Analysing Activity, Language Use, and Social Interactions in an On-line Community from 2002 to 2011

Grace Nathania

INTRODUCTION

This report analyses the activity, language usage and social interaction in an online community. The data used in this analysis consists of metadata and linguistic analysis, obtained using Linguistic Inquiry and Word Count (LIWC) system, in 20,000 posts over the years from 2002 to 2011. LIWC assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words (29 variables) used in communication. Below is the summary of the data;

VARIABLE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
WC	1.00	29.00	63.00	103.40	128.00	5195.00
Analytic	1.00	40.12	63.46	60.08	83.53	99.00
Clout	1.00	41.16	59.64	58.13	79.13	99.00
Authentic	1.00	10.08	30.61	38.26	62.75	99.00
Tone	1.00	13.93	25.77	43.92	79.01	99.00
ppron	0.00	4.35	7.30	7.64	10.53	50.00
i	0.00	0.00	2.30	3.34	4.92	50.00
we	0.00	0.00	0.00	0.92	1.19	33.33
you	0.00	0.00	0.00	1.39	1.94	50.00
shehe	0.00	0.00	0.00	0.64	0.00	33.33
they	0.00	0.00	0.00	1.36	2.05	33.33
number	0.00	0.00	0.56	1.83	2.20	100.00
affect	0.00	2.70	4.76	5.65	7.23	100.00
posemo	0.00	0.42	2.41	3.43	4.35	100.00
negemo	0.00	0.00	1.47	2.17	3.17	100.00
anx	0.00	0.00	0.00	0.64	0.00	33.33
anger	0.00	0.00	0.00	0.95	1.31	100.00
social	0.00	5.26	8.70	9.14	12.31	100.00
family	0.00	0.00	0.00	0.38	0.00	66.67
friend	0.00	0.00	0.00	0.38	0.00	50.00
leisure	0.00	0.00	0.00	1.22	1.37	100.00
money	0.00	0.00	0.00	0.55	0.32	50.00
relig	0.00	0.00	0.00	0.61	0.00	50.00
swear	0.00	0.00	0.00	0.30	0.00	100.00
QMark	0.00	0.00	0.00	1.05	0.85	400.00

*summary is rounded into 2 decimal places.

Based on summary above, we can see that Analytic, Clout, Authentic and Tone have high proportion in each post. This is because:

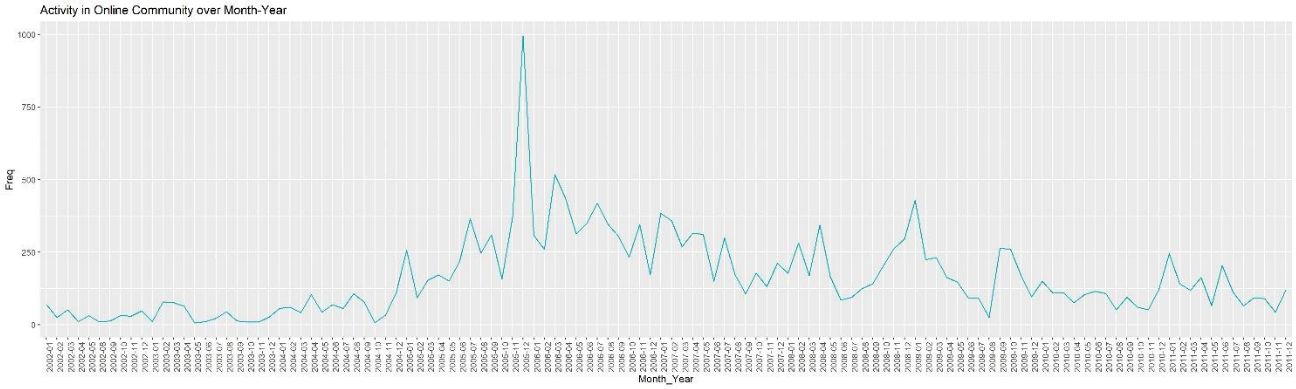
1. Analytic represents analytical thinking used in the post
2. Clout represents power, forced, impact used in the post
3. Authentic represents an authentic tone of voice used in the post
4. Tone represents emotional tone of voice used in the post

which indicate that other variables are dependent to them. We will further analyse the data, by focusing on the average (mean) proportion of each variable to posts, to know, but not limited to, the trend of posts over the years, relationship of variables and posts, and the social network using R studio.

DATA ANALYSIS AND EXPLORATORY

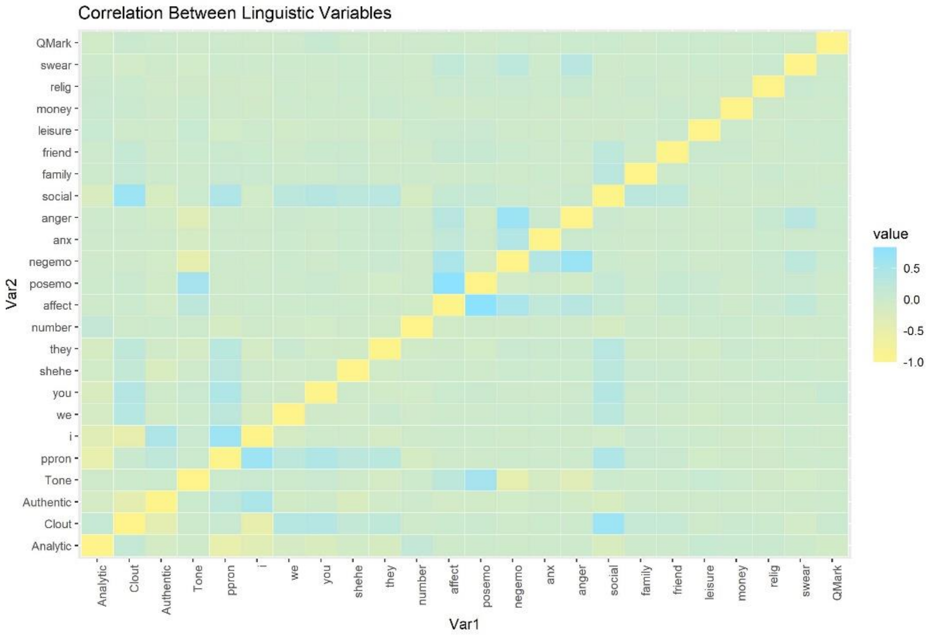
PART A – Analysing Activity and Language on the Forum over Time

The data used is built up of posts on each day, thus to analyse the activity over time, the data is grouped by month-year from the beginning of 2002 to the end of 2011.



Graph above shows us that the internet users' activity always goes up at the end of the year. This number, however, only lasts until January the next year and it will rise and fall moderately afterwards. The number of posts only goes up at this period of time mainly because people have a long holiday, celebrating Christmas, new year, other events, where they have more time to interact and celebrate events together with people online.

Looking at the high peak in December 2005, when internet users were mostly American and European[1], there were many events happened in America and Europe at that time: Ronaldinho was named best football player in Europe in November 2005, 3 football championship games in America, and Latvia, European country, was against LGBT in December 2005[2]. These factors had possibility in affecting the number of posts. Furthermore, in some countries, big events were held only at the end of the year such as great sale, championship games, and holiday packages which increased the tendency of people talking about these kind of events in the online community.



Gaining an insight about participants' activity in the online community, it is important to know the relationship among the linguistic variables. It is beneficial to tell us which variables are dependent to another variable based on their relationship's power and to further analyse the change of the variables over time.

Looking at the heat map on the left and the

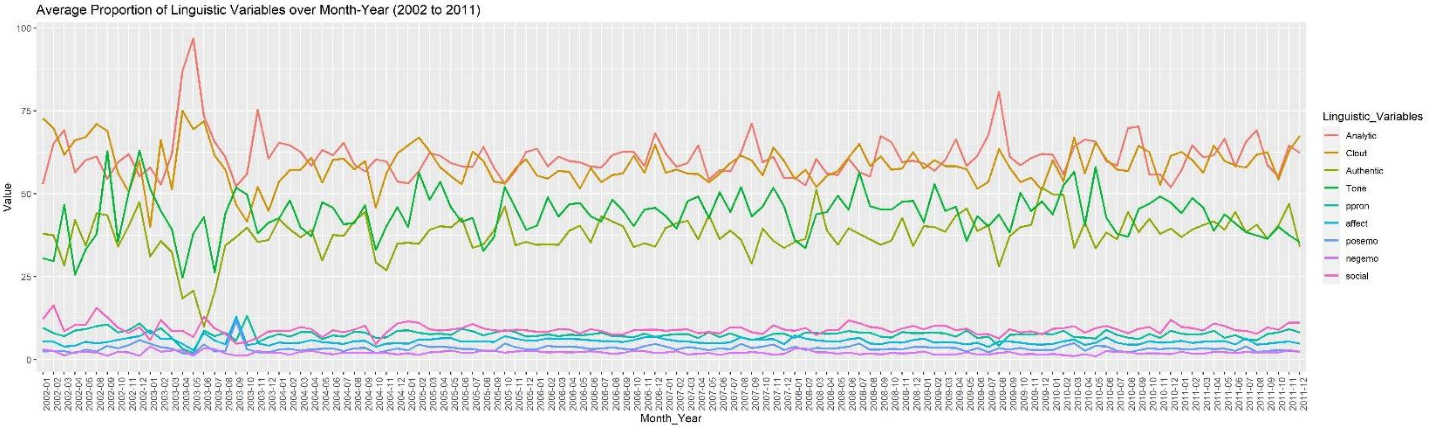
[1] https://www.internetworldstats.com/pr/edi008.htm
[2] https://www.onthisday.com/events/date/2005/november

variables which have high mean value; Analytic, Clout, Authentic and Tone, we will now be able to see from the heat map if other variables are dependent to these 4 variables.

Analytic has relationship with most of the variables which indicates that a post is produced based on analytical thinking. Clout has relationship with all pronouns (i, we, you, shehe, they) and a strong relationship with words referring to social process (social). This tells us that power and forced are used by people in their words when they are talking about social process to bring impacts. Authentic has a strong relationship with i which indicates that when someone talks about himself, his opinions or anything related to one self, he tends to use his own style of communication. Tone is mostly related with positive and negative emotions since it represents the proportion of emotional tone.

Inspecting another variable, all personal pronouns (ppron) definitely has a relationship with all the variables referring to pronouns; i, we, you, she, he, they. Additionally, Affect, where words show sentiments, is related to positive and negative emotions (posemo and negemo) and emotions is related to certain words such as anxiety and anger (anx and anger). Furthermore, words referring to social process (social) have relationship with most of the variables, which probably means the posts mostly talk about the pattern of growth and change in a society over the years.

Knowing all of the information above, Analytic, Clout, Authentic, Tone, ppron, affect, posemo, negemo and social will be taken as the variables to be further investigated. The data will be grouped by month-year to learn if the variables change over time.



Over time, ppron, affect, posemo, negemo and social change steadily while Analytic, Clout, Authentic and Tone rise and fall significantly over 9-year period. The change in proportion of Analytic and Clout rise and fall closely while the proportion of Authentic and Tone also fluctuate closely. At some period of time, such as in 2003, these 4 variables' proportion are in the opposite; Analytic and Clout rocket while Authentic and Tone drop. In a conclusion, each variable does change over time, and Analytic, Clout, Authentic, and Tone are the most insightful information since they show observable change from 2002 to 2011.

To better support the observation of change in linguistic variables over time, t-test is performed with 95% confidence level. t-test is done for 3 variables: Analytic, Authentic and negemo from the data in 2002 and 2011. Analytic and Authentic are chosen since their proportion starts at different value and they keep fluctuating over time. Negemo is chosen to represent variables that rise and fall stably.

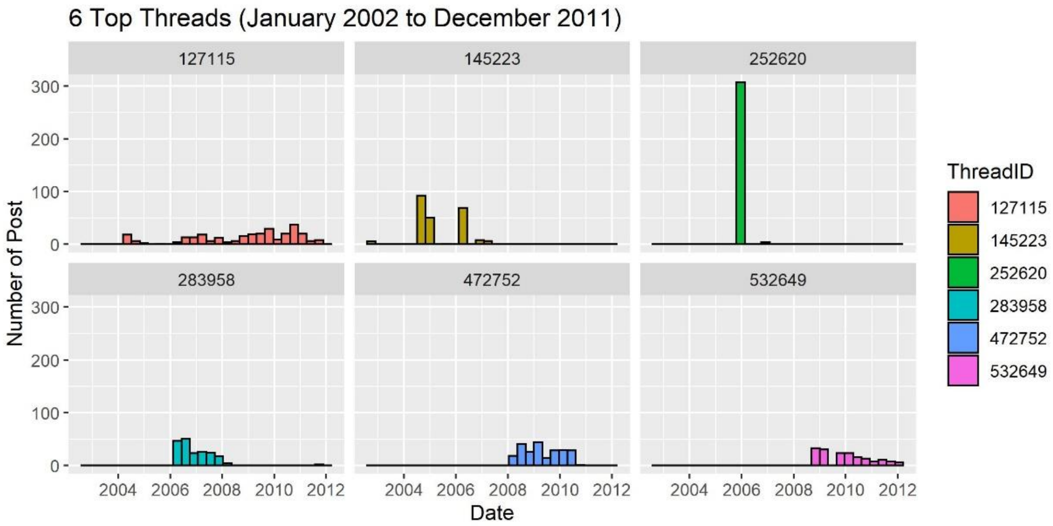
Based on t-test, below is the data of p-value for each variable:

95% C.I.	**Analytic**	**Authentic**	**negemo**
H_0: $\mu_{2002} \neq \mu_{2011}$ vs H_A: $\mu_{2002} = \mu_{2011}$			
p-value	0.3186	0.3326	0.4683

Since all p-values are greater than α = 0.05, we have insufficient evidence to reject null hypothesis. Thus, we can conclude that linguistic variables change over time.

PART B – Analyse the Language Used by Group

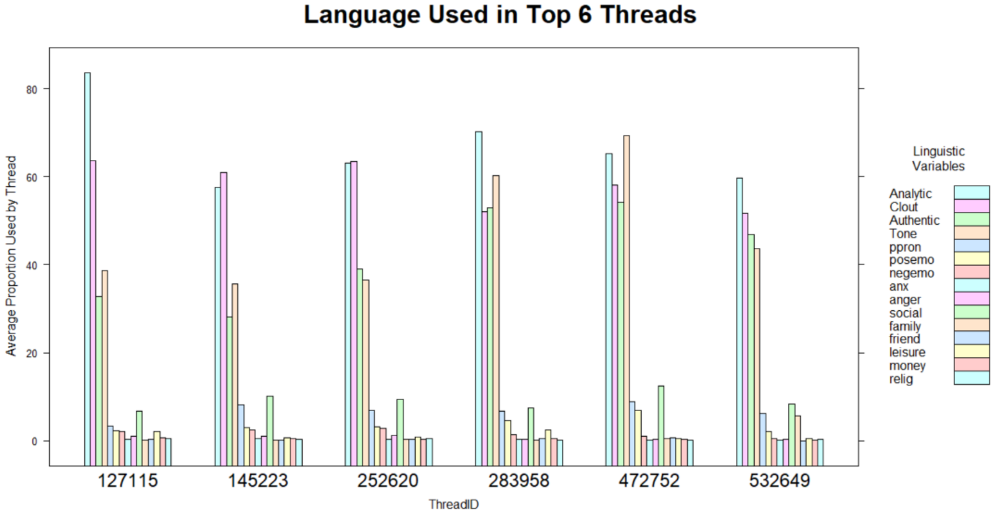
The data is built up by a lot of threads, a group of people communicating on the same topic. Between January 2002 and December 2011, there were some threads continuously grew, some new threads came up or some threads stopped discussing. To further observe the language used by group of people, the 6 top most active threads over 9-year period will be examined. Based on data exploration, threads with the following ID are the most active: 127115, 145223, 252620, 2839, 472752, 532649. Below is a histogram showing each of the thread's distribution in 9 years;



Most threads start to be active in 2004 onwards except thread with an ID of 145223 which is active for a short period of time in 2002 (that small piece of distribution can be called as an outlier and be ignored). Based on the histogram above, thread number 252620 has the highest distribution among all yet the distribution is uneven since the data is only in 2006 and 2007. Thread number 145223 seems to have a bimodal distribution since it has 2 peaks. This is possible to happen since a topic might be back on the trend such as Apple launches new products (iPhone, iPad, Macbook) in an adjacent time.

On the other hands, thread number 283958, 472752, and 532649 have a skewed-right distribution since the threads have the possibility not to be 100 percent active all the time. A lot of factors affect threads' activity such as topic is not up to date anymore. Moreover, this indicates that these 2 threads are active in the beginning of the time and starts to lose people's interest about the topic until the topic is not talked about anymore. Thread number 127115 is a bit skewed-left with an even distribution over the years. This indicates that the topic talked is something people can always talk about such as the continuous change and update in technology.

Examining these 6 top threads and their various distribution over the years, we are now interested to know how linguistic variables take a proportion of the topics talked.



The linguistic variables taken to be analysed are:

1. Analytic, Clout, Authentic, Tone since they have large proportion,
2. ppron since it is related to all personal pronouns,
3. posemo and negemo since they indicate if the post involves positive or negative emotion,
4. anx, anger, social, family, friend, leisure, money and relig since these variables referring to words used in each thread's topic.

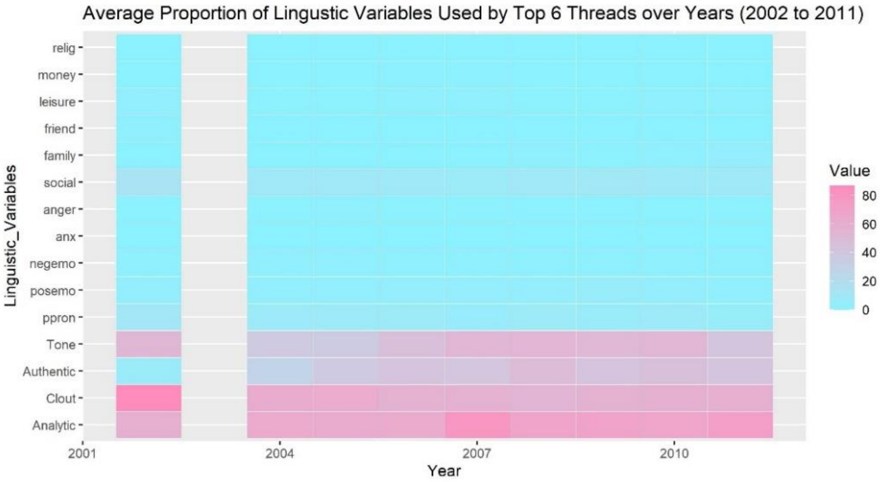
Observing the graph above, there are differences in the language used by each thread. Each thread has different proportion of Analytic, Clout, Authentic and Tone while the variables referring to the words do not show much difference except for words referring to social process (social). Since all threads have higher proportion of positive emotion, we can tell that these 6 top threads are more towards positive rather than negative. Furthermore, we can say that mostly each group talks about the pattern of growth and change in a society over the years (social) since the proportion of this variable is the highest among variables referring to words on certain topics.

Further observing the graph above, thread number 127115 has the highest proportion of Analytic among all, which means people in this group communicate based on analytical thinking. Thread number 472752 has the highest proportion of Tone among all, which we can say the topic in this group involves emotional feeling. Last but not least, thread number 532649 has the highest proportion of words referring to family, which the topic within this group of people is about family. To prove these statement for these 3 threads are true, t-test is conducted on the data of Analytic and Tone (2 t.tests) in thread 127115 and 472752, and on the data of family in thread 472752 and 532649. Based on t-test, below is the data of p-value for Analytic, Tone, and family:

95% C.I.	Analytic	Tone	family
H_0	$\mu_{127115} \geq \mu_{472752}$	$\mu_{127115} \leq \mu_{472752}$	$\mu_{532649} \geq \mu_{472752}$
p-value	1	1	1

Since all p-values are greater than α = 0.05, we have insufficient evidence to reject null hypothesis. Although t-test is not conducted on all thread, sample is sufficient to conclude that: thread number 127115 has the highest Analytic proportion, thread number 472752 has the highest Tone proportion and thread number 532649 simply has the highest proportion of family.

Understanding that each group used different language, the change of language used over time will give an insightful information. Since all 6 top threads started after 2004 onwards except for thread number 145223, the data in 2002 on heat map below solely talks about the language used in it. Additionally, there is no data recorded in 2003.



Looking at the heat map of data in 2002 from thread number 145223, Clout has the highest proportion which is aligned with the bar chart above. Considering the whole heat map, the language used within threads did change over time especially the involvement of analytical thinking, force/power/impact, authentic tone of voice, and emotional tone. The proportion of power/force/impact, analytical thinking and words referring to social process are relatively stable. The proportion of voice's authenticity and emotional tone have slight change over time. Other variables' proportion, such as those referring to certain topics (family, religion, leisure) cannot really be seen from the heat map since they have small proportion, thus t-test is done on one of the variables, friend, with 95% confidence level from the data in 2004-2007 (data A) and 2008-2011 (data B).

$$H_0 : \mu_A \neq \mu_B \quad \text{vs} \quad H_A : \mu_A = \mu_B$$

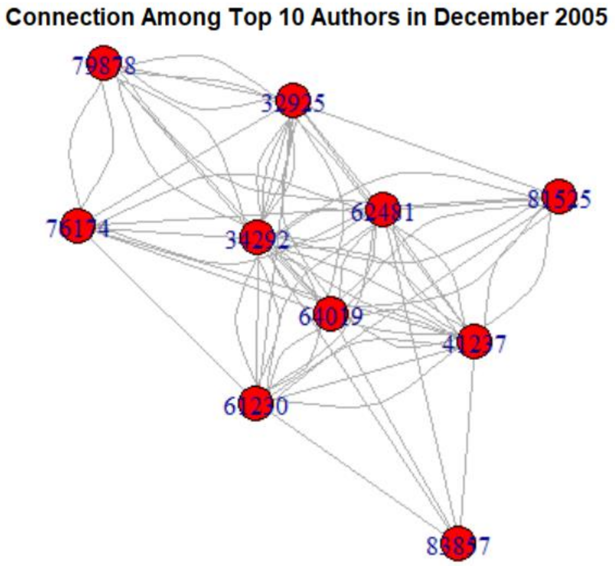
p-value for this test is 0.4794, which is greater than $\alpha = 0.05$, and it makes us have insufficient evidence to reject null hypothesis. Thus, we can conclude friend and other variables referring to certain topic (anx, anger, social, family, leisure, money, and relig) do change over time as well.

PART C – Social Networks Online

Referring to observation in part A first graph that users' activity always goes up at the end of the year and the highest number of post is in December 2005, we choose December 2005 as the period of time to learn the social networks online. We will take the top 10 most active authors and their connections are represented in the adjacency matrix below:

	34292	**76174**	**41237**	**61230**	**64019**	**79878**	**83857**	**62481**	**32925**	**81525**
34292	0	1	4	3	3	3	1	2	5	2
76174	1	0	1	1	1	2	0	2	1	1
41237	4	1	0	3	2	0	1	3	1	2
61230	3	1	3	0	2	0	1	2	1	1
64019	3	1	2	2	0	1	1	2	2	2
79878	3	2	0	0	1	0	0	2	3	0
83857	1	0	1	1	1	0	0	1	0	0
62481	2	2	3	2	2	2	1	0	1	1
32925	5	1	1	1	2	3	0	1	0	1
81525	2	1	2	1	2	0	0	1	1	0

Based on the adjacency matrix above, we can obtain a network graph and necessary data about degree, betweenness, closeness, and eigenvector of each node (author).



Observing the graph above, there are 10 vertices and 69 edges, and the graph is not simple since there are multi-edges. We are interested to know if the network is connected, robust or fragile by analysing the diameter and average path length of the network. This social network has a diameter (the longest distance between any 2 vertices) of 2 and an average distance between any 2 vertices over the whole network of 1.16 (2 d.p.). The histogram of degree distribution, broken into 5, shows a skewed-right histogram which means most vertices only have a few edges while there exist some vertices which are extremely linked. These network analysis data and graph above can tell us that the network is both connected and robust.

In further ado, we want to know who the most important author in this network is. To determine the most important author, centrality measure (Degree, Closeness, Betweenness and Eigenvector) of each vertex is calculated. Below is the data ordered by degree:

AuthorID	Degree	Closeness	Betweenness	Eigenvector
34292	24	0.11	3	1.00
41237	17	0.10	1	0.77
64019	16	0.11	1	0.71
62481	16	0.11	1	0.68
32925	15	0.10	0	0.72
61230	14	0.10	0	0.65
79878	11	0.08	0	0.53
76174	10	0.10	0	0.44
81525	10	0.09	0	0.48
83857	5	0.08	0	0.25

Observing the table above, author number 34292, followed by author number 41237, are the most important authors in the network. Author number 34292 is ranked first in all measure (Degree, Closeness, Betweenness, and Eigenvector) while author number 41237 is ranked second in most measure except Closeness. In addition, both of these authors are involved in the largest clique, strengthening the fact that they are 2 of the most important authors.

In the following months, the social network surely changes since in part B first graph, we can see that threads' activity changes over time, do does the authors' activity. Looking at the closest month to December 2005, January 2006, the top 10 most active authors has changed. In January 2006, the top 10 most active authors are those with ID: 54960, 83344, 39170, 84618, 34292, 47875, 53655, 32249, 62481, and 79334. In January 2006, only author with number 34292 and 62481 from top 10 authors in December 2005 who were still one of the most active authors. Author number 41237 posted 16 posts in December 2005 and he only posted 3 posts in January 2020. Therefore, in conclusion, social networks will always change in the following months.

CONCLUSION

Derived from the metadata and linguistic analysis between 2002 and 2011, we can conclude that internet users' activity varies. The activity tends to increase at the end of the year until January the next year and start dropping afterwards. Considering 24 linguistic variables in the linguistic analysis, variables have relationship among each other and there are some of them being dependent to another variable. Linguistic variables do change over time as well.

Further observing the data, threads, group of people talking on the same topic, do change over time. From the observation of the 6 top most active threads, a thread mostly starts from 2004 onwards and not all threads last for a long period of time. Furthermore, each thread discusses different topic which indicates that there is a difference in the language used. Since threads' activity changes over time, so does the language usage.

In addition, the analysed data has an online social network among the users involved in threads. Examining a social network in December 2005, the top 10 most active users are mostly connected to one another. However, since there is change in users' and threads' activity, the social network changes in following months, not only in this period of time (2002 to 2011), but even after 2011 as well.

To sum up, there are a lot of factors, both external and internal, affect the activity, language usage, and social interaction in online community. With the growth of the internet and the development of social media platform, the number of internet and social medias' users are indeed increasing. Because of this increasing number, the activity, language usage and social interaction online will continue to change intermittently depending on the world's situation. Considering current pandemic of COVID-19, where people have to stay at home and rely on the news online (or offline through the news on television) to receive faster information regarding current condition and situation, the activity in online community and post in social medias surely surge dramatically which will influence the language used and the social interaction.[3]

[3] https://www.bbc.com/news/technology-52052502

APPENDIX – Data’s Description and R Code

Attributes’ Description

ThreadID	: Unique ID for each thread (a group of posts on a theme)
AuthorID	: Unique ID for each author (-1 is anonymous)
Date	: Date
Time	: Time
WC	: Word count of the text of the post
Analytic	: LIWC Summary (analytical thinking)
Clout	: LIWC Summary (power, force, impact)
Authentic	: LIWC Summary (using an authentic tone of voice)
Tone	: LIWC Summary (emotional tone)
ppron	: LIWC (all personal pronouns)
i	: LIWC ("I, me, mine" words) First person singular
we	: LIWC ("We, us, our" words) First person plural
you	: LIWC ("You" words) Second person
shehe	: LIWC ("She, he, her, him" words) Third person singular
they	: LIWC ("They" words) Third person plural
number	: Quantities and ranks
affect	: LIWC (expressing sentiment)
posemo	: LIWC (Positive emotions)
negemo	: LIWC (Negative emotions)
anx	: Words indicating anxiety
anger	: Words indicating anger
social	: Words referring to social processes
family	: Words referring to family
friend	: Words referring to friends/friendship
leisure	: Words referring to leisure
money	: Words referring to money
relig	: Words referring to religion
swear	: Swear words
QMark	: Question Mark (Punctuation)

R Code

```
library(ggplot2)
library(lattice)
library(reshape2)
library(lubridate)
library(igraph)
library(igraphdata)

rm(list = ls())
set.seed(30241510)
#importing data
webforum = read.csv("webforum.csv")
webforum = webforum [sample(nrow(webforum), 20000), ] #20000 rows

#tidying data by removing anonymous authorID and 0 word count
webforum = subset(webforum, AuthorID != -1)
webforum = subset(webforum, WC != 0)

#Formating Date to Date since it was previously a string
webforum$Date = as.Date(webforum$Date)

#Formating ThreadID as factor
webforum$ThreadID = as.factor(webforum$ThreadID)

#Creating Month-Year by extracting from Date
webforum$month_year = format(as.Date(webforum$Date), "%Y-%m")

#Getting summary of the data
variables_summary = summary(webforum[,5:29])



######################## Part A ########################
#Part A.1: How active are participants, and are there periods where this increases or decreases?
#Calculating Frequency of post in each month_year
freq_count = as.data.frame(as.table(tapply(webforum$ThreadID,list(webforum$month_year),length)))
colnames(freq_count) = c("Month_Year","Freq") #rename the columns' name

#Plotting the freqcount
timeplot = ggplot(data = freq_count, aes(x=Month_Year,y=Freq, group = 1)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_line(color = "#00AFBB") + ggtitle("Activity in Online Community over Month-Year")

ggsave("PartA_Point1.jpg",timeplot,width = 50, height = 15, units = "cm")

#Part A.2.1: Looking at the linguistic variables, do these change over time?
#Calculating the mean of each variable group by year - only some variables are used based on the correlation
variable_variety = as.data.frame(aggregate(webforum[,c(6,7,8,9,10,17,18,19,22)],list(webforum$month_year),mean))
colnames(variable_variety)[1] = "Month_Year"

#Melting our data to create a graph
variable_variety = melt(variable_variety, id = c("Month_Year"))
variable_variety$value = round(variable_variety$value, digits = 2)
colnames(variable_variety) = c("Month_Year", "Linguistic_Variables", "Value")

#Plotting a multiple lines graph to know the usage of linguistic variables over time
time_variable = ggplot(data = variable_variety, aes(x = Month_Year, y = Value, group = Linguistic_Variables,
color = Linguistic_Variables)) + geom_line(size = 1) + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Average Proportion of Linguistic Variables over Month-Year (2002 to 2011)")

ggsave("PartA_Point2.1a.jpg",time_variable,width = 50, height = 15, units = "cm")



#Part A.2.2: Is there a relationship between them?
#Calculating each variable's correlation into each other by first filtering the linguistic variables
correlation_data = webforum[,6:29]
correlation_data = round(cor(correlation_data),2)
diag(correlation_data) = -1
correlation_data = melt(correlation_data)
cor_heatmap = ggplot(data = correlation_data, aes(x=Var1, y=Var2, fill=value))
cor_heatmap = cor_heatmap + geom_tile(color = "white") + scale_fill_gradient(high = "#8ce2ff", low = "#fff48c")
cor_heatmap = cor_heatmap + theme(axis.text.x = element_text(angle = 90, hjust = 1))
cor_heatmap = cor_heatmap + ggtitle("Correlation Between Linguistic Variables")
cor_heatmap = cor_heatmap + theme(legend.position = "right")

ggsave("PartA_Point2.2.jpg",cor_heatmap,width = 25, height = 17, units = "cm")
```

```
#Proving that linguistic variables do change over times. Variables taken: Analytic, Authentic, and negemo
#Data A would be from 2002, and DataB would be from 2011
webforum$Year = year(webforum$Date)
data_2002 = webforum[webforum$Year == 2002,]
data_2011 = webforum[webforum$Year == 2011,]

#Perform t-test for Analytic
t.test(data_2002$Analytic,data_2011$Analytic, conf.level = 0.95)

#Perform t-test for Authentic
t.test(data_2002$Authentic,data_2011$Authentic, conf.level = 0.95)

#Perform t-test for negemo since it's considered to be stable
t.test(data_2002$negemo,data_2011$negemo, conf.level = 0.95)



######################## Part B ########################
#Part B.1: Describe the threads present in your data.
#We are now analysing the distribution of the top 6 most active threads from 2002 to 2011

#Take top 6 Threads over month-year
top_thread = as.table(by(webforum$Date,webforum$ThreadID,length))
top_thread = sort(top_thread, decreasing = TRUE)
top_thread6 = as.data.frame(head(top_thread,6))
colnames(top_thread6) = c("ThreadID","Freq")
top_thread6 = webforum[(webforum$ThreadID %in% top_thread6$ThreadID),]

#Creating a histograph for each ThreadID
g = ggplot(data = top_thread6) +
  geom_histogram(mapping = aes(x = as.Date(Date), fill = ThreadID), color = "black") +
  ggtitle("6 Top Threads (January 2002 to December 2011)") +
  xlab("Date") + ylab("Number of Post") + facet_wrap(~ ThreadID, nrow = 2)

ggsave("PartB_Point1.jpg", g, width = 20, height = 10, units = "cm")



#Part B.2: By analysing 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?

#We will be using the top 6 threads from previous point to analyse the language used by each group
language_in_thread = as.data.frame(aggregate(top_thread6[,6:29],list(top_thread6$ThreadID),mean))
colnames(language_in_thread)[1] = "ThreadID"

#Since there exist some variables that are dependent to another variable, we will only select:
#Analytics, Clout, Authentic, Tone, ppron, posemo and negemo, as well as words referring to topics
language_in_thread = language_in_thread[,c(-7,-8,-9,-10,-11,-12,-13,-24,-25)]
language_in_thread = melt(language_in_thread, id = "ThreadID")
language_usage_graph = barchart(value~ThreadID,data = language_in_thread,
                                  main=list(label = "Language Used in Top 6 Threads", cex = 2),xlab = "ThreadID",
                                  ylab = "Average Proportion Used by Thread", group = variable,
                                  scales=list(x=list(cex=1.5)), auto.key = list(space = "right", title =
                                                                         "Linguistic\nVariables",cex.title = 1))

language_usage_ggplot = ggplot(language_in_thread, aes(x = ThreadID, y = value, fill =  variable))+
                          geom_bar(position = "dodge", stat = "identity")

#Performing t-test on Analytics and Tone in ThreadID 127115 and 472752
data_127115 = webforum[webforum$ThreadID == 127115,]
data_472752 = webforum[webforum$ThreadID == 472752,]
data_532649 = webforum[webforum$ThreadID == 532649,]

#t-test for Analytics in ThreadID 127115 and 472752
t.test(data_127115$Analytic,data_472752$Analytic, "less", conf.level = 0.95)

#t-test for Tone in ThreadID 127115 and 472752
t.test(data_127115$Tone,data_472752$Tone, "greater", conf.level = 0.95)

#t-test for family in ThreadID 472752 and 532649
t.test(data_532649$family,data_472752$family, "less", conf.level = 0.95)



#Part B.3: Does the language used within threads change over time?
#Calculating the mean of each variable in top 6 threads group by year
language_over_years = as.data.frame(aggregate(top_thread6[,6:29],list(top_thread6$Year),mean))
language_over_years = language_over_years[,c(-7,-8,-9,-10,-11,-12,-13,-24,-25)]
colnames(language_over_years)[1] = "Year"

#Melting our data to create a graph
language_over_years = melt(language_over_years, id = c("Year"))
language_over_years$value = round(language_over_years$value, digits = 2)
colnames(language_over_years) = c("Year","Linguistic_Variables","Value")
```

```
thread_time_variable = ggplot(data = language_over_years, aes(x = Year, y = Linguistic_Variables, fill = Value))
+
  geom_tile(color="white") +
  ggtitle("Average Proportion of Lingustic Variables Used by Top 6 Threads over Years (2002 to 2011)") +
  scale_fill_gradient(low ="#8cf2ff", high = "#ff8cbc")

ggsave("PartB_Point3.jpg",thread_time_variable,width = 22, height = 12, units = "cm")

#t-test for friend to prove that variables referring to words also change.
#Choose data in 2004 to 2007 (Data A)
friend_data_0407 = webforum[as.Date(webforum$Date, "%Y-%m-%d") >= as.Date("2004-01-01","%Y-%m-%d"),]
friend_data_0407 = friend_data_0407[as.Date(friend_data_0407$Date, "%Y-%m-%d") < as.Date("2008-01-01",
                                                                                          "%Y-%m-%d"),]

#Choose data in 2008 to 2011 (Data B)
friend_data_0811 = webforum[as.Date(webforum$Date, "%Y-%m-%d") >= as.Date("2008-01-01","%Y-%m-%d"),]

#doing t-test for friend from (Data A and B)
t.test(friend_data_0407$friend, friend_data_0811$friend, conf.level = 0.95)



######################## Part C ########################
#Part C.1: Social Networks Online
#Choosing a period of time to get top author from that period of time
#Time: December 2005 when  of post is the highest
SNO = webforum[as.Date(webforum$Date,"%Y-%m-%d") >= as.Date("2005-12-01","%Y-%m-%d"), ]
SNO = SNO[as.Date(SNO$Date,"%Y-%m-%d") < as.Date("2006-01-01","%Y-%m-%d"), ]

top_author = as.table(by(SNO$Date, SNO$AuthorID, length))
top_author = sort(top_author, decreasing = TRUE)
top_author10 = as.data.frame(head(top_author,10))
colnames(top_author10) = c("AuthorID", "Freq")
top_author10_data = SNO[(SNO$AuthorID%in%top_author10$AuthorID),]
top_author10_data = top_author10_data[,1:2] #Selecting ThreadID and AuthorID only

#Finding on which ThreadID(s) top 10 authors posts at. This is to create a matrix for social network
for (i in 1:10){
  author = top_author10[i,1]
  author = as.character(author)
  cat("AuthorID: ",author,"\n")
  author_data = top_author10_data[top_author10_data$AuthorID == author,]
  cat("ThreadID(s): ",as.character(unique(author_data$ThreadID)),"\n","\n")
}



#Creating the graph for the social network and Analysing it
author = read.csv("author10_sno.csv", header = TRUE, row.names = 1)
colnames(author) = c(34292,76174,41237,61230,64019,79878,83857,62481,32925,81525)
author_matrix = as.matrix(author)

g1 = graph_from_adjacency_matrix(author_matrix, mode = "undirected")
plot(g1, vertex.color = "red", main = "Connection Among Top 10 Authors in December 2005")

#Calculating the number of vertex and edges in the network
print(vcount(g1)) #10
print(ecount(g1)) #69

#Checking whether the network is simple or not
print(is.simple(g1)) #False

#Calculating the graph's diameter and avg path length
print(diameter(g1))
print(average.path.length((g1)))
hist(degree(g1), breaks = 5, col = "grey")

#Calculatig graph's clique
table(sapply(cliques(g1), length))

#Calculating the degree, closeness, betweenness and eigenvector of each vertex
#and round it to 2 and put it together as a dataframe
deg = as.table(round(degree(g1), digits = 2))
cl = as.table(round(closeness(g1), digits = 2))
bet = as.table(round(betweenness(g1)), digits = 2)
eig = as.table(round(evcent(g1)$vector, digits = 2))

tab = as.data.frame(cbind(deg,cl,bet,eig))
colnames(tab) = c("Degree","Closeness","Betweenness","Eigenvector")
```

```
#Order by Degree
print(tab[order(-tab$Degree),])

#Order by Closeness
print(tab[order(-tab$Closeness),])

#Order by Betweeneess
print(tab[order(-tab$Betweenness),])

#Order by Eigenvector
print(tab[order(-tab$Eigenvector),])

#Looking at the largest clique to determine that author 34292 is the most important author
cliques(g1)[sapply(cliques(g1), length)== 8]

#Checking top 10 most active authors in January 2006
SNO_jan = webforum[as.Date(webforum$Date,"%Y-%m-%d") >= as.Date("2006-01-01","%Y-%m-%d"), ]
SNO_jan = SNO_jan[as.Date(SNO_jan$Date,"%Y-%m-%d") < as.Date("2006-02-01","%Y-%m-%d"), ]

top_author_jan = as.table(by(SNO_jan$Date, SNO_jan$AuthorID, length))
top_author_jan = sort(top_author_jan, decreasing = TRUE)
top_author10_jan = as.data.frame(head(top_author_jan,10))
colnames(top_author10_jan) = c("AuthorID", "Freq")
print(top_author10_jan)
```