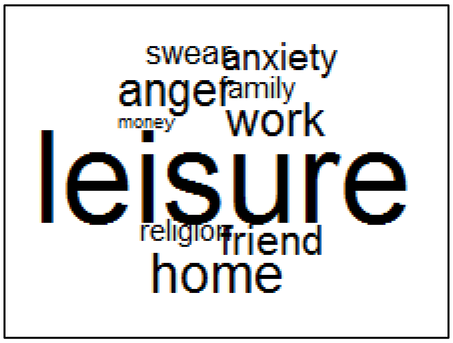
Insights gained from the visualization

The word cloud for the whole forum



As we can see here, there is most emphasis on work, closely followed by leisure and anger in the forum. This gives us a brief overview on what topics are being discussed.

Here is a word cloud from a specific thread



Here, leisure is the unanimous winner. We can make an educated guess that this thread mostly has content concerning leisure talks.

Reason of choice of visualization

A word cloud was used because

- They're fast and very easily perceived as it reveals the essentials.
- They delight and provide emotional connection.

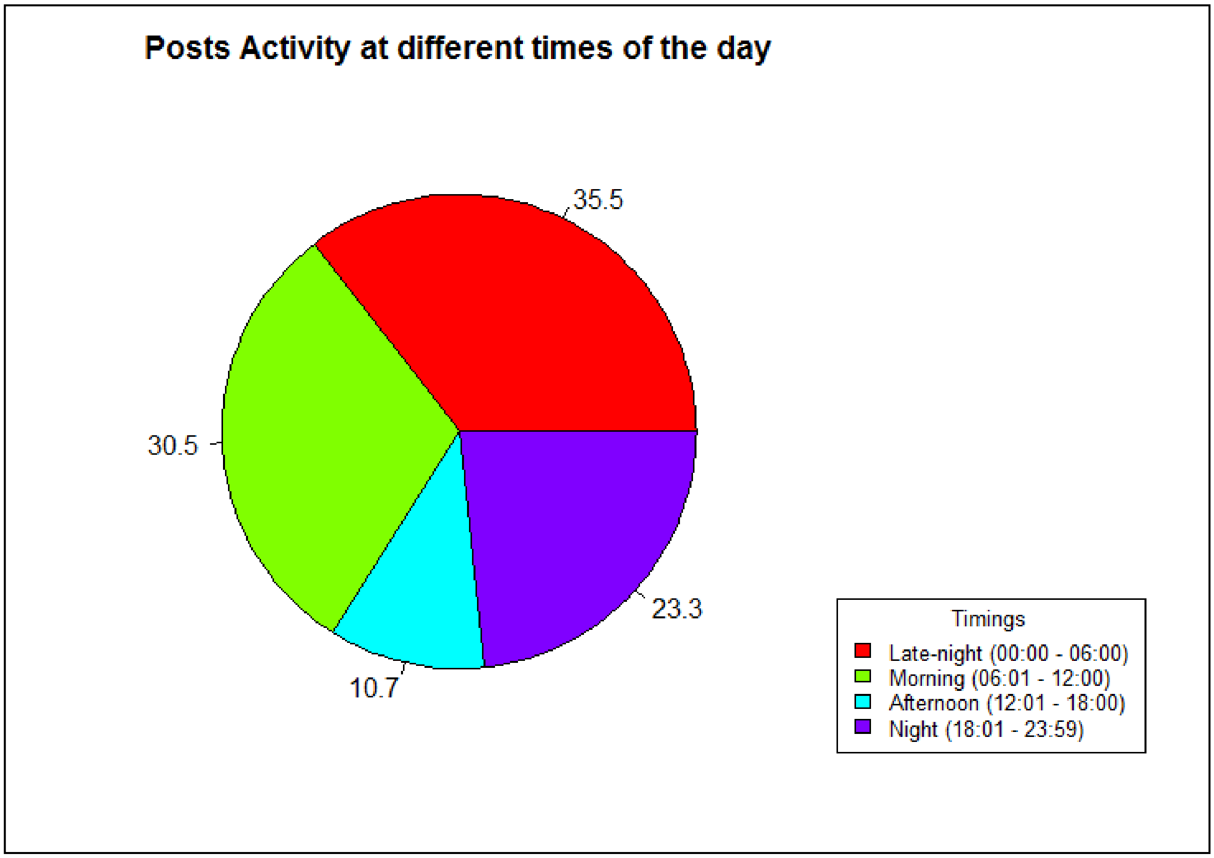
INVESTIGATION ON HOW POST ACTIVITY VARIES AT DIFFERENT TIMES OF THE DAY IN A FORUM.

Procedure followed

So, time in the dataset was given as the raw time. We converted the raw time into a more abstract form of "Times of the day". Times of the day we included were

1. Late-Night (00:00 – 06:00)
2. Morning (06:01 – 12:00)
3. Afternoon (12:01 – 18:00)
4. Night (18:01 – 23:59)

Then we created a pie chart to visualize the data which is shown below.



Insights Gained from the visualization

We can see that time of the day which has the most amount of posts is at “late-night” followed by “morning”. This could be useful in many number of ways like:

1. People can post for example at “late-night” so that they get a wider audience of people to see their post. For example, if we took our Moodle forum and applied the same visualization, students could figure out at what time of the day would be best for them to post in the forum to get a reply as fast as possible.

2. Also, the people who handle the backend side of the servers can optimize resource use by handling the traffic on peak times of the day better using Load Balancing techniques.

Reasons for using a Pie Chart

It is good way of displaying classified data and they are a very good visual way to compare proportions and are also easy to interpret.

INVESTIGATION ON TYPE OF LANGUAGE USED BY AN AUTHOR.

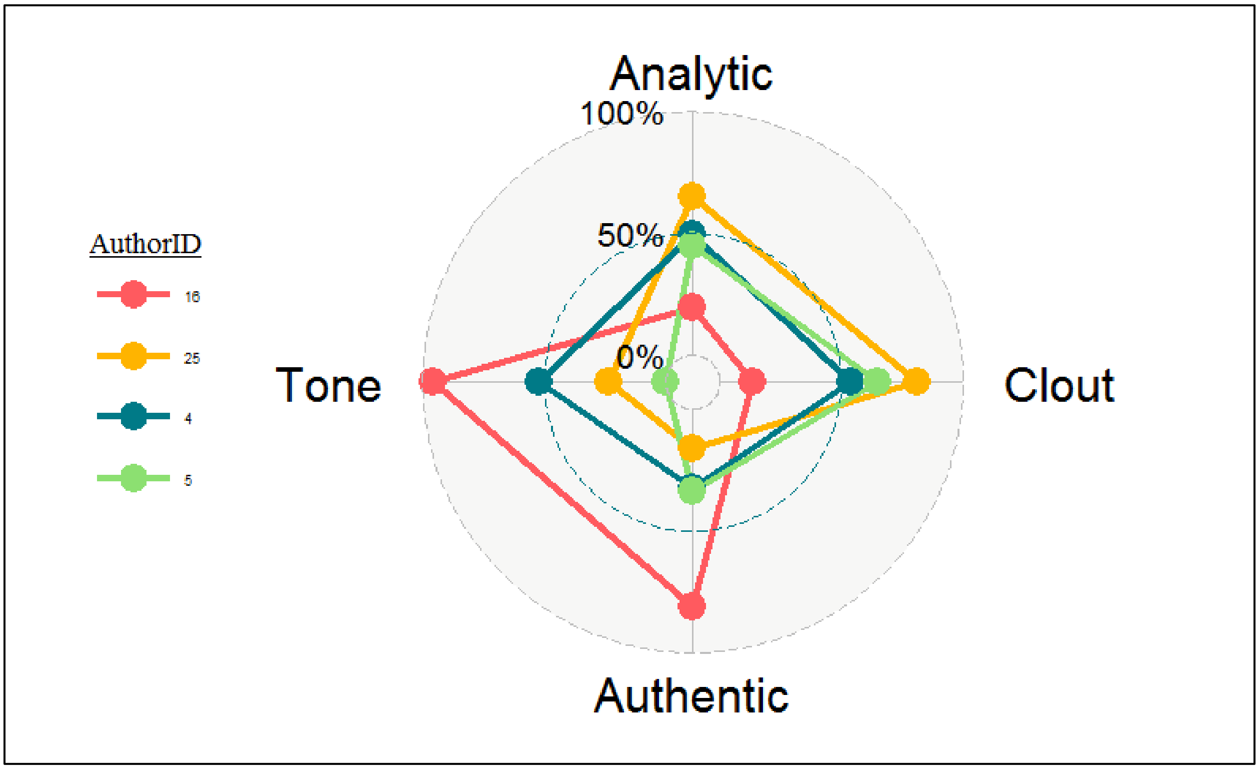
Procedure followed

So, the attributes to find the type of language used by an author were

1. Analytic
2. Clout
3. Authentic
4. Tone

We used these attributes because these attributes give the summary of the language the author uses. So, for each author, we took the mean of each language attribute from all the posts.

Then we visualized this data with a radar chart which is shown below.



Insights Gained from the visualization

For example, we could guess if a person is trustworthy by seeing how authentic he is from the visualization. We could also say how influential a person is by looking at his “clout” value from the visualization.

Reasons for using a Radar Chart

Radar charts are a very good way of showing different types of performance metrics.

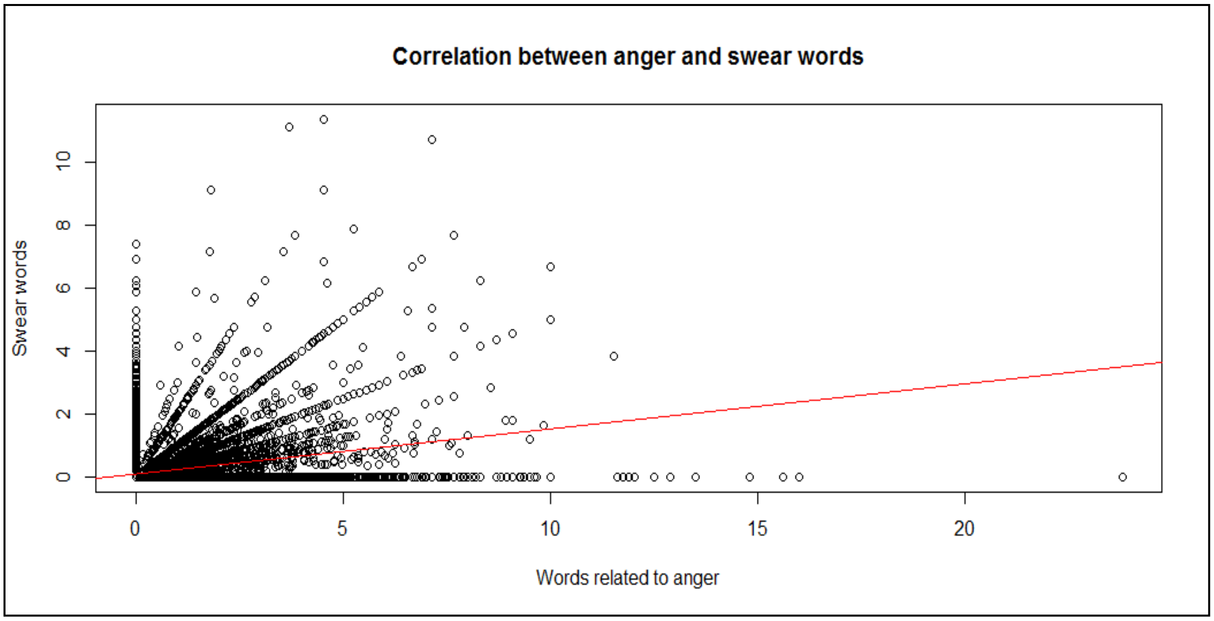
Recommendation

If we had better attributes to deal with it, a lot more information could be found out about the author. For example, if we had attributes related to personality traits like openness, conscientiousness, extraversion, agreeableness etc., we can find out the personality of each author.

INVESTIGATION ON RELATION BETWEEN DIFFERENT CATEGORIES OF WORDS.

- Swear and Anger

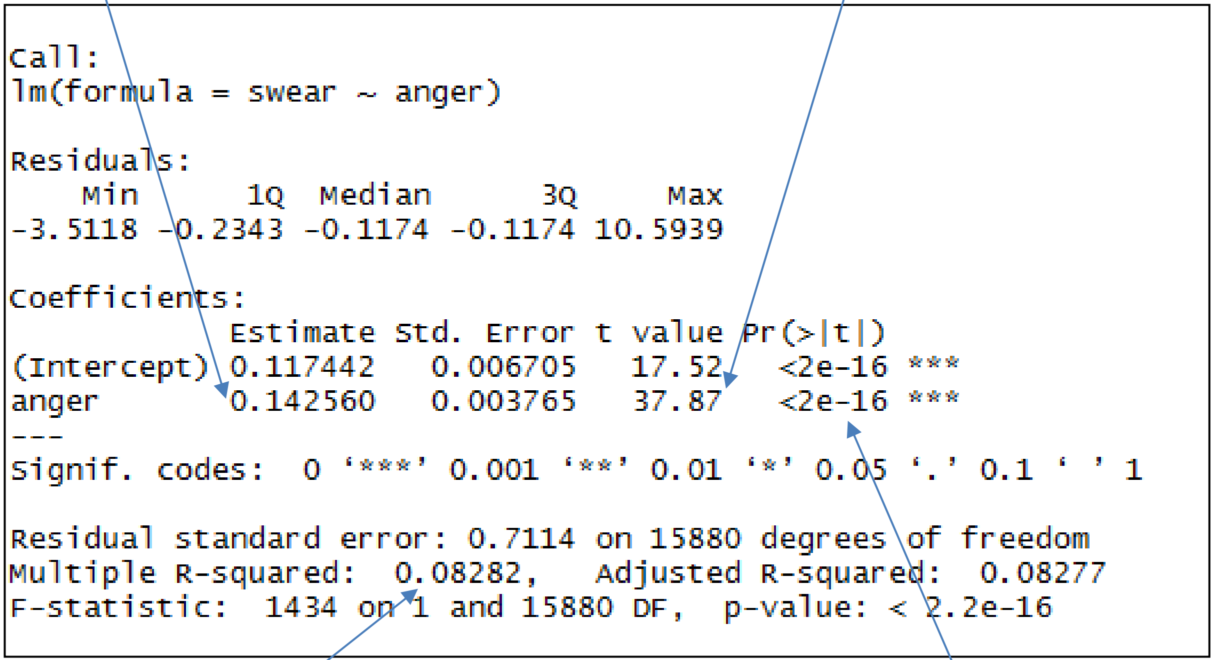
Trying to find out if anger has any effect on swear. In general, we would expect that anger would have a very strong relation with swear. To find a relation we used a scatter plot and did a linear regression. The visualization is shown below



Just looking at the regression line, it seems that as words related to anger increases, it has an increase in the number of swear words. To prove a relation exists, we used the statistical summary of the linear regression which is shown below

This is the slope which is positive. This shows there is a positive trend.

The t-value is far away from 0. So, this indicates we could reject the null hypothesis. Hence, we could declare a relationship between anger and swear.



```
Call:
lm(formula = swear ~ anger)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5118 -0.2343 -0.1174 -0.1174 10.5939 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.117442   0.006705   17.52   <2e-16 ***
anger       0.142560   0.003765   37.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7114 on 15880 degrees of freedom
Multiple R-squared:  0.08282,	Adjusted R-squared:  0.08277 
F-statistic:  1434 on 1 and 15880 DF,  p-value: < 2.2e-16
```

The multiple r squared value says that 8% of the variance found in swear can be explained by anger. This value is very small. Hence our initial assumption that anger has a very strong relation is not true. Hence this means there would be other emotions that influence swear. Other emotions could include frustration, joy, surprise etc. Also swear words can also be used in a positive manner, in the form of jokes and humor.

The p-value is very close to zero which indicates that it is unlikely that we will observe a relationship between anger and swear due to chance.

LIBRARY PACKAGES USED

Library Used	Reason
Tibble	They are data frames that forces you to confront problems earlier leading to cleaner and more expressive code.
Dplyr	Used for easy data manipulation
Chron	Used for representing dates and times of the day.
Wordcloud	To draw word cloud
ggplot2	To draw line chart
Ggradar	To draw radar chart
Scales	Easy way to convert from data values to perceptual properties like legends for visualisations.

R code

Please refer to the attached R file.