FIT3152 Data analytics Assignment 1

Student Name: Kuah Jia Chen

Student ID: 32286988

Question a.1)

How active are participants over the longer term (that is, over months and/or years)? Are there periods where activity increases or decreases?

To figure out how active the participants are over the longer term (i.e., over months and years), I would need to pre-process the data to produce the suitable graph to analyse the activeness of participants over time. Firstly, I would need to group by the whole dataset based on the months and years (i.e., columns with the same months and years will be grouped together) and calculate the number of occurrences for each group. By doing so, I will be able to receive a dataset that consists of the number of occurrences of posts for each month of years.

```
> head(date_table)
# A tibble: 6 x 2
  Date       No_of_post
  <date>          <int>
1 2002-01-01        118
2 2002-02-01         55
3 2002-03-01         66
4 2002-04-01         86
5 2002-05-01        125
6 2002-06-01         34
> mean(date_table$No_of_post)
[1] 166.6667
```

Figure 1. The first six rows of the dataset that will be used to generate graph later and the average of post

Note: Please ignore the day (i.e., xxxx-xx-01) for each row, it does not represent anything. The day also will not be considered when plotting the graph.

In figure 1, we can see that the dataset is now grouped based on months and years and the second column indicates the number of posts in that month. Therefore, the dataset is now ready to produce graphs. **The first six rows of the dataset that will be used to generate graph later and the average of post**

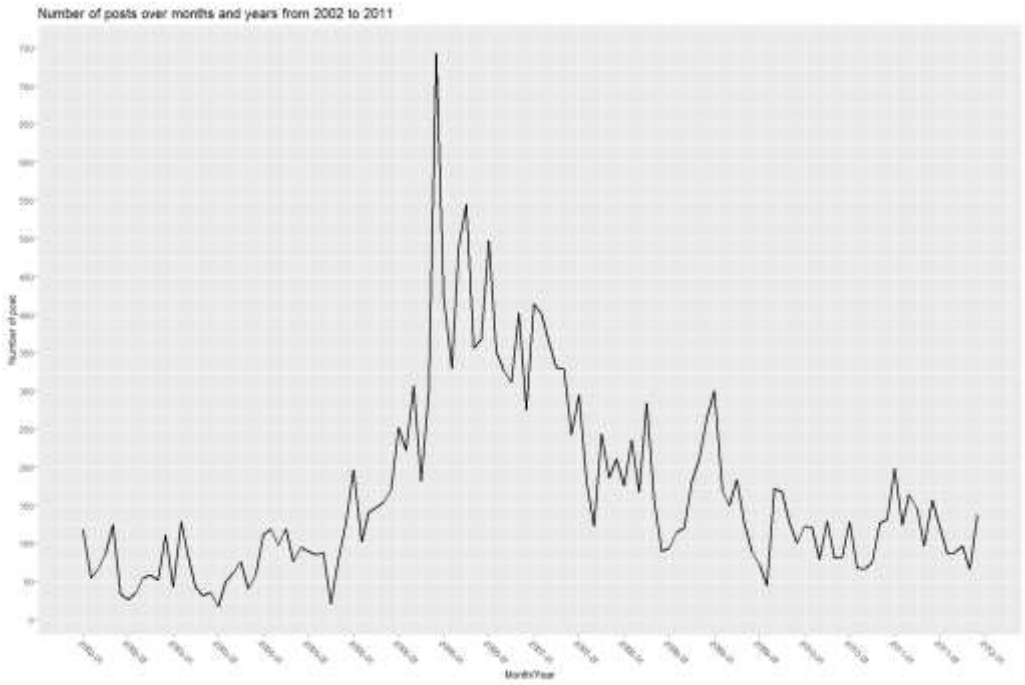

Figure 2. Time series graph that shows the number of posts over months and years from 2002 to 2011

As in figure 1, we know that the average number of posts for each month/year is around 167. Therefore, I would say the participants are quite active looking at this number. From figure 2, we can see the activeness of the participants by considering the number of posts throughout

the whole period. Initially from Jan 2002 to July 2004, the activeness of the participant is stable with some variation. However, the time series graph shows that during the period from July 2004 to Jan 2006, there was a continuous increase in the number of posts (although there was some fluctuation) and even reached roughly 750 posts around Jan 2006 eventually. This indicates that the activeness of the participant is rising during that period. Nevertheless, as shown in figure 2, I realized that after Jan 2006, the number of posts started to decrease. Especially the period from Jan 2006 to July 2009, although there was still some fluctuation, overall, the number of posts kept decreasing on average. Nonetheless, after July 2009, the number of posts started to be relatively more stable.

In the previous paragraph, I had said that in the period from July 2004 to Jan 2006, there was a continuous increase in the number of posts compared to the period before July 2004. To support my finding, I will do a hypothesis testing, where the null hypothesis is the mean of posts in the period from July 2004 to Jan 2006 is greater than the mean of posts before July 2004. The confidence level is 95% as default.

Figure 3. shows that result of the hypothesis testing using t.test in R

According to the result shown in figure 3, we can see that the p-value is 0.9972. It means that if the null was true, then the chance of observing a sample with an extreme or more extreme difference from the null as the one we observed is about 99.7%. There is insufficient/weak evidence against the null hypothesis that the mean of posts in the period from July 2004 to Jan 2006 is greater than the mean of posts before July 2004, since p-value > 0.05. Therefore, we cannot reject the null hypothesis. Hence, this supports my statement that the activity does increase during the period from July 2004 to Jan 2006.

Similarly, I will do a hypothesis testing to prove that the activity does decrease in the period from Jan 2006 to July 2009. However, I will only compare the mean of posts of this period with the period from Oct 2005 to Jan 2006, the reason is because the second period (i.e., from Oct 2005 to Jan 2006) is roughly the peak of the graph. Therefore, it is fairer to compare these two periods as this tells us whether there is a decrease after that instead of just getting the mean of post before Jan 2006. In this case, the null hypothesis is that the mean of posts in the period from Oct 2005 to Jan 2006 is greater than the mean of posts in the period from Jan 2006 to July 2009. The confidence level is 95% as default.

Figure 4. shows that result of the hypothesis testing using t.test in R

According to the result shown in figure 4, we can see that the p-value is 0.7539. It means that if the null was true, then the chance of observing a sample with an extreme or more extreme difference from the null as the one we observed is about 75.4%. There is insufficient/weak evidence against the null hypothesis that the mean of posts in the period from Oct 2005 to Jan 2006 is greater than the mean of posts in the period from Jan 2006 to July 2009, since p-value > 0.05. Therefore, we cannot reject the null hypothesis. Hence, this supports my statement that the activity does decrease during the period from Jan 2006 to July 2009.

Is there a trend over time?

By looking at the figure 2, we can easily see the trends of the number of posts over time. The shape of the time series graphs seems to look like a parabola shape. The number of posts kept fluctuating from Jan 2002 to July 2004. After that, the number of posts started to increase significantly until Jan 2006, however it started to decrease from Jan 2006 to July 2009. Eventually, after July 2009, the number of posts remained relatively stable with some fluctuation until the end of 2011.

Question a.2)

Looking at the linguistic variables, do the levels of these change over the duration of the forum?

For me to visualise how all the levels of linguistic variables change over the duration of the forum, I will plot several line graphs to achieve this objective. Moreover, instead of including all 18 lines in one graph, I had decided to plot three graphs. The first graph is for the four summary variables, the second graph is all the variables related to pronouns, and the third graph is for the rest of the linguistic variables. Nevertheless, I did not analyse word count in this question, the reason is that I do not think word count will be useful to figure out the relationship between linguistic variables as word count does not indicate any functional properties of a post to answer this question.

To plot the line graph, I need to pre-process the data beforehand. Firstly, I will group by the linguistic variables for each post based on the month of each year and after that, find the mean of the value for each month of each year. The reason why I use the mean of the percentile/proportion of the posts within a specific month of a specific year of a linguistic variable is that this helps me to visualise better how the level of the variables changes over time as it makes the fluctuation of values to be more obvious in the graphs.

```
> head(webforum_four)
        Date Average_of_Analytic Average_of_Clout Average_of_Authentic Average_of_Tone
1 2002-01-01             57.89390          68.62008              32.45754         36.70695
2 2002-02-01             61.02491          69.83873              33.41600         42.19982
3 2002-03-01             61.59879          67.16364              32.04273         37.12485
4 2002-04-01             58.47233          59.73709              39.54360         40.99802
5 2002-05-01             59.69344          60.69296              33.36536         46.97160
6 2002-06-01             65.14029          65.73765              23.58382         51.67588
```

Figure 5. shows the first 6 rows of data frame after pre-processed the data

In figure 6, we can see that after doing the group by operation, I will eventually obtain this data frame and now I could use it to plot the line graph, which will be shown in figure 7. **Please note that I will do the same group by operation to all the linguistic variables.**

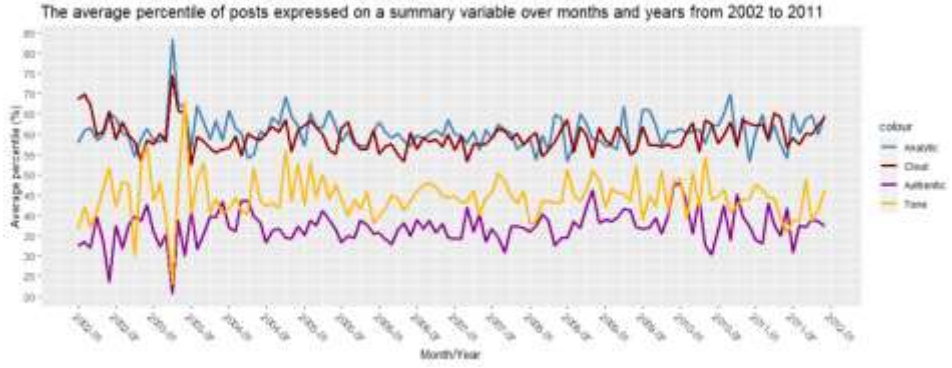

Figure 6. shows the average percentile of posts expressed on a summary variable for each month of each year

In figure 7, I realized that the mean percentile for Analytic and Clout agree over time as the lines are quite close to each other. Besides that, most of the time, the average percentile for Analytic and Clout are between 55% to 70%. While for Authentic, its mean percentile is lower than Analytic and Clout over time (i.e., frequently in the middle of 35% to 50%), except that there was a significant rise in around July 2003, where the mean percentile reached about 70%. The mean percentile for Tone often was the lowest among summary variables throughout the period. It's mean percentile usually among 30% to 45%. Therefore, overall, we can say that the level of the four summary variables does change over time but most of the time is still within a specific range.

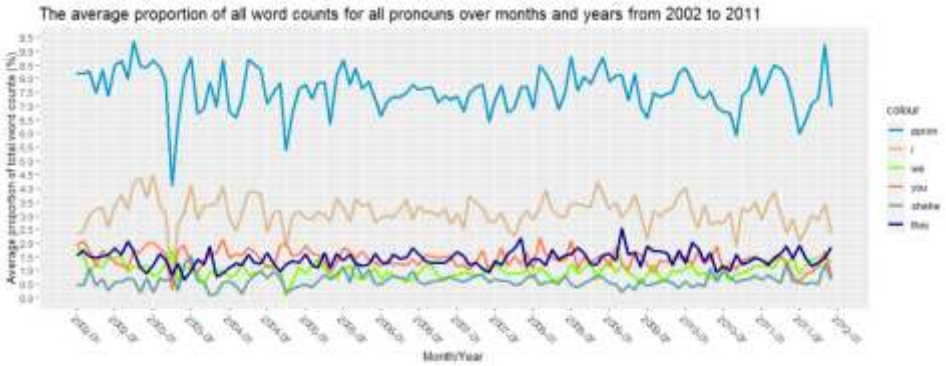

Figure 7. shows the average proportion of all word counts for all pronouns for the posts over months and years from 2002 to 2011

The figure 8 illustrates the average proportion of total word counts for all pronouns for the posts over months and years from 2002 to 2011. From figure 8, we can see that the average proportion for ppron is often between 6.0% to 9.0% throughout the period and its value is the highest among all other pronouns. This is reasonable as ppron is the sum of all other pronouns shown in the graph. The second highest will be the proportion for i variable, its value is often between 2.0% to 4.0% throughout the whole period, except that there was a significant drop in April 2003, where the value is roughly 1.0%. It is interesting to note that for the rest of the average proportion of pronouns (i.e., we, you, shehe and they), their line is very close to each other, hence we can see that their percentage is in the middle of 0% to 2.0% over the long term. Hence, in conclusion, the level of the pronoun variables does fluctuate over the duration of the forum, however, the value is often still within a specific range.

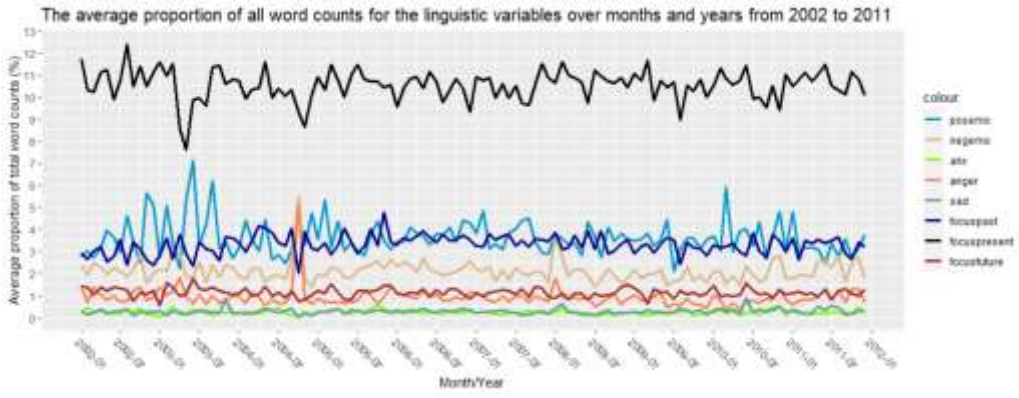

Figure 8.shows the average proportion of all words for the linguistic variables for the posts over months and years from 2002 to 2011

The figure 9 demonstrates the average proportion of total word counts for a specific linguistic variable for the posts over months and years from 2002 to 2011. We can clearly see that

most of the posts are focusing on the present and as a result, the average proportion for focuspresent is frequently between 9% to 12%, and hence it is the highest among the eight variables shown in the graph, followed by posemo and focuspast. The line of posemo and focuspast are adjacent to each other, their average proportions are around 2.5% to 5% throughout the whole period. Surprisingly, the line for negemo, anger, focusfuture, anx and sad tend to be close to each other, hence their average proportion is often between 0% to 2.5% approximately over the long term. Thus, the level of all the linguistic variables shown in figure 9 does vary throughout the period as the line graphs are not smooth at all, but their value is usually still within a particular range (i.e., still considered as stable).

Is there a relationship between linguistic variables over the longer term?

To investigate the relationship between linguistic variables over the longer term, I had decided to create a heatmap to find the correlation between each pair of linguistic variables (except word count with the same reason stated above). To create the heatmap I would need to pre-process the previous data frame. I will use the cbind function to combine all the columns from the 3 data frames that were used to create figure 7, 8 and 9 into a single data frame. After that, I will use the melt function to reshape the data frame and eventually use that data frame to produce the heatmap.

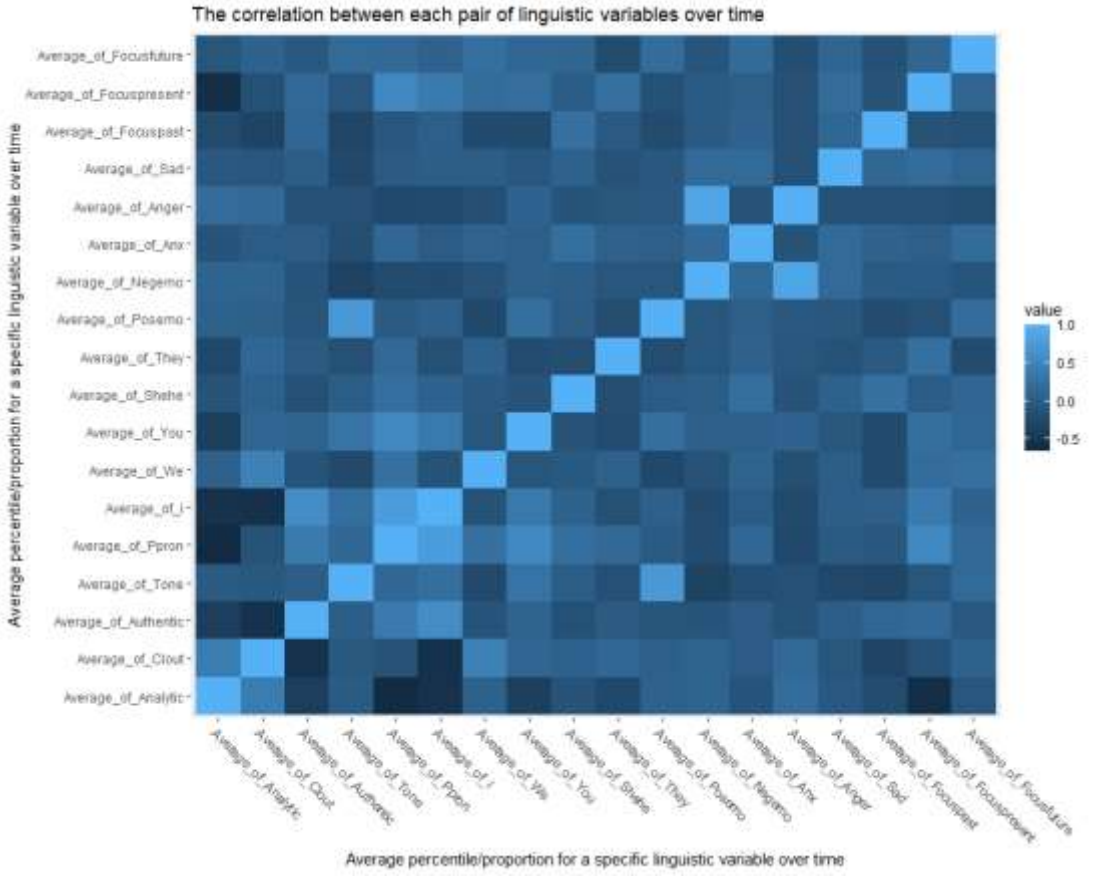

Figure 9. shows the correlation between each pair of linguistic variables over time

The figure 10 shows the heatmap produced by the pre-processed data frame (i.e., the data frame named “correlation_all”). In this figure, we can see that most of the pairs of linguistic variables seem to have no high correlation with each other but there are some exceptions. For instance, the correlation between Average_of_Anger and Average_of_Negemo is around 0.86, the correlation between Average_of_Ppron and Average_of_i is about 0.78 and the correlation between Average_of_Analytic and Average_of_Ppron is approximately -0.65.

Please note that these correlation values are obtained by printing out the “correlation_all” data frame. Nevertheless, it is obvious to know that there might be a high positive correlation between Average_of_Anger and Average_Negemo. (It is true that this pair of variables is not the only pair with high positive/negative correlation, however due to the page limit, I will not be able to plot a scatter plot and do a regression model for each pair) Therefore, I will create a scatter plot with Average_of_Negemo as x-axis and Average_of_Anger as y-axis. Besides that, I also included the Average_of_Tone as this variable provides additional information for us and to figure out if this variable is influenced by the two variables. In addition, I also include a regression model to prove that there is indeed a positive correlation between the two linguistic variables. The red line in figure 11 is the regression line.

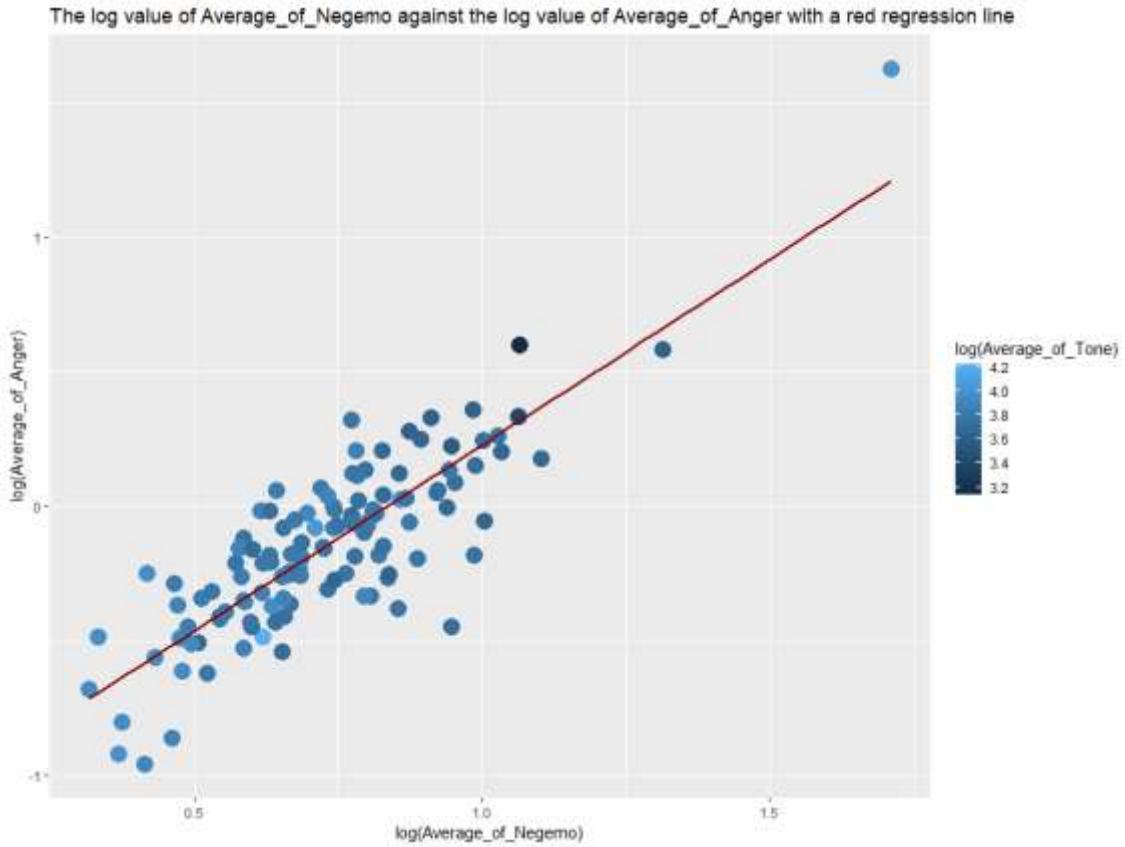

Figure 10. shows the log value of Average_of_Negemo against the log value of Average_of_Anger with a red regression line

From figure 11, we can see that the log value of Average_of_Negemo and Average_of_Anger does produce a linear model. Initially, I plotted the scatter plot without using the log value for both variables, however, I realized that it does not really produce a linear model, therefore I had decided to log them. Same goes with the Average_of_Tone, the colour difference is not obvious without using the log value of it. Therefore, from the figure 11, it is obvious to say that log (Average_of_Anger) and log (Average_of_Negemo) do have positive correlation to each other as the slope is positive. Moreover, we can utilise the regression line to discover more about their relationship by looking at the regression model. By looking at the figure 12, we can see the details of the regression model. Hence, we can express the relationship between log (Average_of_Anger) and log (Average_of_Negemo) with the following, E [log (Average_of_Anger)] = -1.15106 + 1.37883*log (Average_of_Negemo). Thus, it is indeed a positive correlation between them. Additionally, by visualizing the **colour** of the points, we can spot that as log (Average_of_Anger) and log (Average_of_Negemo) increases, log (Average_of_Tone) also tends to increase as the colour of the points get darker (except the outlier).

```
> fitted = lm(log(webforum_all_linguistic_variables$Average_of_Anger)~log(webforum_all_linguistic_variables$Average_of_Negemo))
> summary(fitted)

Call:
lm(formula = log(webforum_all_linguistic_variables$Average_of_Anger) ~
    log(webforum_all_linguistic_variables$Average_of_Negemo))

Residuals:
     Min       1Q   Median       3Q      Max
-0.60151 -0.10749 -0.00296  0.10728  0.41739

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      -1.15106    0.06190  -18.59   <2e-16 ***
log(webforum_all_linguistic_variables$Average_of_Negemo)  1.37883    0.08176   16.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1789 on 118 degrees of freedom
Multiple R-squared:  0.7068,    Adjusted R-squared:  0.7043
F-statistic: 284.4 on 1 and 118 DF,  p-value: < 2.2e-16
```

Figure 11. shows the details of the regression model between log (Average_of_Anger) and log (Average_of_Negemo)

In conclusion, we can say that in the long term, there is a positive correlation between Average_of_Negemo and Average_of_Anger, as well as those pairs shown in the heatmap.

Question b.1)

Using the relevant linguistic variables, is it possible to see whether threads are happier or more optimistic than other threads, or the forum in general, at different periods in time.

For this question, since there are too many threads, I had decided to only focus on the **top ten** threads that had the most posts. Therefore, the first thing I did is to group by the data using ThreadID and find the number of occurrences for each thread using summarize (frequency=n ()). After that, I will be able to get the ThreadID for the top ten threads that have the most posts by sorting the result data frame and using the head () function.

After getting the ThreadID that we want to analyse, now is the time to figure out what linguistic variable should be considered in this question. I had read the reference in the assignment specification, and I realized that the sad variable is included and part of the negemo variable. However, it does not mention whether anx and anger is included and part of negemo, but since based on the definition of these variables, I can assume that the value for anx and anger is already included in negemo. Therefore, for this question, to determine whether threads are happier or more optimistic than other threads, I will only consider the average of the posemo and negemo variables over time for each thread as these two variables indicating the level of positive emotions and negative emotions of each post in a thread and hence allow me to identify how happy a thread is.

I realized that if I include the average of posemo and negemo for the top ten threads in a single image it would be overwhelming. Therefore, I had decided to have two images, the first image consists of the graphs for the top four threads, whereas the second image consists of the graph for the top 5 to 10 threads. Hence, I will filter the data in webforum data frame accordingly and use the group by function to find the average of posemo and negemo for each thread in each month of a specific year (i.e., group by using ThreadID and Date). The figure 13 shows how the data frame that will be used to plot the graph looks like. This data frame consists of the average posemo and negemo for the four threads over time. Please note that I will do the same procedure for the top 5 to 10 threads.

```
> top_4_thread
# A tibble: 87 x 4
# Groups:   ThreadID [4]
   ThreadID Date       Average_Posemo Average_Negemo
      <int> <date>              <dbl>          <dbl>
 1   127115 2004-04-01           0              5.98
 2   127115 2004-05-01           2.54           2.54
 3   127115 2004-06-01           1.54           3.08
 4   127115 2004-08-01           3.23           0
 5   127115 2004-09-01           0.21           1.68
 6   127115 2005-01-01           0.57           1.14
 7   127115 2006-05-01           1.78           2.78
 8   127115 2006-08-01           2.31           1.91
 9   127115 2006-09-01           2.24           2.62
10   127115 2007-01-01           2.42           0.82
# ... with 77 more rows
```

Figure 12. shows the top few rows of the result data frame after group by the data using ThreadID and Date to get the average posemo and negemo for each month of a year for each thread

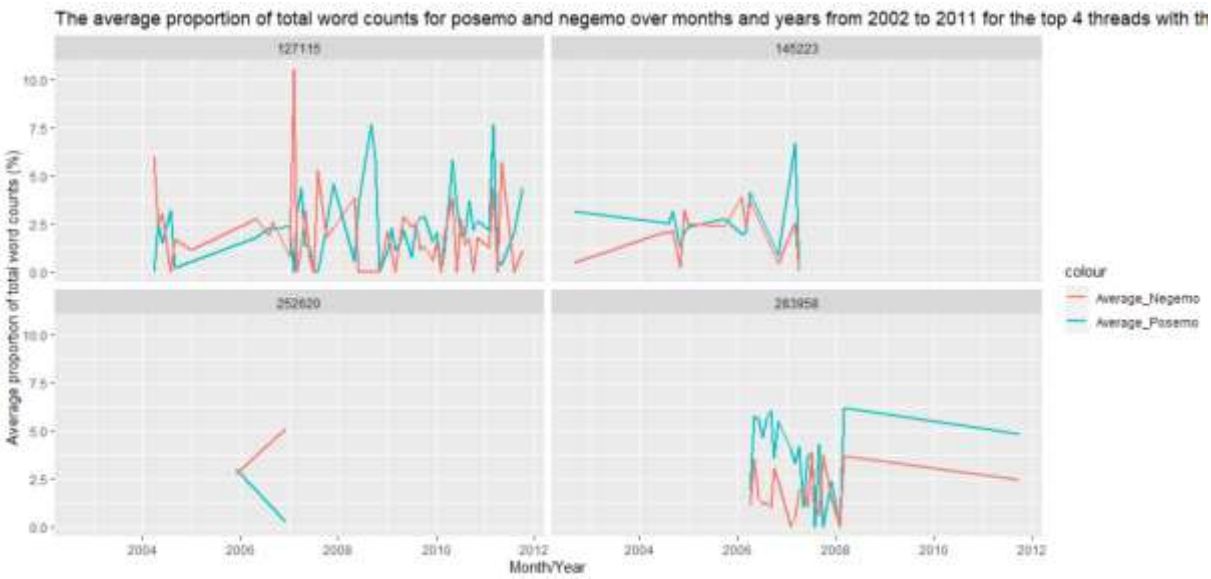

Figure 13. shows the average proportion of total word counts for posemo and negemo over months and years from 2002 to 2011 for the top four threads with the most posts

In figure 14, we can observe the average proportion of posemo and negemo throughout the period for the top four threads with the most posts. For the thread with 127115 as ID, we notice that the two linguistic variables begin to be active from around 2004 until 2012. Throughout this period, it is tough to judge the distinction between the values for the two variables for this thread as the values are quite close to each other. Regardless, we can say that most of the time the average proportion of negemo and posemo are between 0% to 5.5%. Hence, I will not consider this thread to be optimistic although the average posemo is relatively high as the average negemo is also identical. Currently, we will be looking at the second thread with the ID 145223. This thread is surprisingly like the previous thread. Its average posemo and average negemo are fairly like each other as illustrated in the image, besides that the average posemo was slightly higher than the average negemo before 2005. For the thread with ID 252620, throughout the entire period, the average negemo appears to be higher than the average negemo. Consequently, it is evident to say that this thread is not optimistic at all. Finally, the last thread among the top four threads will be the one with ID 283958. For this thread, excluding the year 2007, its average posemo is higher than the average negemo by approximately 2.5%. Accordingly, it seems like the last thread (i.e., 283958) is the most optimistic and happiest thread among the top four.

The reason for me to identify which thread is the most optimistic among these four threads is because afterward I will compare this thread with the thread that is most optimistic among the top 5 to 10 threads and do a hypothesis test regarding it to find out which one is the most optimistic and happiest thread among the top ten threads. Besides that, **please note that I will only compare a pair of threads in this question**, the reason is that due to the page limit, I will not be able to do a hypothesis test for every single pair of threads to identify the happiest thread. Hence, I decided to pick this pair of threads (i.e., the one that is the most optimistic and happiest thread among the top four threads and another one that is the most optimistic and happiest thread among the top 5 to 10 threads).

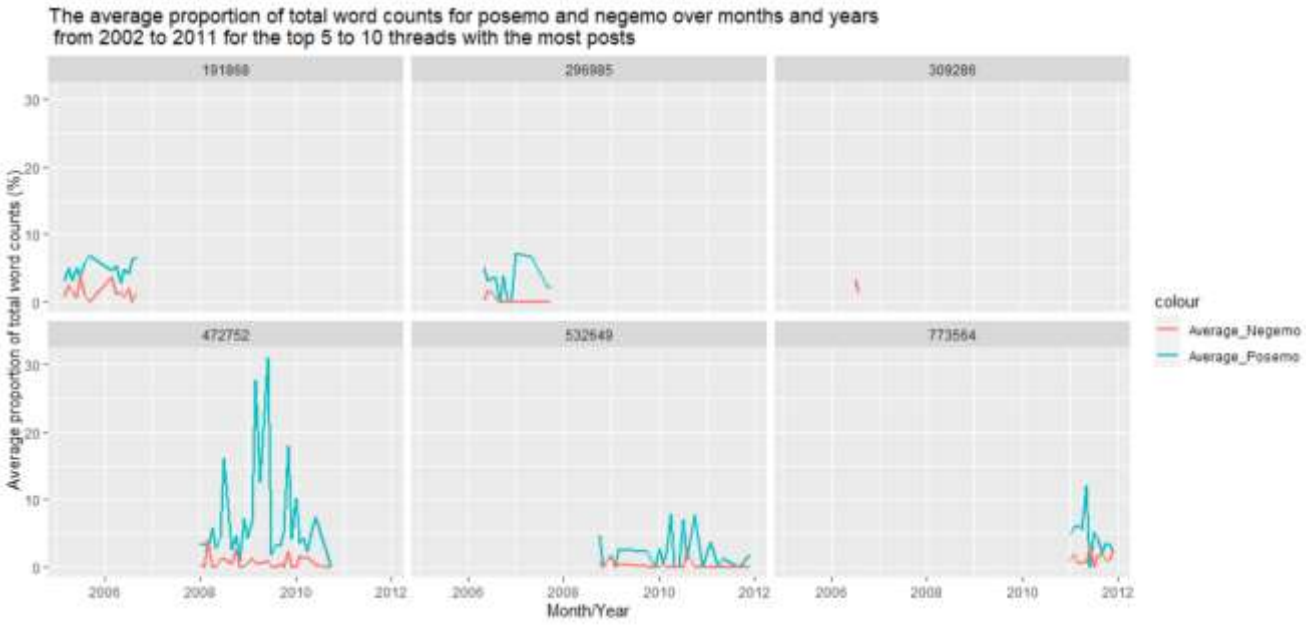

Figure 14. shows the average proportion of total word counts for posemo and negemo over months and years from 2002 to 2011 for the top 5 to 10 threads with the most posts

Figure 15 shows the average proportion of total word counts for posemo and negemo over months and years from 2002 to 2011 for the top 5 to 10 threads with the most posts. Thus, we can determine the most optimistic thread among these six threads by using this figure. Based on this figure, it is interesting to note that other than the thread with ID 309286, the rest of the five threads seem to have a higher value for an average of posemo compared to the average of negemo throughout the entire period. For the thread with ID 191868, 296985, 532649, and 773564, we can see that its average of posemo is slightly higher than its average of negemo within a range of 0% to 10% most of the time. Nonetheless, this doesn't seem to happen for the thread with ID 472752. From the figure, we notice that its average of posemo is considerably higher than other threads over time and relatively higher than its average of negemo. Its average of posemo even reached more than 30% during 2009. Thus, in summary, it is evident to say that the thread with ID 472752 is the most optimistic and happiest thread among these six threads according to the details displayed in the figure.

After getting the most optimistic and happiest thread among the top four threads (i.e., the thread with ID 283958) and among the top 5 to 10 threads (i.e., the thread with ID 472752), I will now compare these two threads to find out which one is more optimistic than another. Based on my current observation, I feel like the thread with ID 472752 is more optimistic and happier than the thread with ID 283958. However, I will need to have evidence to support my finding. To do so, I will do two hypothesis testing, the first one is to compare the average of posemo over time and the second one will be comparing the average of negemo over time.

For the first hypothesis testing, I will be required to obtain the average posemo for each month of a year for both threads by filtering the data frame that I used to plot the graph and I will use the t.test function. In this case, the null hypothesis is that the average of posemo for the thread with ID 472752 is greater than the average of posemo for the thread with ID 283958 throughout the whole period. The confidence level is 95% as default. The result of this hypothesis testing is provided in figure 16. According to the result displayed in figure 16, we can see that the p-value is 0.9889. It means that if the null was true, then the chance of observing a sample with an extreme or more extreme difference from the null as the one we observed is about 98.9%. There is insufficient/weak evidence against the null hypothesis that the average of posemo for the thread with ID 472752 is greater than the average of posemo for the thread with ID 283958 throughout the whole period since the p-value > 0.05. Therefore, we cannot reject the null hypothesis. Hence, from here, we know that the thread with ID 472752 is happier than the thread with ID 283958.

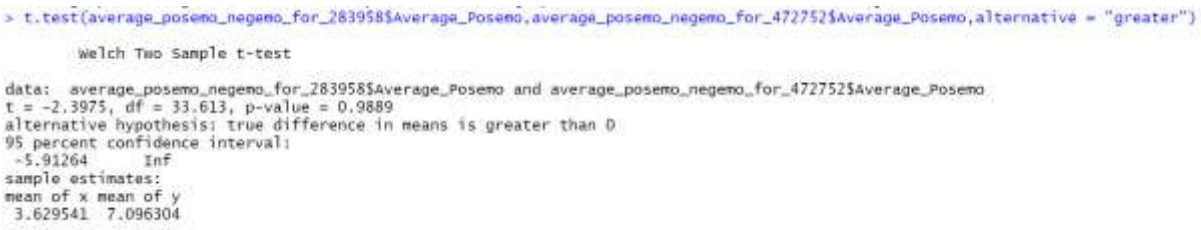

Figure 15. shows that result of the hypothesis testing using t.test in R

Likewise, for the second hypothesis testing, I will be required to obtain the average negemo for each month of a year for both threads by filtering the data frame that I used to plot the graph and I will use the t.test function. The null hypothesis is that the average of negemo for the thread with ID 283958 is greater than the average of posemo for the thread with ID 472752 throughout the whole period. The confidence level is 95% as default. The result of this hypothesis testing is illustrated in figure 17. The result shows that the p-value is 0.9972. It means that if the null was true, then the chance of observing a sample with an extreme or more extreme difference from the null as the one we observed is about 99.7%. There is insufficient/weak evidence against the null hypothesis that the average of negemo for the thread with ID 283958 is greater than the average of posemo for the thread with ID 472752 throughout the whole period since the p-value > 0.05. Therefore, we cannot reject the null hypothesis. Hence, from here, we know that the thread with ID 283958 is more pessimistic than the thread with ID 472752.

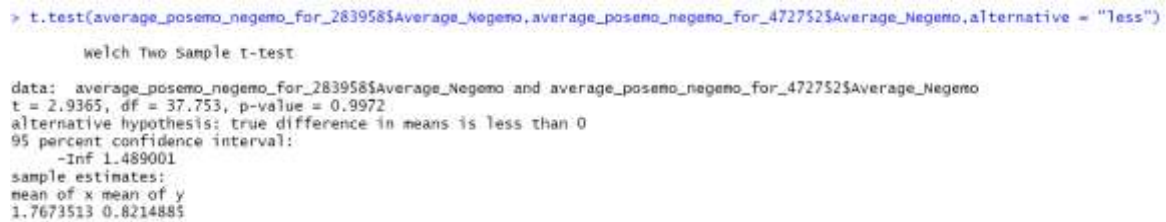

Figure 16. shows that result of the hypothesis testing using t.test in R

Using the result of both hypothesis testing above, we know that the thread with ID 472752 is happier, more optimistic, and less pessimistic than the thread with ID 283958, thus the thread with ID 472752 is the most optimistic thread among the top ten threads.

Question c.1)

Create a non-trivial social network of all authors who are posting over a particular time. For example, over one month. To create this, your social network should include at least 30 authors, some of whom will have posted to multiple (2 or more) threads during this period. Your social network should be connected, although some authors may be disconnected from the main group. Present your result as a network graph.

Before creating the social network, I will need to explore the number of posts for each month of a year. As a result, I will be able to select a particular month such that the number of posts for that month is not too many (i.e., avoid the result network graph to be over occupied) or less (i.e., at least 30 authors had posted over this month). To do this, I will group by the webforum data set by using the Date column and using summarize function to count the number of rows for each group.

Eventually, I decided to build a social network of all authors who are posting over July of 2009. The reason why I choose to build the network graph according to this specific time frame is because it fulfils the precondition of this question (i.e., at least 30 authors) and the number of vertices (i.e., total number of unique AuthorID) and edges are not too excessive, hence the result network graph will not be too crowded. After deciding the time frame, I will create a new data frame and filter the webforum data set such that the new data frame only contains the rows that are posted within this time frame. Now, I will be plotting the social network graph based on this new data frame.

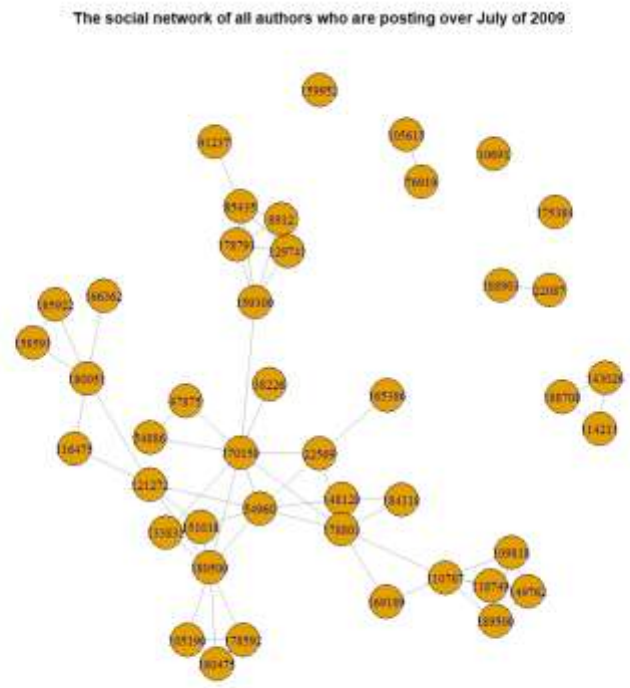

Figure 17. shows the social network of all authors who are posting over July of 2009

The result network graph is shown in figure 18. The vertices represent the AuthorID for each author and the edges between two vertices represent that these two vertices posted to the same thread during this time frame (i.e., July of 2009). There are a total of 67 edges and 44 vertices in this network graph. Thus, it means that there are 44 unique authors included.

Please note that this is an unweighted graph, which means although two vertices may post to the same thread with different ThreadID for multiple times (e.g., Author A and Author B posted to Thread C and Thread D during July of 2009, hence they posted to the same thread twice), however we did not take this into consideration as we are plotting an unweighted graph.

From figure 18 shown above, we can see that most vertices are connected, but despite that, around 10 vertices are disconnected from the main group. In addition, after obtaining the result network graph, we can analyse the network in an overall perspective by finding out its diameter and average path length, graph density, transitivity, and the degree distribution. According to the information shown in figure 19, it shows that the diameter of the network graph is 7, the average path length is around 3.28, the graph density is approximately 0.07 and its transitivity is about 0.52. In addition, we can plot a histogram to visualise the distribution of the degree for the vertices in the network graph. The result histogram is shown in figure 20. We can see that most of the vertices (i.e., around 22 vertices) have a degree of 0 to 2, and around 12 vertices have a degree of 2 to 4, and only a minority of the vertices have a degree between 4 to 10.

```
> diameter(g)
[1] 7
> average.path.length(g)
[1] 3.275618
> graph.density(g)
[1] 0.07082452
> transitivity(g)
[1] 0.5234043
> # get.adjacency(g) # the adjacency matrix is too big to show in the console
> hist(degree(g),breaks=5,col = "grey")
>
```

Figure 18. shows all the factors regarding the network graph shown in figure 18

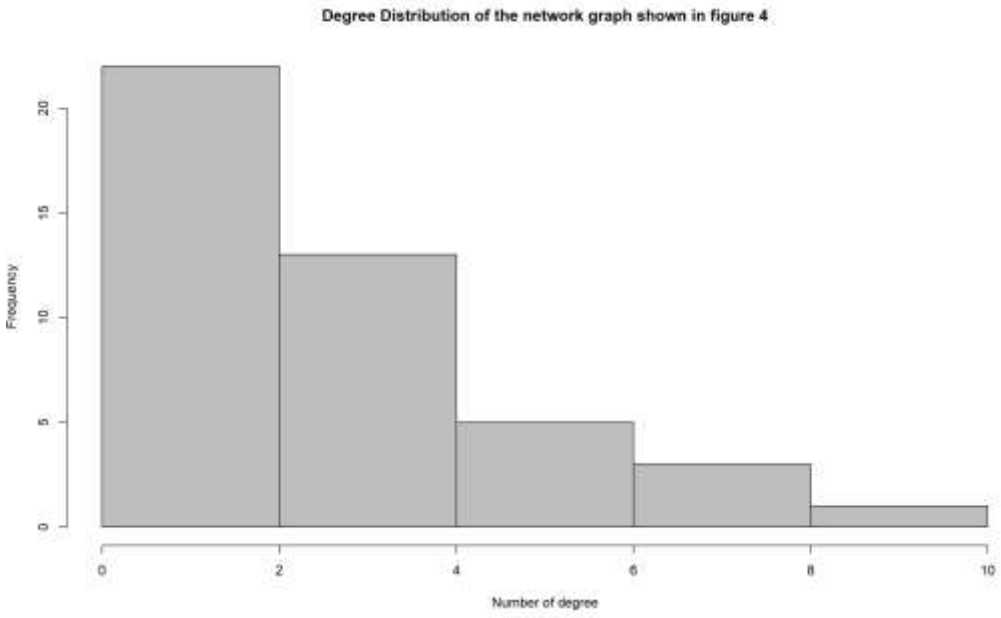

Figure 19. shows the degree distribution of the network graph shown in figure 18

Question c.2)

Identify the most important author in the social network you created.

To identify the most important author in the social network that I created in the previous question, firstly, I will need to calculate the network centrality measures (i.e., degree, betweenness, closeness and eigenvector) for each vertex. The figure 21 shown illustrates the top few rows of the data frame that used to store the network centrality measures for each vertex.

```
> head(stats)
       degree betweenness   closeness eigenvector
149762      4    0.000000 0.007812500   0.2150063
129741      4    0.000000 0.008333333   0.2115239
148128      3    5.333333 0.010526316   0.4267857
170150      9  266.500000 0.014285714   0.9989390
175384      0    0.000000         NaN   0.0000000
105615      1    0.000000 1.000000000   0.0000000
```

Figure 20. shows the top six lines of the data frame that stores the network centrality measures for each vertex

After generating the useful data frame, I can now sort the data frame according to each network centrality measure. After sorting the data frame and studying the result of it, I realized that the most important author in the social network is the author with ID 170150. The reason is because 170150 ranked top 2 for most of the measures (i.e., ranked first in betweenness and degree, ranked second in eigenvector but ranked eighth in closeness), which is better than all other authors in overall. Therefore, the author with ID 170150 is the most important author in the social network that I created in the previous question.

Looking at the language they use, can you observe any difference between them and other members of their social network?

After identifying the most important author, I will need to compare the language used by this author with other authors that posted during this time frame. Essentially, I will first find the average of all linguistic variables for the posts that were posted by the author with ID 170150 during July 2009.

Linguistic variable	Analytic	Clout	Authentic	Tone	ppron	i	we	you	shehe
Average value (%)	70.75	65.9	27.3	37.7	5.00	0.99	0.53	1.34	0.00
Linguistic variable	they	posemo	negemo	anx	anger	sad	focus past	focus present	focus future
Average value (%)	2.13	2.14	1.45	0.17	0.44	0.17	1.83	11.51	0.78

Above table shows the average percentage/percentile of all linguistic variables based on the posts that were posted by the author 170150 during July of 2009. All the values are generated by using R code. Please note that I did not include WC (word count) in this question. The reason is because I do not think WC will give any meaningful details regarding the difference of language between authors as WC only indicates the number of words. Besides that, I am planning to plot histograms to visualise the distribution of the value for all

the linguistic variables for all the authors that are included in the social network but excluding the author 170150. Therefore, another reason to not include WC is because including WC variable will increase the range of x-axis and cause my later graphs (i.e., figure 22 in the following page) to be more difficult to visualize as the bar are extremely small, hence due to these two reasons, I had decided to remove WC column when analysing the difference of language between author 170150 with other authors.

To compare the language used by author 170150 with other authors that posted during this time frame, I will plot a histogram for each linguistic variable (except WC) to show the distribution of the value for all authors (excluding author 170150).

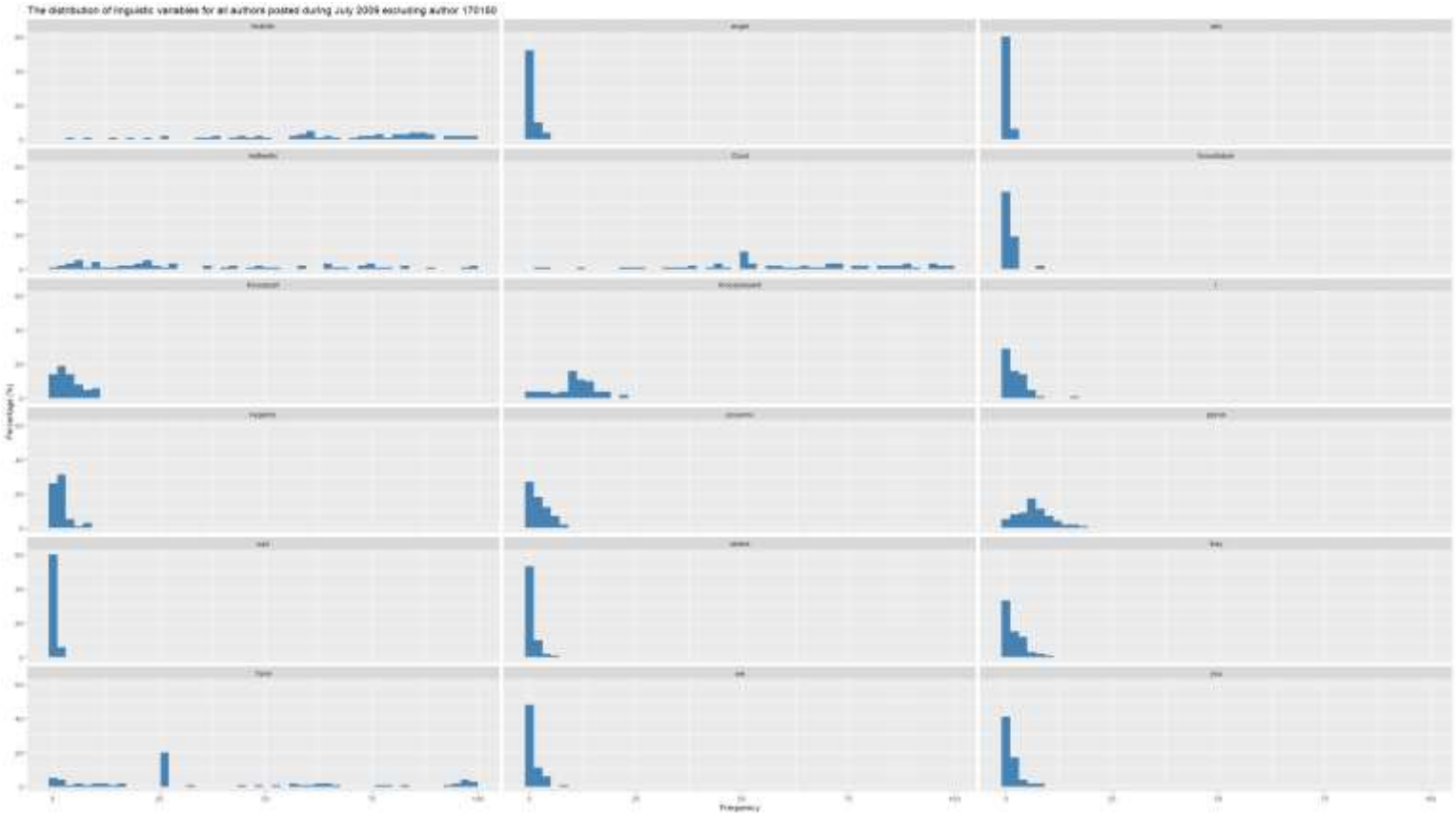

Figure 21. shows the distribution of linguistic variables for all authors posted during July 2009 excluding author 170150

The figure 22 shows the distribution of linguistic variables for all authors posted during July 2009 but excluding author 170150. After obtaining the histograms shown in figure 22, we can now compare the language they use. From the figure above, we can see that there is not any pattern for the distribution of analytic, authentic, tone and clout, therefore we can't really tell the differences between author 170150 and other authors regarding these variables. Moreover, for anger, anx, focusfuture, focuspast, i, negemo, posemo, sad, shehe, they, we and you, the range of these variables for other authors are all/almost within the range of 0% to 12.5%. Despite that, interestingly, the value of these linguistic variables for author 170150 also within this range, hence I cannot really tell the difference of these variables between them. Furthermore, the range of focuspresent and ppron for other author seems to be larger than the previous variables, the value ranges between 0% to 25%. Surprisingly, the average of focuspresent and ppron for author 170150 also within this range. Hence, I am interested to figure out whether there is a difference in mean between author 170150 and other author for focuspresent and ppron variables. Therefore, I will be doing two hypothesis tests to support my finding. Please note that due to page limit, I am not able to do

a hypothesis test for each linguistic variable. Thus, I had selected only two, which are focuspresent and ppron to do the hypothesis testing.

For the first hypothesis testing, I will be required to obtain the value of focuspresent for author 170150 and other authors by filtering the “data_long” data frame (i.e., the data frame that used to plot figure 22) and “post_in_July_of_2009_all_variables” data frame. After that, I will use the t.test function such that the null hypothesis is that the mean of focuspresent does not differ between author 170150 and other authors during July of 2009. The confidence level is 95% as default. The result of this hypothesis testing is provided in figure. Based on the result displayed in figure 23, we can see that the p-value is 0.7064. It means that if the null was true, then the chance of observing a sample with an extreme or more extreme difference from the null as the one we observed is about 70.6%. There is insufficient/weak evidence against the null hypothesis that the mean of focuspresent does not differ between author 170150 and other authors during July of 2009 since the p-value > 0.05. Thus, we cannot reject the null hypothesis.

```
> data_for_focus_present_other_author = data_long[(data_long$name=="focuspresent"),2]
> data_for_focus_present_other_author
 [1]  9.52 10.53 14.67  9.77  7.14 13.22 14.10  6.21 12.50 10.00 10.32  0.82  8.33 13.33 11.59  2.98  3.01 12.64 18.92 13.64 10.71 16.92
[23]  2.73  0.00 12.00 16.67 10.84  2.86 11.29  4.48  5.39  9.38 11.11  8.00  0.00 10.89 10.53 12.50 15.83 13.89  8.70 12.42  3.92 13.51
[45] 15.49 13.79  4.26 22.22  0.00  6.93 12.84 11.76 18.18 14.29  9.30 12.06 15.00 18.52 10.08 10.26 21.62 12.63  5.00 17.58 10.39  9.43
> data_for_focus_present_author_170150 = post_in_July_of_2009_all_variables[(post_in_July_of_2009_all_variables$AuthorID == 170150),22]
> data_for_focus_present_author_170150
[1]  4.76 14.29 12.68  9.26 21.74 13.11  4.76
> t.test(data_for_focus_present_other_author,data_for_focus_present_author_170150,alternative="two.sided")

        Welch Two Sample t-test

data:  data_for_focus_present_other_author and data_for_focus_present_author_170150
t = -0.39258, df = 6.9651, p-value = 0.7064
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.443977  4.610557
sample estimates:
mean of x mean of y
 10.59758  11.51429
```

Figure 22. shows that result of the hypothesis testing using t.test in R

Similarly, for the second hypothesis testing, I will be required to obtain the value of ppron for author 170150 and other authors by filtering the “data_long” data frame (i.e., the data frame that used to plot figure 22) and “post_in_July_of_2009_all_variables” data frame. Next, I will be using the t.test function with a null hypothesis that the mean of ppron does not differ between author 170150 and other authors during July of 2009. The confidence level is 95% as default. The result of this hypothesis testing is provided in figure 24. Based on the result displayed in figure 24, we can see that the p-value is 0.2652. It means that if the null was true, then the chance of observing a sample with an extreme or more extreme difference from the null as the one we observed is about 26.5%. There is insufficient/weak evidence against the null hypothesis that the mean of ppron does not differ between author 170150 and other authors during July of 2009 since the p-value > 0.05. Thus, we cannot reject the null hypothesis.

```
> data_for_ppron_other_author = data_long[(data_long$name=="ppron"),2]
> data_for_ppron_other_author
 [1]  7.94 13.45 10.00  1.50  7.14 12.40  3.85 10.34  6.25  5.56  5.56  1.64 16.67  0.00  8.70  5.36  1.50  6.59  5.41  9.09  0.00 10.77
[23]  2.73  0.00  6.00  8.33  4.82  0.71  6.45  7.46  7.78  4.69  9.26  4.00  1.37  4.95 15.79  7.50  5.00 10.19  5.80  7.45  1.96  5.41
[45]  7.04  8.05  4.96  0.00  5.88  2.97  6.42 11.76 18.18 11.69 13.95 11.28 10.00 11.11  3.36  5.98  2.70  5.26  5.00  6.59  5.19  7.55
> data_for_ppron_author_170150 = post_in_July_of_2009_all_variables[(post_in_July_of_2009_all_variables$AuthorID == 170150),10]
> data_for_ppron_author_170150
[1] 0.00 2.86 5.63 5.56 8.70 9.84 2.38
> t.test(data_for_ppron_other_author,data_for_ppron_author_170150,alternative="two.sided")

        Welch Two Sample t-test

data:  data_for_ppron_other_author and data_for_ppron_author_170150
t = 1.2002, df = 7.8118, p-value = 0.2652
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.585280  4.996579
sample estimates:
mean of x mean of y
 6.701364  4.995714
```

Figure 23. shows that result of the hypothesis testing using t.test in R

Besides that, for both results of the hypothesis test, we also can see that the 95% confidence interval also includes zero, hence we cannot rule out the possibility of there being no difference at a population level between author 170150 and other authors.

Overall, in conclusion, based on the result of both hypothesis testing above and the insight gained from figure 22, I can conclude that the language used by the most important author (i.e., author 170150) does not differ with other authors in the social network.