FIT3152 Data analytics: Assignment 1

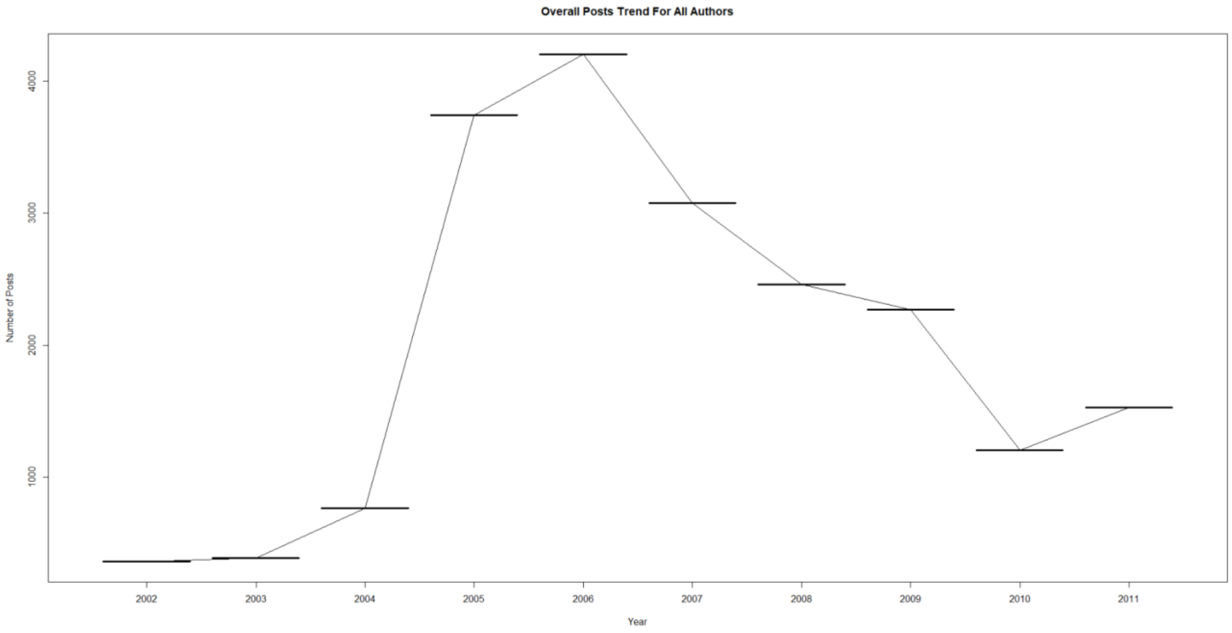
Introduction:

The data consists of the linguistic analysis of posts over the period of years 2002 to 2011. The analysis is processed on 20,000 sample posts from this time period which is the randomly distributed subset of the given data. The linguistic analysis was conducted using Linguistic Inquiry and Word Count (LIWC), which assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication.

The process of analysis of the sampled data is divided into three parts. First, the data is processed into relevant form for the linguistic variables to find if any of them have trend over the specific time period. That would help us in discovering how active participants(authors) are. It would lead us to find relationship (if there is any) between any of the linguistic variables. The second part of the process consists of analysis on the thread section. By dividing threads into groups with other variables, it would give out the distinguish of languages used by different groups. Therefore, it would indicate how each thread are distributed all over the other variables in the data. The final part of the analysis process includes the social network of the participants. In the social media platform, the participants could communicate on the same thread at the same time and it forms a social network. It is high possibility that the participants also communicate on other threads which would extend their social network. Hence, there is the social network designed as part of the process to describe how participants form variety of social network over range of threads. The implemented social network will give us an idea of how participants can make great amount of communication and extend their network which would be within very short duration of time.

Task-a:

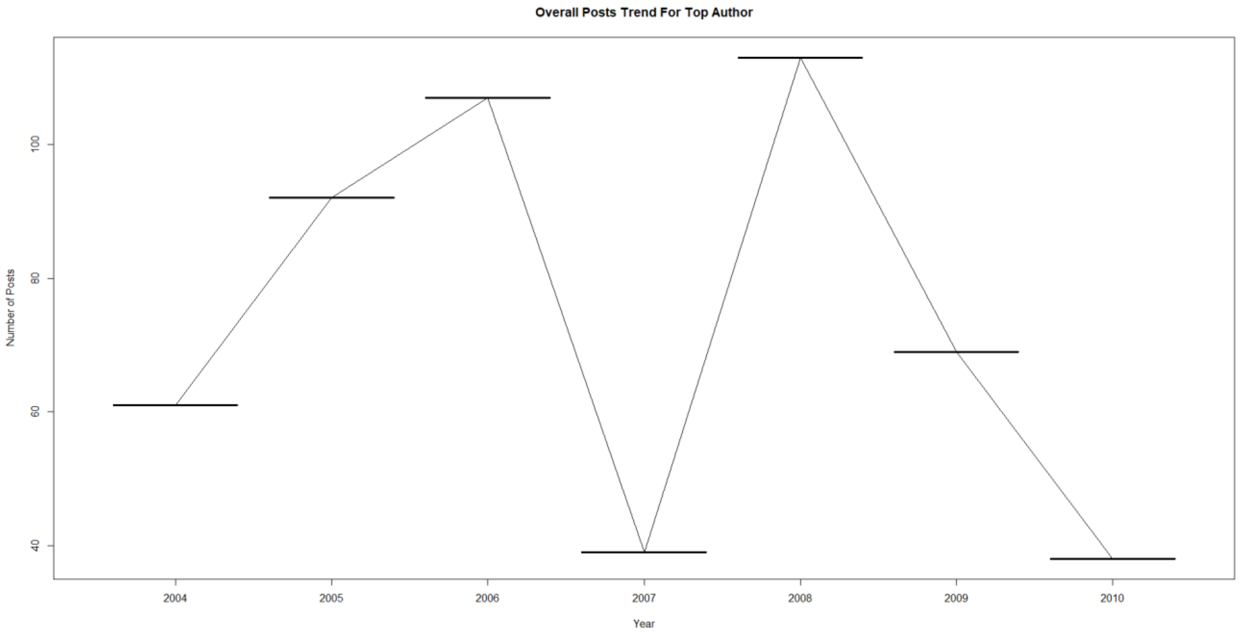
The dataset provided consists of the timeline from year 2002 to year 2011. The participant makes posts frequently throughout the year. However, it is been found that the most overall posts by all author turned out to be in the calendar year of 2006. According to the graph below, the graph was at its peak in 2006. But before 2006, the trend was moving upwards linearly up to 2004 and then it took a huge climb and went up in 2005 by three times of the current value. The trend was on the all-time best value of the graph in 2006 before it started falling rapidly in the following year. It was still falling heading into 2008 and 2009 but with the lower rate as compared to the year 2007. The trend was declining rapidly again into 2010 but it started increasing into 2011 which is the first time the graph started increasing since 2005-06. In terms of the number posts, Participant were most active throughout 2006 on the social media. They were not too involved during 2002 and 2003, However, they started discussing more and more in the following years till 2006. That is when the trend of posting on social media goes upwards very rapidly but the participants have been less active from 2006 to 2010 as the graph decline rapidly and then slowly, and then rapidly heading into 2010. But the trend goes up quietly as participants have decided to be posting online more often in 2011. The average posts per year (population mean) is 2000. Therefore, there is no time trend that affects the average value before 2004 but it will be interesting to see the increase in posts between 2004-2006. The decreasing trend towards 2010 also affects the average value. It will depend upon the linguistic variables as in how they will play part to affect the graph which will come in later part for the further analysis.



There are already plenty of participants as the data above includes the posts from all the authors throughout the ten years. [Appendix-1] However, the most active participant have been throughout the years is 39170 (author_id) and there has been total of 519 posts by the author but the mean of the posts is 8.153 which is close to minimum post number of 1. There have been only 24 out of 2453 authors who have posted more than 100 times in the forum.

Considering the fact that the overall trend for all authors have been going up and down throughout the period, it's believed to be decent spot to see the behavior of any author as in how it's differed from the original overall trend. In other words, we can monitor individual author's behavior considering the overall social media culture that is been changing repeatedly.

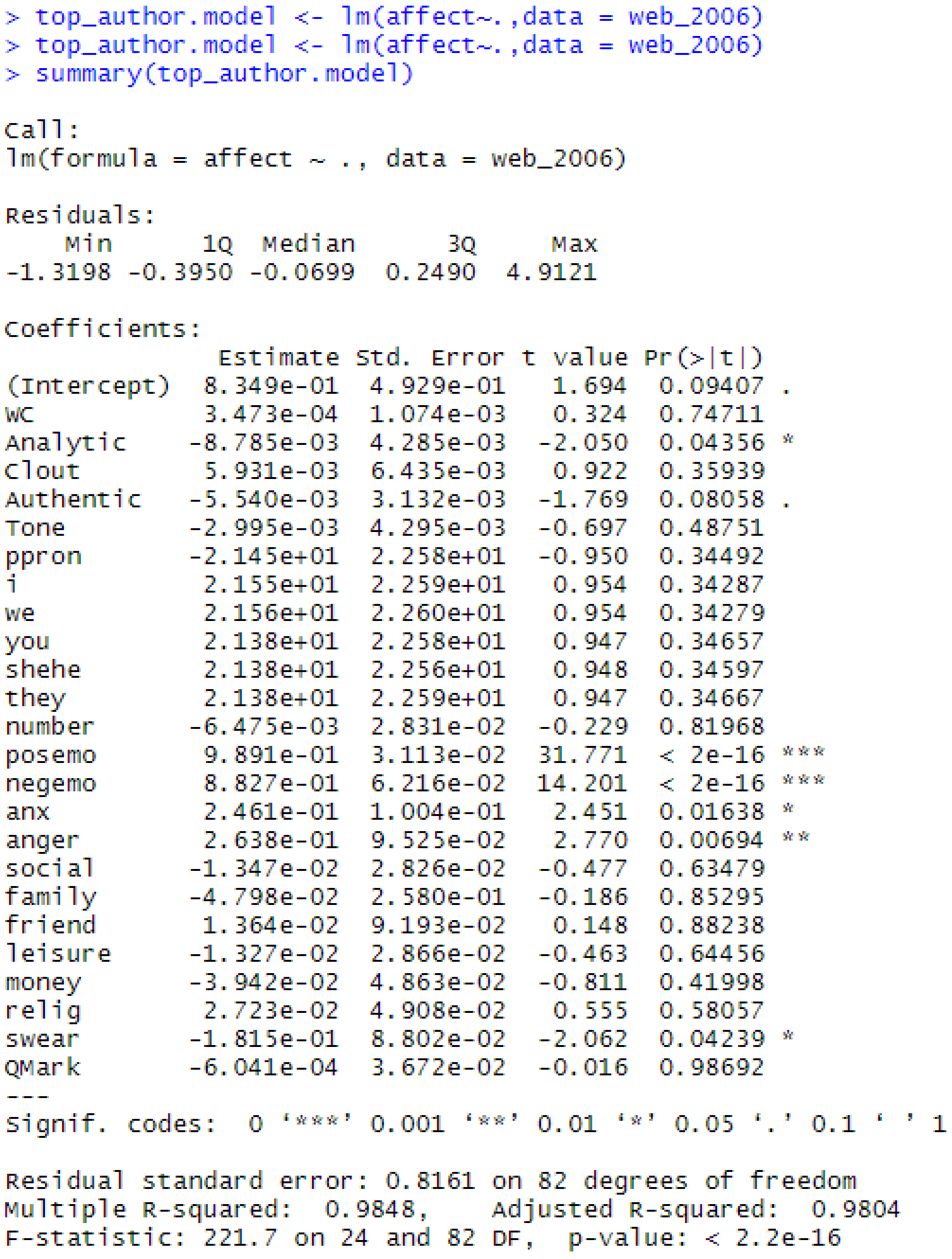
The chosen individual author is the author having the most posts to compare against the overall posts by other authors. As the average frequency per year for the data is 74.14 and it can be seen in the data that the time trends of going up from 2004-2006 makes difference to the mean value. The trend consistently falls in the next year and it consistently climbs up in the year after as the user could possibly be having the allocated time commitment. From 2008 to 2009, the graph gets down towards nearly mean value and it gets down again to below mean value. But overall, it will be interesting insights for the trend of the year (2004-05), (2005-06) and (2006-07).



According to the points made for the overall data for all authors, the trend for both intervals (2004-05) and (2005-06) are both interesting insights which affects the time trend for both graphs. However, they both have an increasing for the same trend and they both make it having the same behavior for all the participants. However, the purpose of this finding is to see the change of behavior in the top participant. If we stick by the overall authors trend, we'll end up with the same activity data as others and the interesting finding will be to find the variables which are affecting the top author's social media status and in the different trend from the others. Therefore, we will investigate interval (2006-07) which satisfies all these criterions to see the insights into data and identify the variables forcing him to behave differently.

As the change of behavior of the individual author falls into sentiment emotion which is the psychological process as the author would do certain things like posting or going online on social media depending on the sentiments meaning the affective emotion process. This process can affect the participant in both positive and negative ways. The effect of the same will be seen in the post they make next. We will use regression analysis to see what variable have significant relationship with affect.

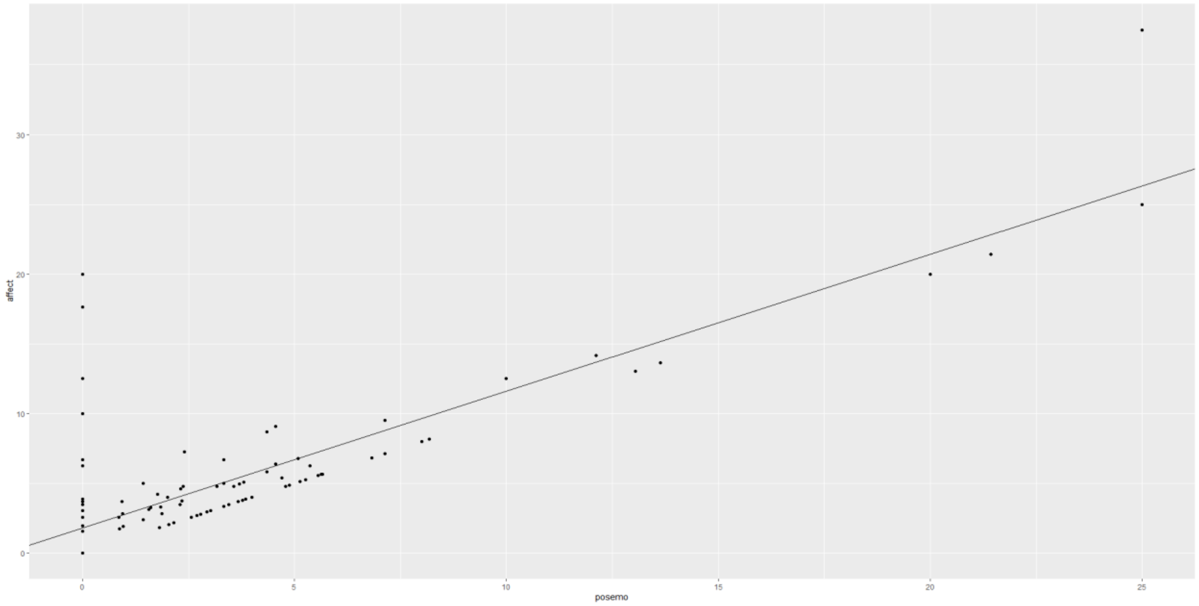
From the table below, we could ignore the variables as we identified that they do not have significant p values. So, from the data below, the significant predictors for the affect variable are 'Analytic', 'Authentic', 'posemo', 'negemo', 'anx', 'anger' and 'swear' where; 'Analytic', 'Authentic', 'anx' and 'swear' are less significant and 'posemo', 'negemo' are the most significant predictors. The rest variables are removed. The Adjusted R-squared value is that 98.04% of the variation in affect can be predicted by the above variable which is a good fit. While the median is close to 0 and the residual standard error is 0.8161 which is pretty good for predicting. Overall P-value getting 2.2*10^-16 means even more significant results for the affect variable. It means it supports the hypothesis that there's a linear relationship possible between affect and the other variables mentioned above.



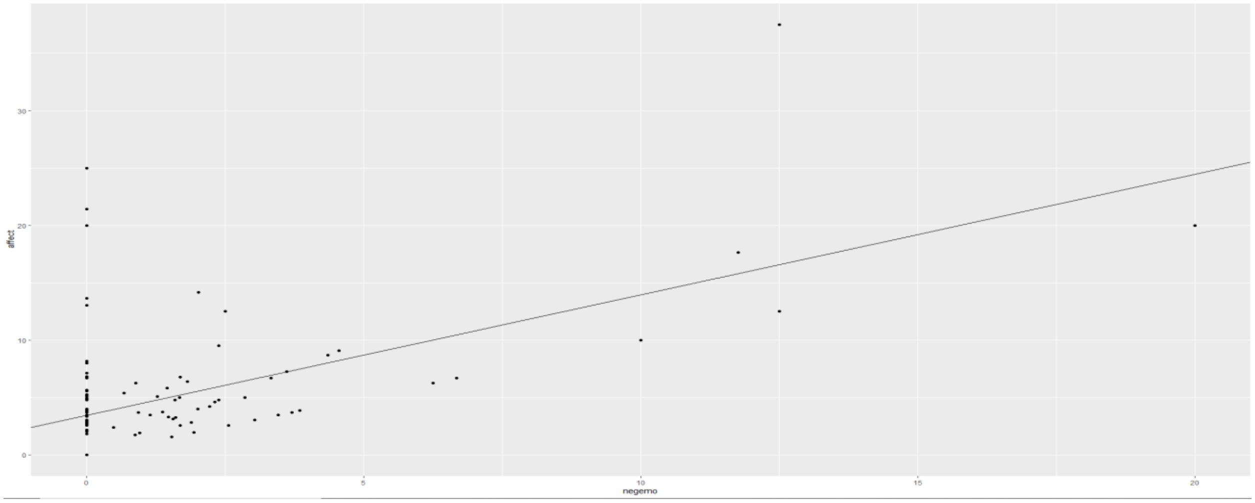
```
> top_author.model <- lm(affect~.,data = web_2006)
> top_author.model <- lm(affect~.,data = web_2006)
> summary(top_author.model)

Call:
lm(formula = affect ~ ., data = web_2006)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3198 -0.3950 -0.0699  0.2490  4.9121 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.349e-01  4.929e-01   1.694  0.09407 .  
WC           3.473e-04  1.074e-03   0.324  0.74711    
Analytic    -8.785e-03  4.285e-03  -2.050  0.04356 *  
Clout        5.931e-03  6.435e-03   0.922  0.35939    
Authentic   -5.540e-03  3.132e-03  -1.769  0.08058 .  
Tone        -2.995e-03  4.295e-03  -0.697  0.48751    
ppron       -2.145e+01  2.258e+01  -0.950  0.34492    
i            2.155e+01  2.259e+01   0.954  0.34287    
we           2.156e+01  2.260e+01   0.954  0.34279    
you          2.138e+01  2.258e+01   0.947  0.34657    
shehe        2.138e+01  2.256e+01   0.948  0.34597    
they         2.138e+01  2.259e+01   0.947  0.34667    
number      -6.475e-03  2.831e-02  -0.229  0.81968    
posemo       9.891e-01  3.113e-02  31.771  < 2e-16 ***
negemo       8.827e-01  6.216e-02  14.201  < 2e-16 ***
anx          2.461e-01  1.004e-01   2.451  0.01638 *  
anger        2.638e-01  9.525e-02   2.770  0.00694 ** 
social      -1.347e-02  2.826e-02  -0.477  0.63479    
family      -4.798e-02  2.580e-01  -0.186  0.85295    
friend       1.364e-02  9.193e-02   0.148  0.88238    
leisure     -1.327e-02  2.866e-02  -0.463  0.64456    
money       -3.942e-02  4.863e-02  -0.811  0.41998    
relig        2.723e-02  4.908e-02   0.555  0.58057    
swear       -1.815e-01  8.802e-02  -2.062  0.04239 *  
QMark       -6.041e-04  3.672e-02  -0.016  0.98692    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8161 on 82 degrees of freedom
Multiple R-squared:  0.9848,	Adjusted R-squared:  0.9804 
F-statistic: 221.7 on 24 and 82 DF,  p-value: < 2.2e-16
```

According to the graphs below, it’s clear that the positive emotion and negative emotion have clear relationship trend with Affect. Affect and positive emotion have a very strong positive correlation of 0.8185 which clearly means that the change in any value of positive emotion will positively change in the affect value. It interprets that the individual top author who think or post positively within his/her posts are likely to also engage in more affectional(sentimental) language. However, in case where author posting negative emotion are less likely to engage in sentimental language because the correlation is 0.5544 which is not as strong as positive emotion does. Still, negative emotion will have decent amount of impact while deciding for the affect variable for the given Author. In the result, there have been trends where the graph is falling and the reason behind that is the overuse of negative emotion which results in the author not being engaging into Affect or sentimental emotion. Not engaging into the posts would also keep author away from social media which makes graph falling in a few months but it is overall good positive. Remember, there is also further but less significant relationships as Authentic, Analytic, Anx, Anger and Swear can make an impact on Affect. They are not strong relation but they still have their part to play as Anger and Swear could potentially lead to negativity and Analytic, Authentic could spread positiveness in most cases so the given author would also be impacted by these variables.

AFFECT VS POSEMO- Correlation: 0.8185



AFFECT VS NEGEMO- Correlation: 0.5544



Task-b:

Analyse the language used by groups. Some starting points:

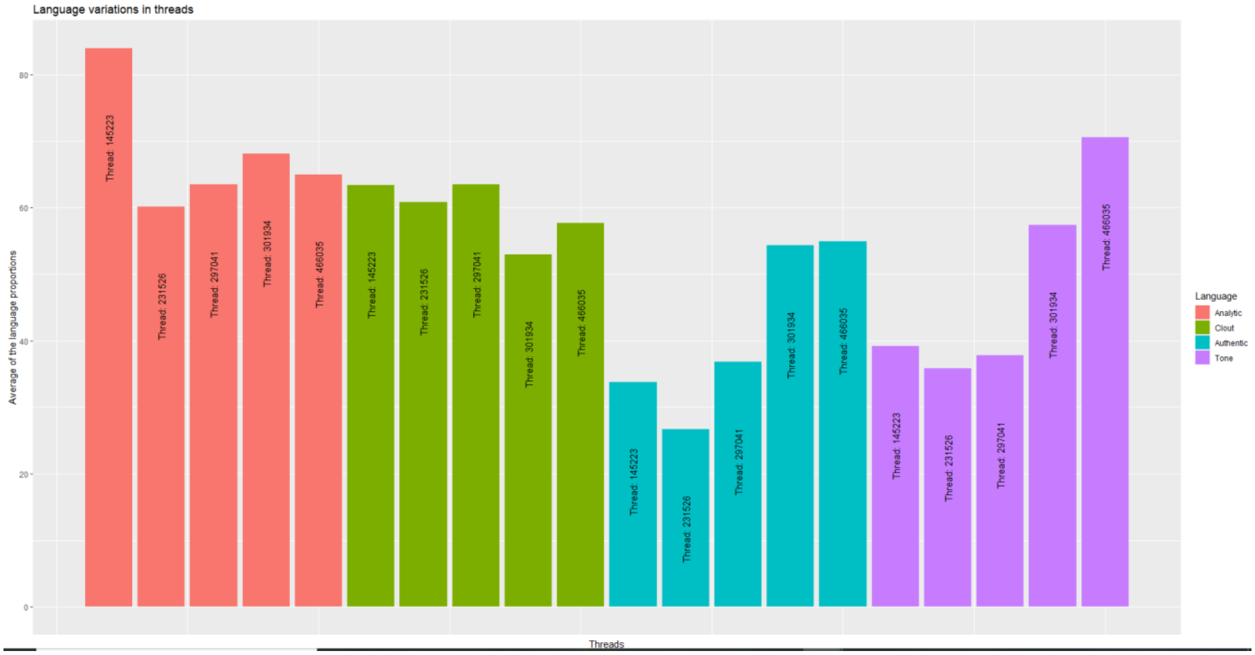
• Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.

In the given data, there are in-total 600 threads consisting which means there are 600 topics that groups of participants are communicating upon. So, there are 2453 unique authors exists in the data provided

Thread 252620 (Thread_id) is the topic which is repeated more than any other topic as it is been 335 times participants have shown interest in it. The average value for the number of times authors talk on the same topic is 33.33 while the least time the same topic is discussed 12.

We could also find the first and last post of the given thread. For thread id=252620, the first and last date of the posts on the thread are 07-12-2005 and 20-12-2006 respectively.

• By analyzing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by these different groups?

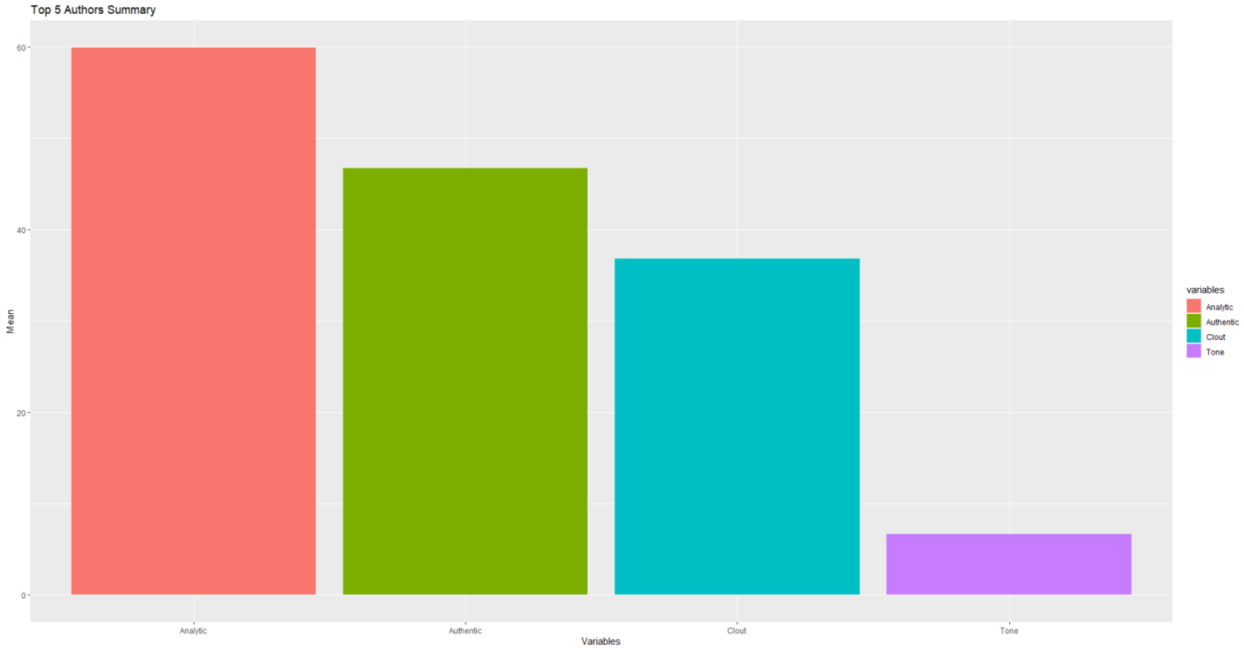


There are plenty of languages used in all threads but, to see the difference in the language, It would be best to show multiple threads in the same graph and the language being changed will show the difference. Here, the top-5 most frequent threads have been selected to check for the language difference. As the top-5 frequent threads mean the most popular forum posts it has been on social media, there will be variety of languages used over such posts. Hence, taking the five threads and seeing the language variation would make it easier to see the pattern.

As in for Clout, there is not much of difference between the proportion values of different threads. All of the threads average around 55 which wouldn't let us see the drastic change in Clout language. For Analytic class, it seems the values of all threads lie between 60 and 70 except for thread 145223. Thread 145223 has moved way too high in terms of analytic variable values which means this thread would have had much more analytical and logical side of discussion than any other threads. There's certainly interesting data varying for that part. The top five threads have seen changes in Tone class of languages as the first three threads have the around average value which does not have much variation. But for the other two threads 301934 and 466035, they both vary from the other threads so it's much clear difference between each thread. Lastly, for Authentic language, the same thing happens as for the Tone language where first three threads have the same amount of data and the other two make difference equally. Therefore, At the end, it is not easy to see the difference for all the threads as some of the threads have the similar discussion going on which results in the same proportional number in some languages. However, more the number of threads you choose, the higher chances the graph will have to distinguish between threads and from there to distinguish between languages, finally.

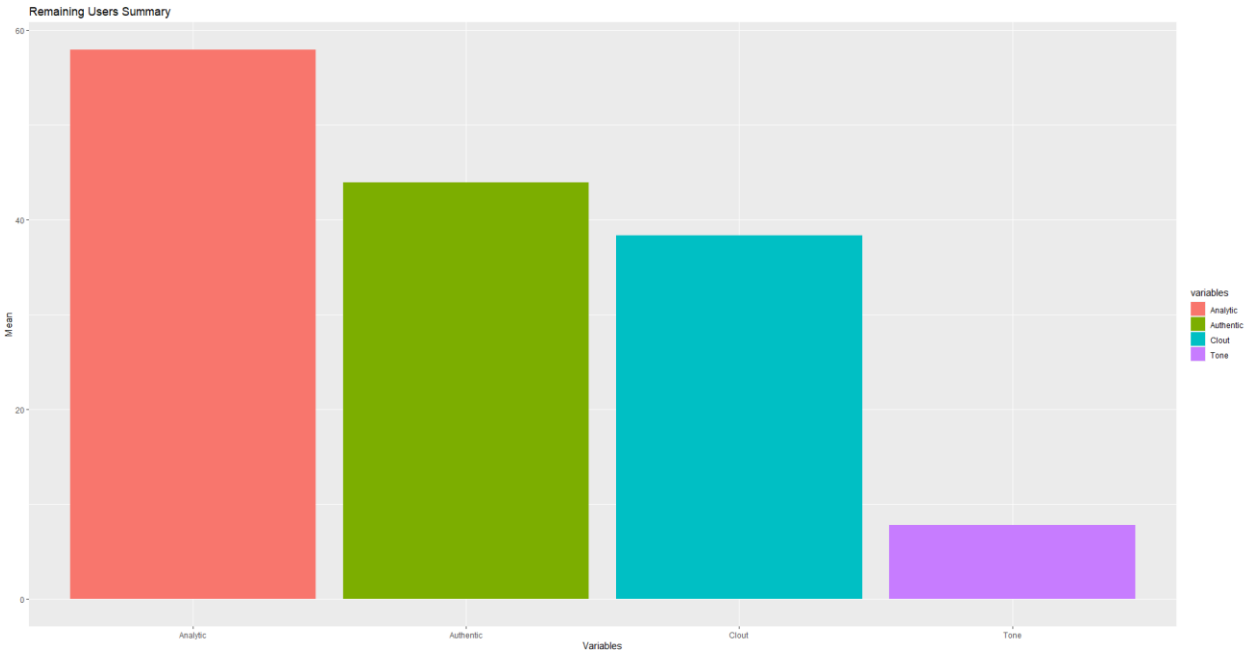
- Does the language used within threads change over time?

The language used within threads can be plenty of way to represent that. Now, one way we can test if the top authors use the same language as any other users. For that, we split data into two parts where one represents first five most frequent authors and the rest are the remaining authors in the data. Then the two sets of data can be compared to see the result for languages importance.



Again, we took Analytic, Authentic, Tone, Clout to make visualization and results clear. The above graph is for the top- 5 users vs the languages being used while making posts on social media. It can be seen that the mean value for Analytic is better than Authentic followed by Clout and Tone.

It follows the same path for the Remaining authors as well. However, the mean value lies slightly lower than top author graph.

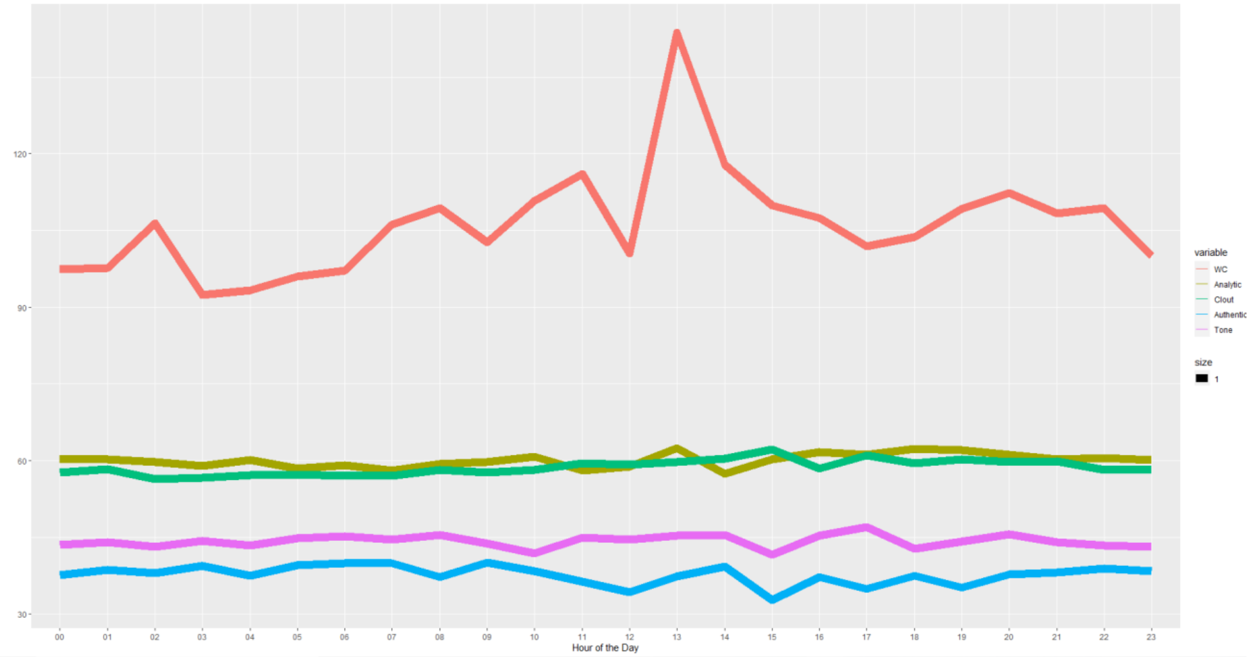


To check for the difference between the top-5 and the other users, we will conduct a one-sided t-test. There is the difference between clout mean values which we can see by comparing two graphs. Therefore, the test finds if the clout is smaller in the top-5 authors than other authors and being 99% confident about it.

The p-value for the test is 2.2*10^-16 and confidence interval does not include zero which rules out the possibility of being zero. However, since we have the smaller p-value which is significant, we can ignore the null hypothesis and can support the alternative hypothesis that is that the clout in top 5 authors is greater than clout in the other authors on average with 99% significance. To test the mean difference for authentic, another t test is conducted. And the p-value is again 2.2*10_16 which again supports the alternative hypothesis that mean value of authentic in top 5 users is higher than the remaining ones. The results above neatly states that the top 5 users are trying to be authentic as they would try to behave very nicely on the forum but at the same time, they might try to get power in their writing as compared to the remaining users.

As it was proved that the top-5 authors have the higher mean value than the rest of the authors overall, we would like to extend this form to state that the higher frequency of authors are most likely to fall into higher mean values. For the thread, if we take the most frequent thread in the same way to find the interactive results, we will deepen our analysis into the thread which would possibly get us the most languages available in the frame. In that we could deepen our understanding to see how the trend for languages in thread are behaving as the time changes.

,



The graph above shows the thread analysis on the hourly basis on any given day. The time is chosen small to increase the insights into data. The data was worked out by getting hour constraint from the time variable and then using it for all posts.

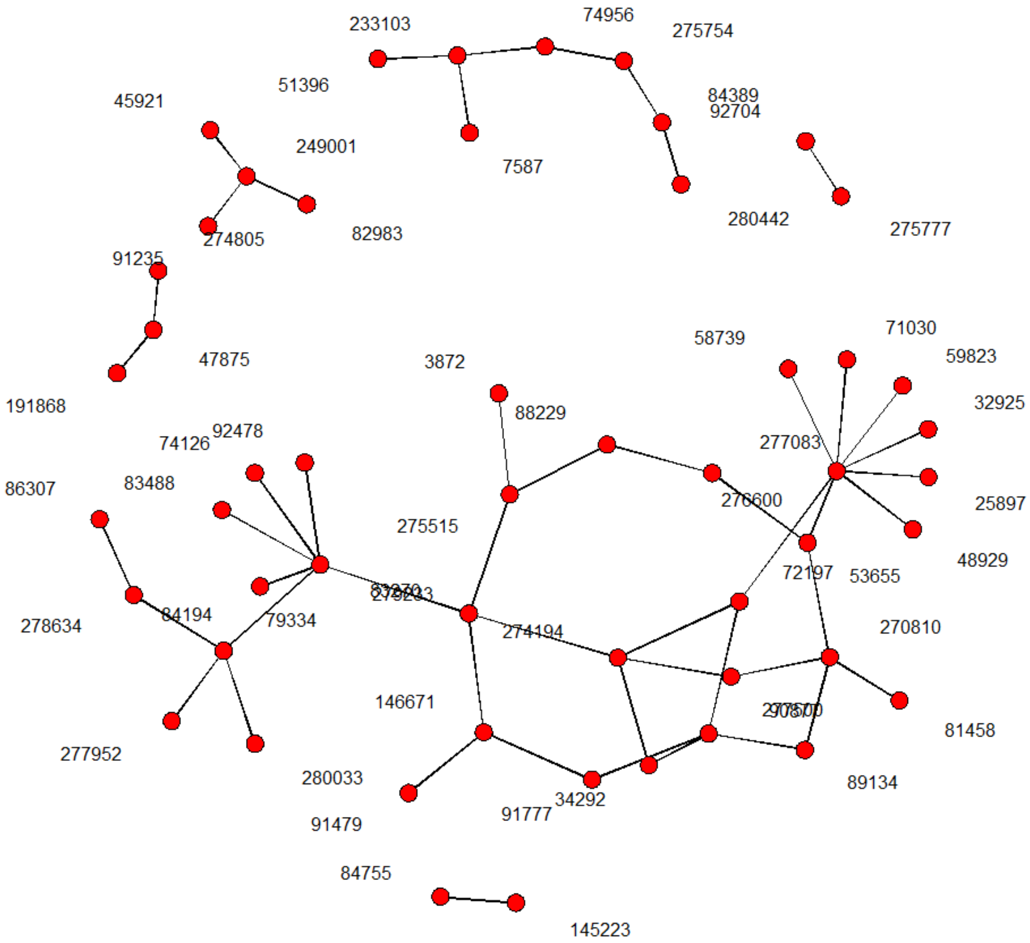
One of the most interesting part about this graph is that the word count which is number of words in the text is rising up towards the late morning and it is on top at 1 pm. We understand that the mid afternoon must be the busiest time at work but it's surprising to see people writing more and long post at this time although it's for all days including weekends and public holidays. Another thing you would expect is to have the least posts during midnight which is exactly what's happening.

For the analytic part, people are expected to be productive the most in the morning or early afternoon. However, surprisingly, there is no trend outlies for analytic part of the graph as it's straight and it falls down slowly at 11 am and rise back up around 1 pm. So, in the afternoon, the trend has been inconsistent but there's no usual trend. Whereas, the same thing takes place for the clout variable as it doesn't show anything more than unstable straight line. Authentic language is suffering below average in some times. We could see that after certain time in the morning the trend starts falling down till midday. Sometimes people just don't feel like behaving soft on social media especially if they are busy or involved in some other activity which is exactly the reason to see the trend having below values. There are participants who slowly start posting authentic parts in their posts and the afternoon trend is just how it's going but, in the evening,, people certainly find time to post more about authenticity on social media. Tone has the exact same trend as Authentic except the average mean value. It really makes it depending on Authentic as participants do feel like making emotional messages while at the same time sending out authentic points.

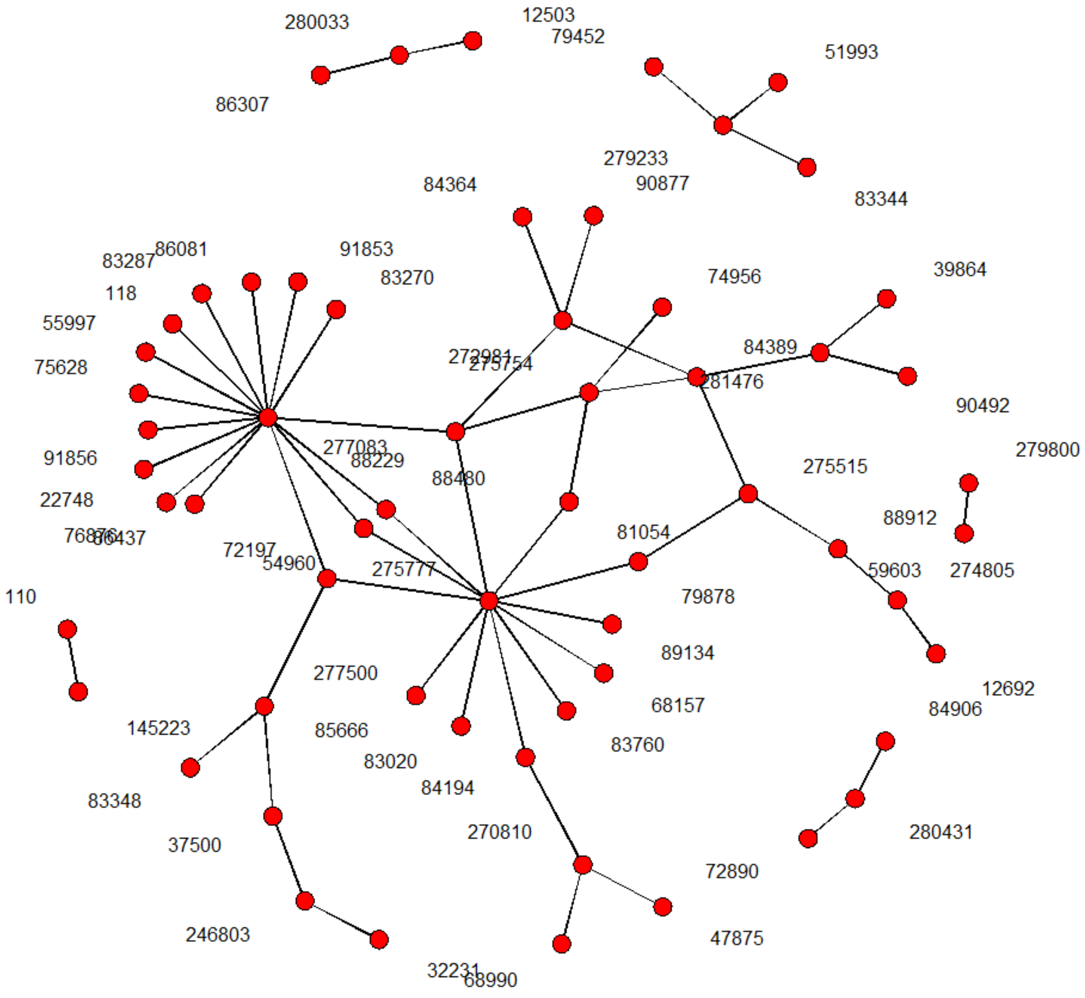
Hence, the hour to hour analysis for the linguistic variables does not give the significant changes into trend as it’s just changing by hours and it included 365 days a year. Therefore, the variables like Authentic, WC, Tone still changes over time but not at the significant rate.

Task-c:

Network connection on Sunday:



Network communication channel on Monday:



We can think of participants communicating on the same thread at the same time (for example during the same month) as forming a social network. When these participants also communicate on other threads, they extend their social network.

According to the data, there have been so many threads forming up each month and each year. Being a little extra careful while choosing the timeline for the network is one of most important tasks. In this case, comparing between two time period would make it reasonable approach. Therefore, the idea is to choose the busiest (most frequent posts) month of the busiest year. That will make sure you have range of thread options available to include most of them. Here, it is comparing between all the Sunday's and Monday's of the busiest month of the year. So, we are doing analysis for the posts made on Sunday and Monday of march,2006. The inner join function is used to merge author id and thread id into adjacency matrix for the connected vertices. Therefore, Authors get connected whenever they make a communication or message on the same thread and once, they go into another thread meaning another vertex, it will extend the graph. [The network diagram based on posts on all Sunday in March,2006]

For both threads, there has been a lot of topics been discussed on social media. As there are a vertex degree of 30 and 24 for thread_id 277083 and 275777 on Monday threads correspondingly which means people have been interested in communicating on the same topic on Monday more than Sunday as these are really high degrees that this data has got. While for Sunday, the maximum degree is 16 which can tell how participant might want to just relax over the weekends and maybe discussing on the new topics for a few participants.

Based on the analysis of all the authors and threads, thread id= 276600, 83270 are the most important vertices in the graph as 276600 rank no. 1 in degree and no.2 in eigen vector while 83270 ranks no. 1 in betweenness. Hence, these threads are the most important vertices in the graph of the values on Sunday in March,2006.

While for the other graph on Monday, thread id= 275777, 277083 and 88229 are the most important vertices in the graph as they ranked no. 1 in betweenness, degree and eigen vector distance correspondingly. There has been one thing clear that the threads participants discuss a lot on Sunday or Monday did not affect on them discussing the topic on any other day. It means that they did not have any of the most important threads in common which means they start the new week (Monday) with the fresh new day and new topic. And there will be another new topic to make a thread on. In conclusion, there have been significant changes in the threads and their connection with Authors as you check for the next day. The reason behind taking all thread_ids on Sunday and Monday was to see the difference and check for any similarities. However, in this case, we can conclude by saying that the networking connections on Sunday on the given thread does not necessarily carry over on Monday into weekdays.

The data given was only based on the limited variables for psychological point of view and more research with time allowance could make a huge difference as it can allow us to deep further into the topic.

Assumptions:

The anonymous authors indicated by authors id -1 have been removed from the data in the initial stage of the process. It's believed that anonymous authors could lead to effective results but at the same time, they are not same all the time so considering them all as a same person could lead to major change in the data relevancy, knowing the fact that it's posted by different people in real. Hence, taking out the anonymous authors would add more honesty into data, and we can perform the analysis on the rest of the data.

There is a thing to notice about the data set that there are entries with a zero word count- this would not provide any meaningful insights into our data and zero word count might mean the deleted post or just any random symbol or reaction to the thread. It makes insufficient to get track of the data since there are varieties of variable so removing such posts would make analysis easier and will provide the real meaning behind the data.

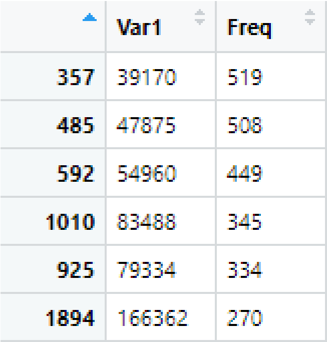
While doing the t-testing 99% of confidence interval is always used. To get the values of date, month, year, day of the week, etc., the data and time variables have been factorized.

All the linguistic variables including summary language and the other variables including grammar, psychological processes values are in percentages scale (0-100) %.

For part-b, assuming we are only dealing with Analytic, Clout, Authentic, Tone variables. The reason behind removing the other variables having small proportion rate is that these four variable cannot be scaled properly with the smaller variables and focusing on these four would give more effective and clean result while it would include the effect from the rest of the variables at the same time.

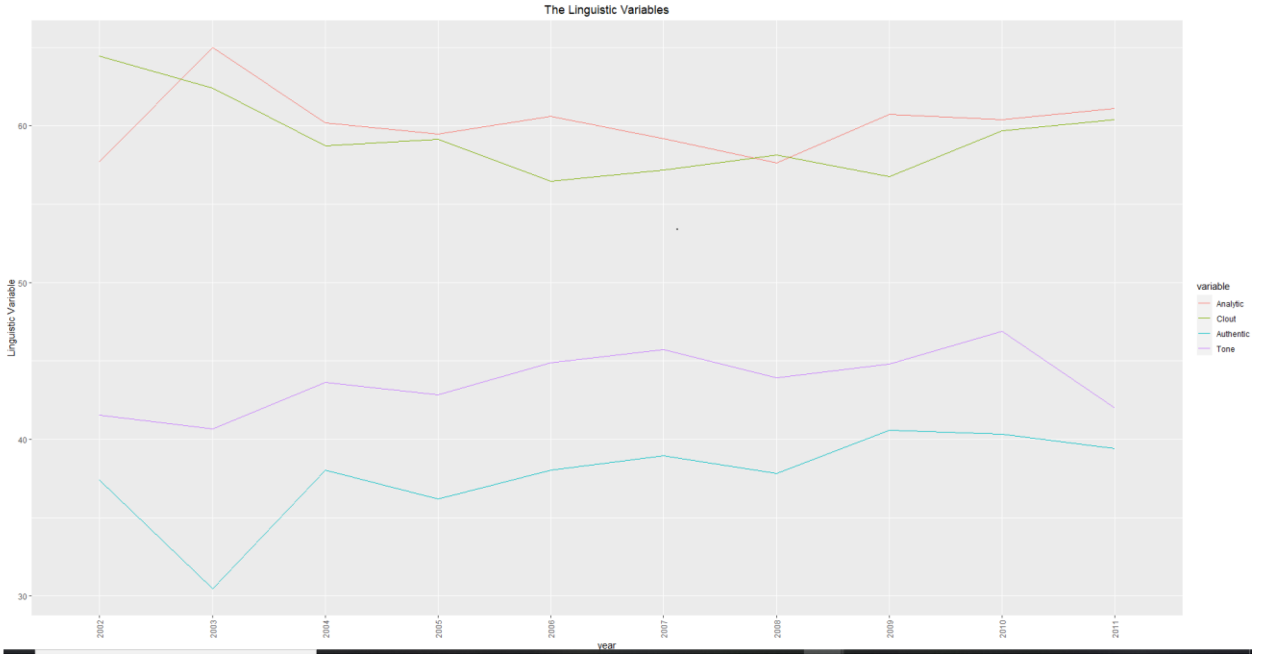
Appendix:

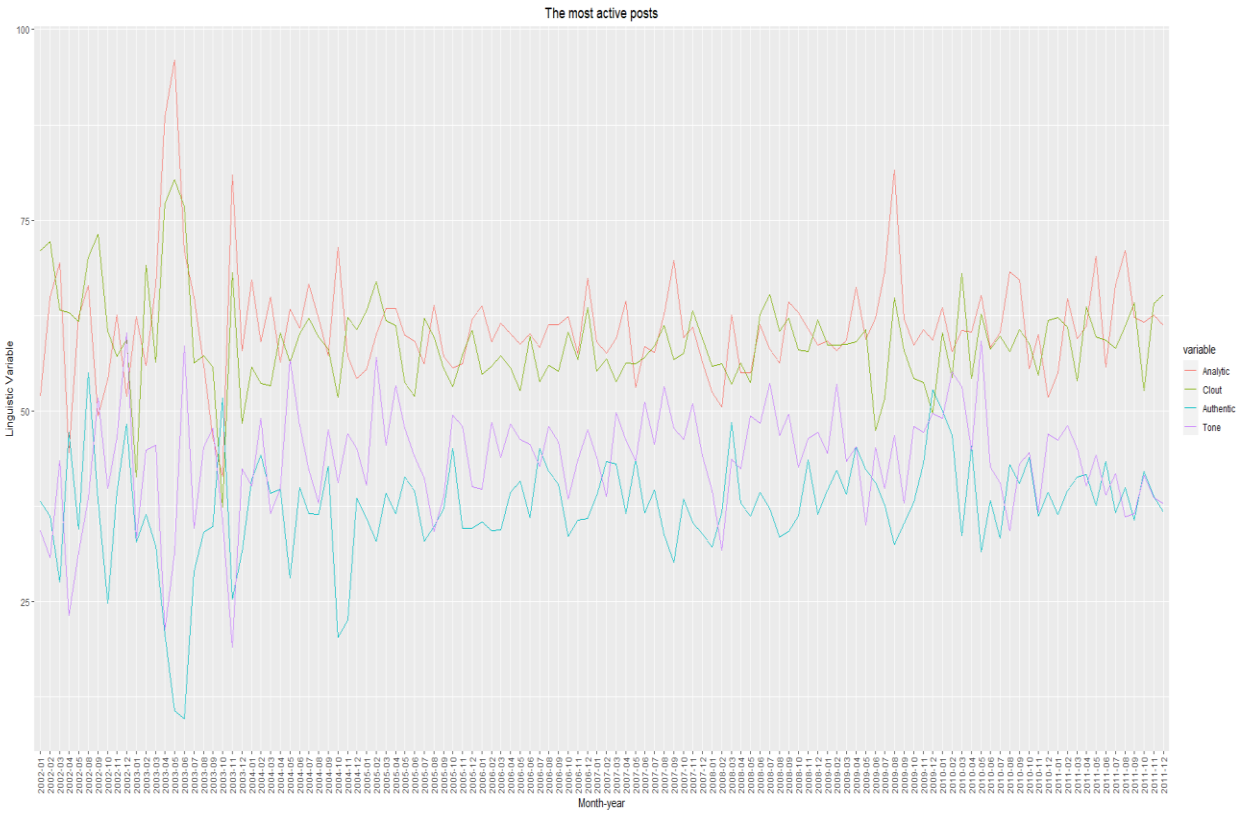
1.

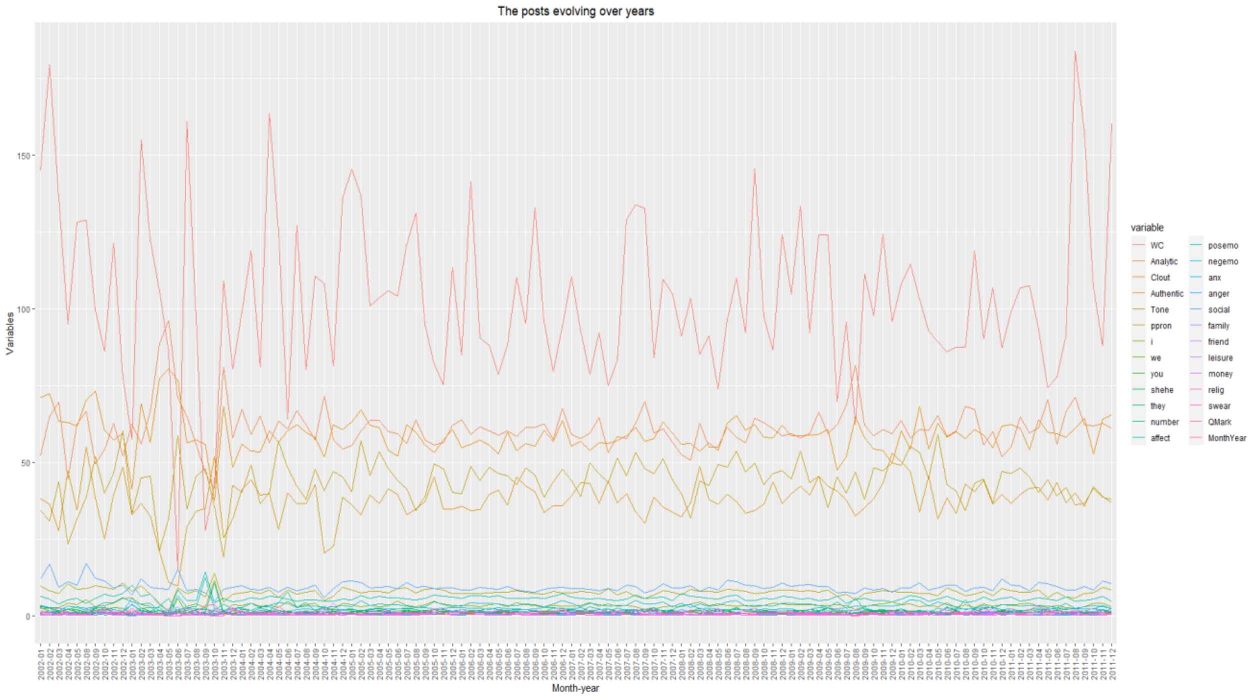


	Var1	Freq
357	39170	519
485	47875	508
592	54960	449
1010	83488	345
925	79334	334
1894	166362	270

2.




The most active posts
Linguistic Variable
100
75
50
25
Month-year
2002-01
2011-12
variable
Analytic
Clout
Authentic
Tone


The posts evolving over years
Variables
150
100
50
0
Month-year
2002-01
2011-12
variable
WC
Analytic
Clout
Authentic
Tone
ppron
i
we
you
shehe
they
number
affect
posemo
negemo
anx
anger
social
family
friend
leisure
money
relig
swear
QMark
MonthYear

R code:

```
install.packages("reshape2")
install.packages("ggplot2")
library(reshape2)
library(ggplot2)
library(plyr)
library(dplyr)
library(lubridate)
library(ggpubr)
library(igraph)
library(igraphdata)
library(network)
library(sna)


#Task-1: Graph-1
#clean the data and generate the new datasets with the random sample size 20,000
rm(list = ls())
set.seed(29143926) # XXXXXXXX = your student ID
webforum <- read.csv("webforum.csv")

#removing anonymose auther and zero word posts from the data
webforum <- subset(webforum, AuthorID != -1)
webforum <- subset(webforum, WC != 0)

webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows

webforum$Date <- as.Date(webforum$Date)
```

```
webforum$Year <- format(webforum$Date, "%Y")
webforum$MonthYear <- format(webforum$Date, "%Y-%m")


#1.1
#Using the author ids to to group into by year
posted_by_auther <- as.table(by(webforum,webforum$Year,nrow))
posted_by_auther <- as.data.frame(posted_by_auther)
mean(posted_by_auther$Freq)

#Trend for author frequencies vs year
plot(posted_by_auther$webforum.Year,posted_by_auther$Freq, main="Overall Posts Trend For All Authors",
     xlab="Year", ylab="Number of Posts")
lines(posted_by_auther$webforum.Year,posted_by_auther$Freq)


#Get the id for the maximum posts by the auther
counting_Author <- as.data.frame(table(webforum$AuthorID))
counting_Author <- counting_Author[order(-counting_Author$Freq), ]

max_auther_id <- counting_Author[which.max(counting_Author$Freq), 1]
m <- mean(counting_Author$Freq)

authr_top <- webforum[webforum$AuthorID %in% max_auther_id,]
author_with_most_posts <- as.table(by(authr_top,authr_top$Year,nrow))
author_with_most_posts <- as.data.frame(author_with_most_posts)

#plotting values for maximum autor only
plot(author_with_most_posts$authr_top.Year,author_with_most_posts$Freq, main="Overall Posts Trend For Top Author",
    xlab="Year", ylab="Number of Posts")

lines(author_with_most_posts$authr_top.Year,author_with_most_posts$Freq)



#------------------------------------------------------

#Applying the regression for the year 2006 for the Affect variable to see the linear relationship

web_2006<-authr_top[authr_top$Year=="2006",]


delete_columns <- c("ThreadID","AuthorID","Date","Time","MonthYear","Year")

web_2006 <- web_2006[,!(names(web_2006) %in% delete_columns)]


attach(web_2006)

authr_top.model <- lm(affect~.,data = web_2006)



author_plot <- ggplot(data=web_2006, aes(x=posemo,y=affect))+geom_point()

fitted <- coef(lm(affect~posemo,data = web_2006))

author_plot + geom_abline(intercept = 1.7836, slope=0.98135)

cor(web_2006$affect,web_2006$posemo)



author_plot2 <- ggplot(data=web_2006, aes(x=negemo,y=affect),labs(subtitle="Affect vs Nege- correlation:0.5564"))+geom_point()

fitted <- coef(lm(affect~negemo,data = web_2006))

author_plot2 + geom_abline(intercept = 3.4392, slope=1.0501)

cor(web_2006$affect,web_2006$negemo)



#--------------------------------------------------------------------------------------------------------------------------

#Appendix[2]
```

```
# Frequency of unique threads

counting_threads <- as.data.frame(table(webforum$ThreadID))

counting_threads <- counting_threads[order(-counting_threads$Freq), ]



thread_id_with_maxposts <- counting_threads[which.max(counting_threads$Freq), 1]

max_thread_posts <- webforum[webforum$ThreadID == thread_id_with_maxposts,]

means_max_thread <- aggregate(max_thread_posts, list(max_thread_posts$Date), mean)



# Remove columns we won't need to plot

deleted_columns <- c("Time", "Date", "AuthorID", "ThreadID", "PostID", "WC", "Month", "MonthYear", "ppron", "i", "we", "you", "shehe", "they",
"number", "affect", "posemo", "negemo", "anx", "anger", "social", "family", "friend", "leisure", "money", "relig", "swear", "Qmark")

means_max_thread <- means_max_thread[ , !(names(means_max_thread) %in% deleted_columns)]



# Melting data to get into format to get plotted

melted_means_max_thread<- melt(means_max_thread, id="Group.1")



# Plot variables against time

ggplot(melted_means_max_thread, aes(x=melted_means_max_thread$Group.1, y=melted_means_max_thread$value, colour=variable)) +
geom_line(size=0.2) + labs(x="Date", y="Value")



#---------------------------------------------------------------------------

#--------------------------------Task-2-------------------------------------------------

#---------------------------------------------------------------------------------------



#------------------------------------------------------------------------------------

#2.1



#get the frequency for the threads in the data and ordering them to extract top-5 for further analysis

counting_threads <- as.data.frame(table(webforum$ThreadID))
counting_threads <- counting_threads[order(-counting_threads$Freq),]

summary(counting_threads)



head(counting_threads)



top_threads <- as.numeric(as.character(head(counting_threads$Var1, n=5)))

top_threads <- webforum[webforum$ThreadID %in% top_threads,]



keep_columns <- c("Analytic", "Tone", "Clout", "Authentic", "Group.1")



#Get the mean average for the whole data and grouped by thread_id

top_threads_averages <- aggregate(top_threads, list(top_threads$ThreadID), mean)

top_threads_averages <- top_threads_averages[, (names(top_threads_averages) %in% keep_columns)]

top_threads_averages_melted <- melt(top_threads_averages, id="Group.1")



labels <- c("Thread: 127115", "Thread: 145223", "Thread: 252620", "Thread: 283958", "Thread: 472752", "Thread: 127115", "Thread: 145223", "Thread: 252620", "Thread: 283958", "Thread: 472752", "Thread: 127115", "Thread: 145223", "Thread: 252620", "Thread: 283958", "Thread: 472752", "Thread: 127115", "Thread: 145223", "Thread: 252620", "Thread: 283958", "Thread: 472752")

ggplot(data = top_threads_averages_melted, aes(x = seq(1:length(top_threads_averages_melted$value)), y = top_threads_averages_melted$value, fill = top_threads_averages_melted$variable)) +

  geom_bar(stat = 'identity', position = 'dodge') +

  theme(axis.text.x = element_blank(), axis.ticks.x=element_blank()) +

  geom_text(aes(label=labels), angle =90, hjust=2) +

  scale_fill_discrete(name = "Language") +

  xlab("Threads") +

  ylab("Average of the language proportions")+

  ggtitle("Language variations in threads")



#--------------------------------------------------------------------------------------------------------------------------------------------
```

```
#2.2


#Extracting the top 5 frequent authors

counting_Author <- as.data.frame(table(webforum$AuthorID))

counting_Author <- counting_Author[order(-counting_Author$Freq), ]



authr_top_top_5 <- as.numeric(as.character(head(counting_Author$Var1, n=5)))

authr_top_top_5_posts <- webforum[webforum$AuthorID %in% authr_top_top_5,]




#The remaining author values

remaining <- counting_Author[counting_Author$Freq<300,]

remaining_Auth <- webforum[webforum$AuthorID %in% remaining$Var1,]



#mean for the values and storing them

top5mean <- as.data.frame(colMeans(authr_top_top_5_posts[6:10]))

colnames(top5mean) <- c("mean")


remaining_mean <- as.data.frame(colMeans(remaining_Auth[6:10]))

colnames(remaining_mean) <- c("mean")



variables <- c("Analytic", "Clout", "Authentic", "Tone")


top5summary <- top5mean[c(2:5),]

top5summary <- data.frame(variables, top5summary)
```

```r
colnames(top5summary) <- c("variables", "means")


remaining_summary <- remaining_mean[c(2:5),]

remaining_summary <- data.frame(variables, remaining_summary)

colnames(remaining_summary) <- c("variables", "means")



#Histogram plotting

ggplot(top5summary, aes(variables, means, fill = variables)) +

  geom_bar(stat = "identity") +

  labs(x = "Variables", y = "Mean", title = "Top 5 Authors Summary")


ggplot(remaining_summary, aes(variables, means, fill = variables)) +

  geom_bar(stat = "identity") +

  labs(x = "Variables", y = "Mean", title = "Remaining Users Summary")




#t-test to determine what's the greater value

t.test(authr_top_top_5_posts$Authentic, remaining$Authentic, "greater", conf.level = 0.99)

t.test(authr_top_top_5_posts$Clout, remaining$Clout, "greater", conf.level = 0.99)



authr_top_top_5_posts$Year <- format(authr_top_top_5_posts$Date, "%Y")


by_year <- aggregate(authr_top_top_5_posts, list(authr_top_top_5_posts$Year), mean)
```

```
keep_columns <- c("Group.1", "Analytic", "Clout", "Tone", "Authentic")

by_year <- by_year[ , (names(by_year) %in% keep_columns)]

authr_top_thread_averages <- melt(by_year, id="Group.1")



authr_top_thread_averages$value <- format(round(authr_top_thread_averages$value, 2), nsmall = 2)



#Appendix[2]


ggplot(data= authr_top_thread_averages, aes(x=Group.1, y=authr_top_thread_averages$value, colour=variable, group=variable)) +

  geom_line() +

  theme(axis.text.x = element_text(angle = 90, vjust=0.5)) +

  theme(plot.title = element_text(hjust = 0.5))+

  ggtitle("The most active posts")+

  xlab("Month-year") +

  ylab("Linguistic Variable")



#--------------------------------------------

#2.3


#plotting hour by hour after changing the time format to hour


webforum$Date <- as.Date(webforum$Date)

webforum$Day <- weekdays(webforum$Date)

webforum$Hour <- substr(webforum$Time, 1, 2)

webforum <- webforum[order(webforum$Day), ]

webforum$Month <- format(webforum$Date, "%Y-%m")
```

```
keep_columns <- c("Analytic", "WC", "Clout", "Authentic", "Tone", "Group.1")


by_hour <- aggregate(webforum, by=list(webforum$Hour), mean)

by_hour <- by_hour[ , (names(by_hour) %in% keep_columns)]

by_hour <- melt(by_hour, id="Group.1")


ggplot(by_hour, aes(x=Group.1, y=value, colour=variable, group=variable)) + geom_line() + xlab("Hour of the Day") + ylab("")


#----------------------------------------

#3.1*

#---------------------------------------


#Performing analysis for march,2006 Sundays and getting substring data for it

webforum$Year <- format(webforum$Date, "%Y")

webforum$MonthYear <- format(webforum$Date, "%Y-%m")

webforum$Day <- weekdays(webforum$Date)

webforum$Hour <- substr(webforum$Time, 1, 2)


no_post <- as.table(by(webforum,webforum$Year,nrow))

no_post<-as.data.frame(no_post)


web_2006<-webforum[webforum$Year=="2006",]

no_post2<-as.table(by(web_2006,web_2006$MonthYear,nrow))

no_post2<-as.data.frame(no_post2)
```

```
web_2006_03<-webforum[webforum$MonthYear=="2006-03",]
no_post3<-as.table(by(web_2006_03,web_2006_03$Date,nrow))
no_post3<-as.data.frame(no_post3)

web_2006_day <- web_2006_03[(web_2006_03$Day=="Sunday"),]
web_2006_day <- select(web_2006_day, AuthorID,ThreadID)

#inner joing the auther and thread id data to get the data into adjacency form later
g1 <- dplyr::inner_join(web_2006_day, web_2006_day, by = "ThreadID")[,-1]
g2 <- apply(g1, 2, as.character) #AuthorID as character will become vertex ID

library(network)
G <- network(g2, directed = FALSE)
adjMatrix1 <- as.sociomatrix(G)
adjEdge <- as.edgelist(G)
plot(G, label = G%v%"vertex.names")


#Finding all the graph related values like degree, betweenness, closeness, diameter
degree = as.table(degree(adjMatrix1))
betweenness = as.table(betweenness(adjMatrix1))
#closeness = as.table(closeness(adjMatrix1))
eig = as.table(evcent(adjMatrix1))

#averagePath = average.path.length(G)
#diameter = diameter(G)

tabularised = as.data.frame(rbind(degree, betweenness, eig))
```

```
tabularised= t(tabularised)


#----------------------------------------------------------------

#Performing analysis for march,2006 Mondays and getting substring data for it
web_2006_Monday <- web_2006_03[(web_2006_03$Day=="Monday"),]
web_2006_Monday <- select(web_2006_Monday, AuthorID,ThreadID)

#inner joing the auther and thread id data to get the data into adjacency form later
h1 <- dplyr::inner_join(web_2006_Monday, web_2006_Monday, by = "ThreadID")[,-1]
h2 <- apply(h1, 2, as.character) #AuthorID as character will become vertex ID


library(network)
H <- network(h2, directed = FALSE)
adjMatrix2 <- as.sociomatrix(H)
adjEdge2 <- as.edgelist(H)
plot(H, label = H%v%"vertex.names")



#Finding all the graph related values like degree, betweenness, closeness, diameter
degree = as.table(degree(adjMatrix2))
betweenness = as.table(betweenness(adjMatrix2))
#closeness = as.table(closeness(adjMatrix2))
eig = as.table(evcent(adjMatrix2))

#averagePath2 = average.path.length(H)
```

```
#diameter = diameter(H)


tabularised2 = as.data.frame(rbind(degree, betweenness, eig))

tabularised2 = t(tabularised2)
```

References:

1.)https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf

2.)https://www.vitalsmarts.com/crucialskills/2015/06/the-differences-between-behavior-and-culture/

3.) https://unamo.com/blog/social/sentiment-analysis-social-media-monitoring