Activity, Language, Use and Social Interactions in an Online Community

Dania Ismadi

8/05/2020

Contents

1 Activity and language on the forum over time

This data is read from `webforum.csv` which is a file containing all the posts on an online forum. The thread ID, author ID, date, word count, and a number of language variables measured through proportions are used to describe each post. My personal sample of 20,000 posts contained quite a few anonymous authors. I decided to remove all these authors and their posts from the data for consistency as it would have been inaccurate to use these values for certain calculations. After doing so, I discovered that the data set spanned 10 years, from January 6, 2002 until December 31, 2011. During this time, there were a total of 2395 authors who posted in 600 different threads. The following subsections will go into a further exploration of the data set by analysing the activity of participants and the linguistic variables in the online forum. Note that all graphs or plots mentioned are included in the appendix.

1.1 Analysis of the activity of participants

First, we will examine the relationship between authors, the number of posts and word count. With 95% confidence, we can say that the average number of posts for each authors ranges from approximately 7 to 9 posts. However, the distribution is extremely right skewed as seen in the histogram. Most authors actually only post once and 75% of authors post less than 5 times. We see the same right-skewed distribution with word count. The average number of words for each authors ranges from approximately 701 to 941 words but 75% of authors wrote less than 515 words in total.

A t.test was performed in order to examine if authors who post more have a higher word count on average than authors who post less. Since the t.test requires an approximately normally distributed data and our data is extremely right-skewed, we have to modify the data a little bit. First, we will only include data among active users (those who post more than a certain amount) and exclude those who post a significant amount more than everybody else. A log transformation is then performed which results in an approximately normal distribution as shown in the histogram. We define authors who post the most as those who are in the top 10 percentile of this data and authors who post the least as those who are in the bottom 10 percentile of the this data. After performing a t.test, we find that there is no evidence that is strong enough to suggest that those who have post more have a higher word count on average than those who post the least. Thus, for the rest of the report, the most active authors are defined as those with the most posts and the least active authors as those with the least amount of posts. This is because from the t.test, we see that there is no reason to suggest that authors with high word counts interact with the forum more through posting than those that do not.

We can say with 95% confidence that the average number of threads authors participate in is between 4 and 5 threads but this is not representative of the population because the distribution is extremely right skewed. Most authors only participate in 1 thread. However, from high the t-statistic and low p-value obtained from performing a t.test, there is strong evidence to suggest that the most active authors participate in more threads on average than the authors who are least active.

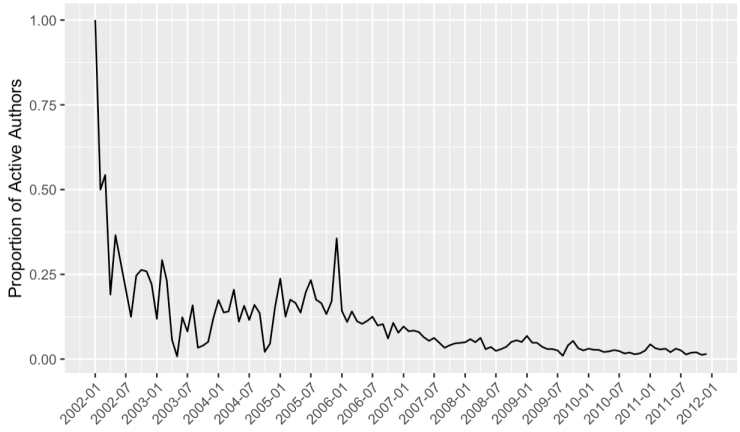

Figure 1: Time Series of the Proportion of Active Authors Over Time

Figure 1 is a time series of the proportion of active authors out of all the authors who ever interacted with the forum up until that time. We notice a generally decreasing trend, suggesting that most authors do not remain active on the thread for long. Most authors are usually only active for one day. With 95% confidence, we can say that authors are active for 209 to 247 days on average but the distribution is right skewed as 75% of authors are active for 197 days or less. Interesting enough, we cannot conclude the more active authors stay active in the forum for a longer time on average than less active authors.

We define an author's responsiveness as the average duration between their subsequent posts. We can say that with 95% confidence, authors take 81 to 101 days of break on average between subsequent postings. So since we could not conclude that more active authors stay active on the forum for long, then are they more responsive? This seems like a reasonable conclusion to make but from performing a t.test, the evidence is not strong enough to suggest that this is the case.

We know that most authors are not active on the forum for long but to get a high level view of the overall activity of participants over time, a time series graph of the number of posts posted per month onto the forum can be plotted (shown in figure 2). I have chosen to plot per month instead of per day as plotting posts per day leads to a messy graph and a trend would be difficult to observe. Furthermore, plotting per year is too low of a resolution as it would only yield us 10 data points. Therefore, plotting per month is the best option.

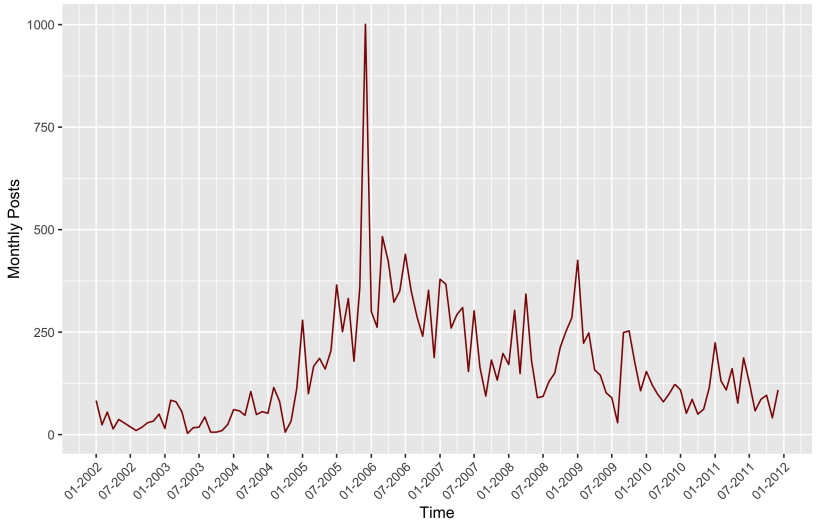

Figure 2: Number of Monthly Posts Over Time

From analysing the time series in figure 2, we can see that the number of posts start of low and gradually increase until it reaches a sharp spike in 2005. More specifically, it reaches an all time high of 1001 posts in October 2005. After, the number of posts begin to decrease. We also notice that 2005 is an outlier as the number of posts stay below 500 posts in other years. This trend is observed in the graph of the decomposition of this time series. Thus, it is clear that the activity of participants is not constant. However, excluding 2005, the change in the level activity is quite gradual.

1.2 Analysis of linguistic variables

The highest proportion of language variables present in posts belong to analytic, clout, authenticity and tone. The most significant out of those four variables is analytic, referring to words that express analytical thinking. This suggests that the forum might be geared more towards professional, academic or thought-provoking conversations rather than shallow, casual exchanges. A heat map of the proportion of language variables used over time in the forum is shown in figure 3. This heat map includes all threads in the web forum. Furthermore, I chose to plot the average proportion by year because plotting by month or day yields a very messy visualization as some months and days have no posts.

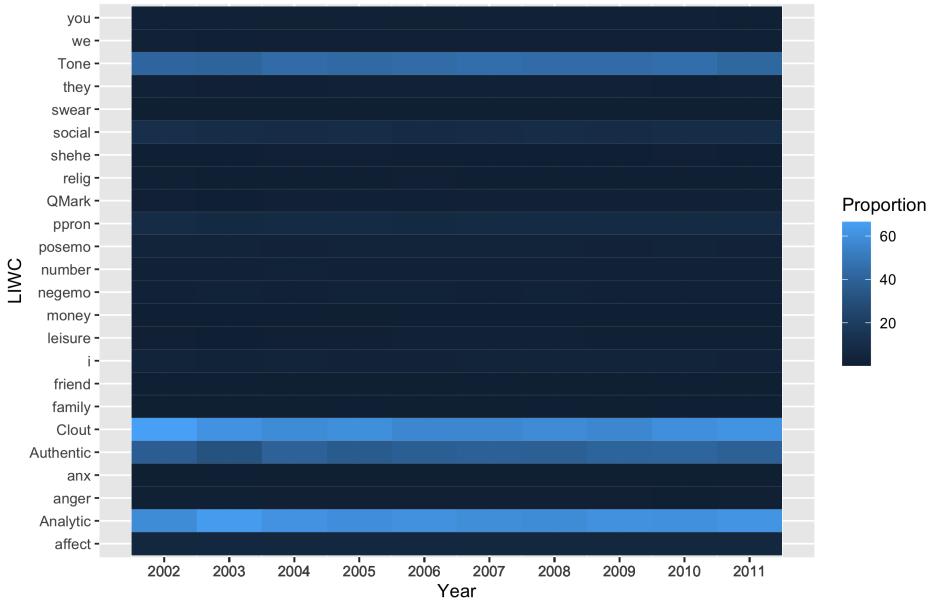

Figure 3: Heat Map of the Language Variables Proportions Over Time

From figure 3, we are able to clearly see the difference in proportions. Analytic, clout, authenticity, tone are significantly higher in proportion than the and other language variables. We also notice that the proportion of language variables generally stay consistent throughout the years. Again, this might suggest that the forum is quite niche regarding the types of conversations that is being discussed.

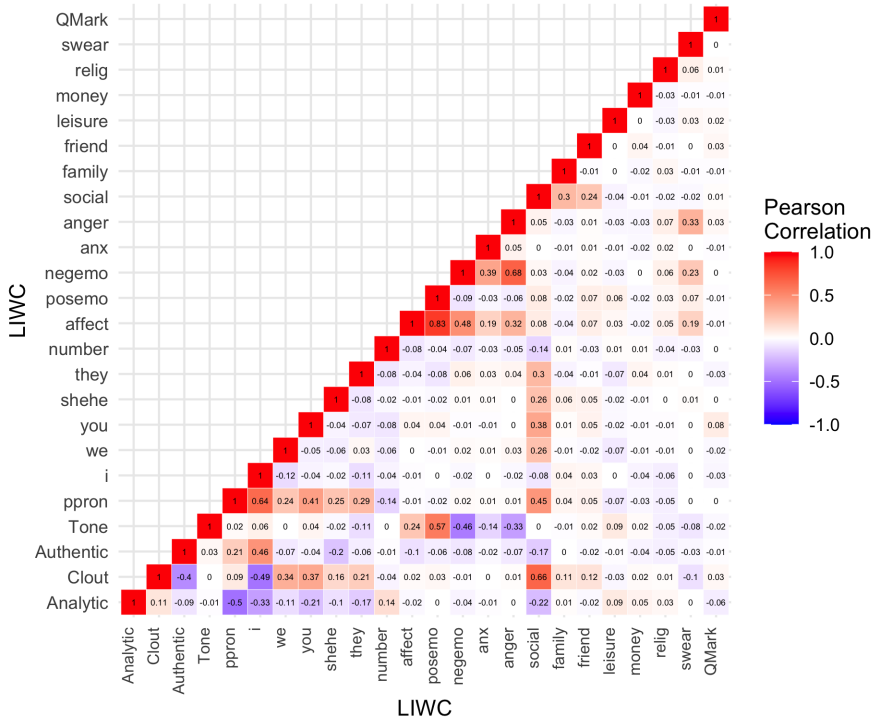

Figure 4: Correlation Matrix between Language Variables

Figure 4 shows a correlation matrix of the language variables in the form of a heat map. We can observe that positive emotion and affect (words that express sentiment) are highly correlated with a value of 0.83. This suggests

that most posts that express an opinion or view in the forum are generally quite positive as well. Some other significant correlations to note are between anger and negative emotion (0.68) and between social and clout (0.66). Anger and negative emotion is quite self explanatory. As for social and clout, it appears that in this forum, authors express more power and force as words refer to social processes increase.

2 Language used by different groups

There are 600 different threads in this web forum. This section will go into a further exploration into the language used in these threads.

2.1 Description of threads

The distribution for the number of posts in each thread is right skewed as seen in the histogram. Most threads have 20 posts, and although the average number of posts is 32, 75% of threads have less than 34 posts. In order to satisfy the condition that the data has to be approximately normally distributed for a t.test, I excluded threads that are not active and threads that have significantly more posts than the rest. A log transformation is then performed resulting in approximately normally distributed data as seen in the histogram. More active threads are defined as the threads that belong in the top 10 percentile in this data and the less active threads are defined at the threads that belong in the bottom 10 percentile. From performing a t.test, there is strong evidence to suggest that threads that have more posts are higher in word count on average than threads with the less posts. Thus, we know that threads in this forum are not usually spammed with just one-line posts like how a digital messaging thread (i.e., Whatsapp, iMessage, Messenger) between people might look.

On average, there are 18 authors (17 to 19 authors with 95% confidence) per thread though most threads have 16 authors. The distribution is right skewed which means there are very few threads with large amounts of authors involved, however the maximum is 145 authors. There is strong evidence to conclude that more active threads have more contributing authors on average than less active threads.

Most threads are active for 3 days only, and 75% of threads are active for less than 112 days. However, the few threads that are active for longer are active for an extremely long time- the maximum is 2767 days. This pulls the average for the duration of active threads up to 178 days. There is strong evidence to suggest that more active threads are on average active for longer than less active threads. However, there is no evidence to suggest that more active threads have more frequent postings on average than less active threads. Thus, this suggests that for more active threads, the number of posts is less likely due to spam but due to actual involvement from the community. Most threads are updated with new posts every day and for 75% of posts the average break between posts is less than 10 days.

2.2 Difference in language between groups

To see if there is a difference in language variables used in different groups, I decided to compare between the most active and least active threads as perhaps, we might be able to observe some kind of pattern where certain threads that express a certain sentiment are more popular than others. A visualisation of this as a bar chart is shown in figure 5. At a glance, we can see that threads with more posts are more analytic, authentic, and revolve more around views and opinions (affect) while threads with the least posts express more force (clout), express more emotion (tone) and refer to social processes. However, to get a more accurate conclusion of the difference in language variables, t.tests were conducted for each variable. From the tests, we can attempt to make conclusions whether more popular threads express more of a certain sentiment on average than less popular threads. Some significant conclusions observed are that more popular threads express more negative emotion and anger on average than least popular threads. This suggests that perhaps it is more likely you will find posts that express strong negative opinion about a certain topic in more popular threads on average than less popular threads. Meanwhile, less popular threads express more force, power, emotion and words that refer to social processes (clout, tone and social). Referring to our previous section, we concluded that clout and social are highly correlated and now, we see that these variables appear more in less popular threads. Nevertheless, from the bar chart, it is clear that in both groups, the most popular language variables are consistent- analytic, clout, authentic and tone, as we also previously concluded.

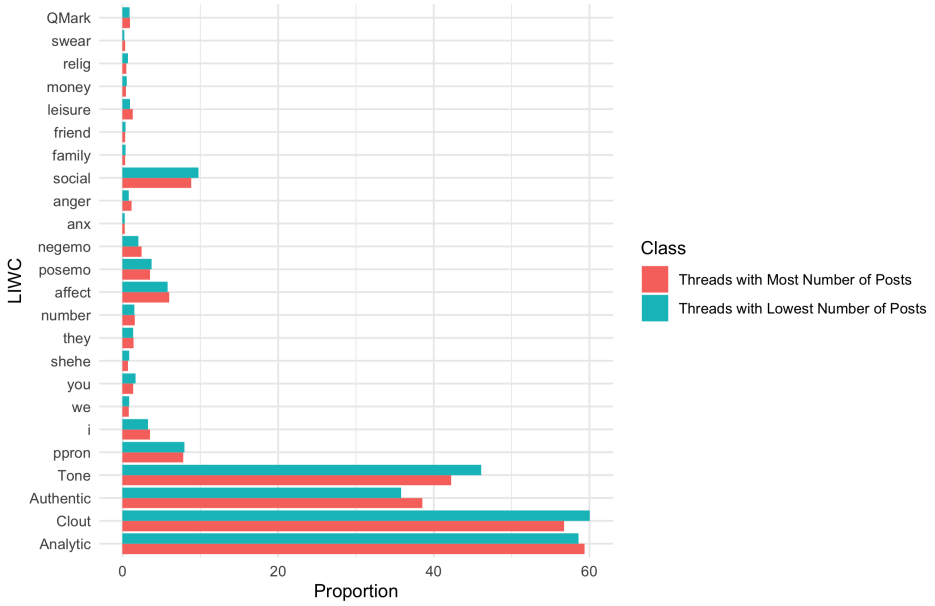

Figure 5: Bar Chart Comparing LIWC Between Most Active and Least Active Threads

2.3 Language in threads over time

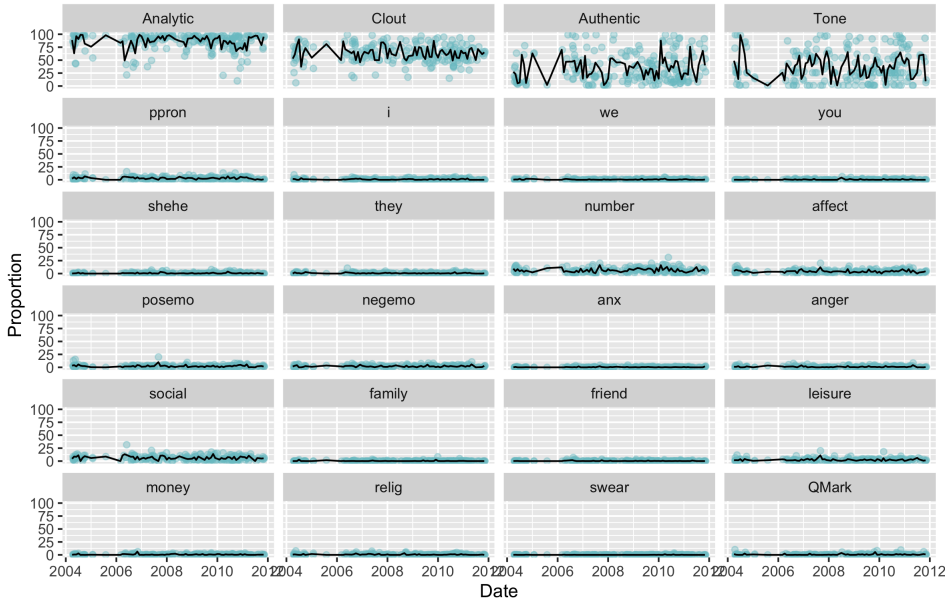

Figure 6: Comparing Language Variables for Thread 127115

To explore the question whether or not language in threads change over time, we will explore the top 6 threads with the most number of posts (Threads 252620, 127115, 472752, 145223, 283958, and 532649). I chose to do this because perhaps it increases the chance for a larger variety of different views, thoughts and opinions to be expressed so we might be able to see more observable changes than viewing a thread with less posts. To do this, I graphed a time series of the average proportion of each language variable used by month. I also mapped the average proportion of each language variable per day shown as the blue points in the graph. This makes it easy to observe how the data is distributed. The graph shown for the thread with thread ID 1277115, the thread with the second most number of posts, is shown in figure 6.

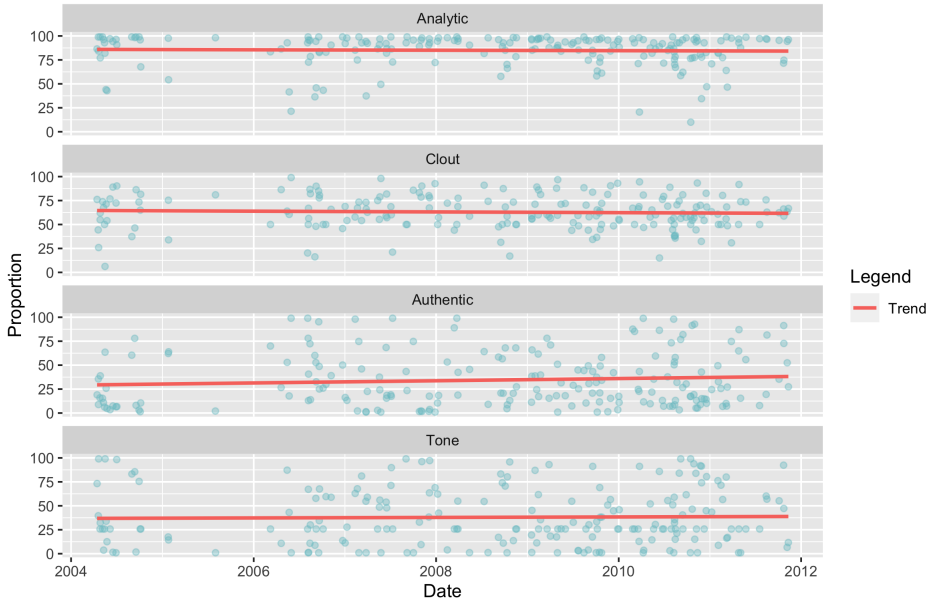

Figure 7: Comparing Language Variables for Thread 127115 with Trend Line

We see that in figure 6, most of the language variables actually stay constant (i.e., family, friend, leisure, QMark and may more) while variables like analytical thinking, clout, authenticity and tone are highly fluctuating showing that language for this thread does change from time to time. However, does this difference persist? The fluctuation makes it very difficult to see a trend over time so I plotted a trend line over the distribution to see if the difference does persist, focusing on the top 4 variables (analytic, clout, authentic and tone). This is shown in in figure 7. We see that the trend line is almost horizontal for all 4 variables. Thus, although these language variables fluctuate a lot, the difference does not persist over time. This suggests that for this thread, the sentiment expressed more or less remains the same during its life. From our sample, we see a similar trend for threads where authors post consistently over time- i.e., in threads 472752 and 532649. All graphs for each thread are included in the appendix in section 4.

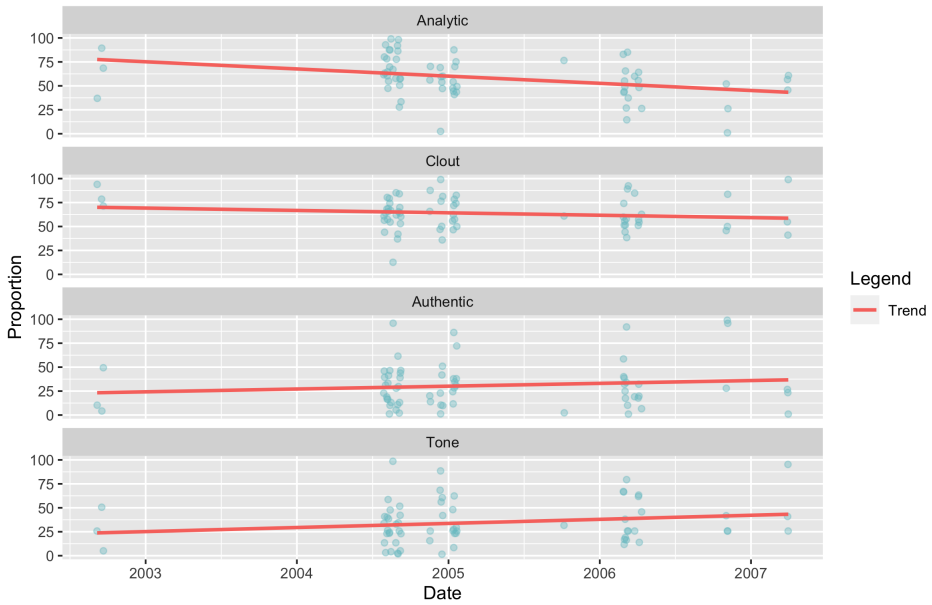

Figure 8: Comparing Language Variables for Thread 145223 with Trend Line

On the other hand, when we analyse threads that have large gap of time between postings, we see a slight difference. An example is shown for thread 145223 in figure 8. We see that for this particular thread, whilst analytical thinking and clout observed a decrease, authenticity and emotional tone saw an increase. Similar trends are observed with thread 252620 and 283958 where we see a slight increase or decrease in language variables. Thus, we can see that from this sample, it is more likely for a difference in language to persist in threads where there are less frequent postings. When threads observe more frequent postings (the duration between subsequent posts are short), the difference in language used in threads rarely persist over time though they do fluctuate. This suggests that for these threads more or less express the same sentiment for the duration that they are active.

3 Social networks online

Participants who communicate on the same network at the same time is a social network. For this analysis, I chose to examine the social network from March, April and May 2005 as this was when we started to see an increase in the number of people posting on the forum.

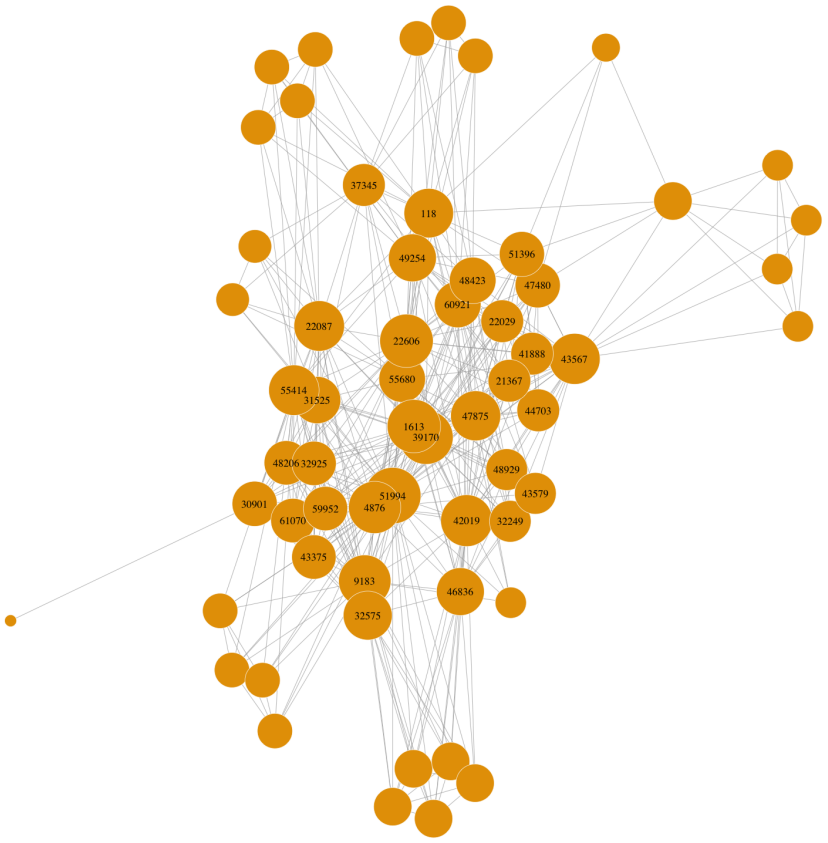

Figure 9: Social Network for March 2005

In the month of March, we see that there are 7 active threads resulting in 7 complete graphs. In these threads, 61 authors participated resulting in 440 connections. The largest clique size during this month is 16 which represents the thread with the most authors participating during March. During this month, all authors are connected to at least one another person so the result is a connected graph. This graph is shown in figure 9. More important authors are represented as bigger vertices while less important authors are smaller. From the histogram of the number of degrees against frequency, we see that the distribution is log normal which is typical of a social network.

There are lots of players with few connections and less players with more connections. The furthest connection between the two most distant people in this graph is 4 and the average path length is 2, thus this graph is very well connected. When observing the number of degrees, closeness centrality measure, betweeness centrality measure and eigencentrality, we notice that author 51994 has the top measurements. This author also belongs in the largest clique size. This suggests that this person is a strong local influencer and act as a bridge between different authors as well. In figure 9, this author is placed in the middle and their vertex is reasonably sized which emphasizes his important. Based on the high eigencentrality measurement, we can say that this person also has high quality connections as they are connected to other important authors.

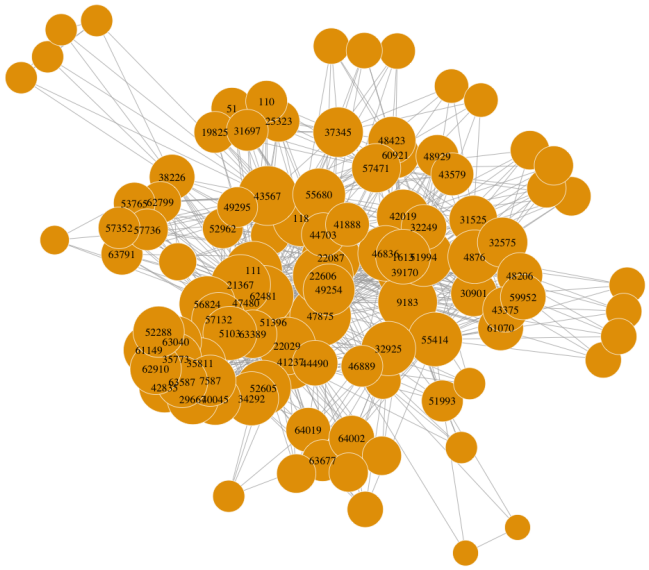

Figure 10: Social Network for March and April 2005

To see how this social network expands over time, we combine this graph with with the graph in April. The resulting graph is shown in figure 10. We see that in this graph, the number of authors increase to 106 and the number of connections to 1011. The average path length stays the same but the furthest connection between the two most distant people decreased to 3 steps. This means that the authors is slightly more connected than before as perhaps some authors posted in many different threads with different authors. When observing the statistics, we see that author 47875 is high in the number of degrees and closeness centrality values. This means that this author has a lot of connections and is very well connected locally. Meanwhile author 435687 is high in betweeness centrality measure meaning that they act as a hub between other authors to connect. Finally, author 22606 has more quality connections due to its high eigencentrality value. However it is important to note that author 47875 follows a close second in these measures as well. This suggests that author 47875 is extremely well connected in this network. We see that this is because author 47875 is active in both March and April, and thus, was able to expand their own personal network. Whilst this person was not one of the most prominent figures in March, the statistics suggest that he was very important in April.

When extending our social network further to May, as expected, we see a large increase in the number of authors and connections to 133 and 1389 respectively. However, the average path length and diameter remains the same. Thus, it does not imply that the authors are significantly more well connected now than before. This social network is represented in figure 11. From analysing the statistics, we see that all authors that are high in the number of degrees, betweeness centrality, closeness centrality and eigencentrality measures have been active in all 3 months. This makes sense as there is a higher chance for them to interact with different authors if they are active throughout.

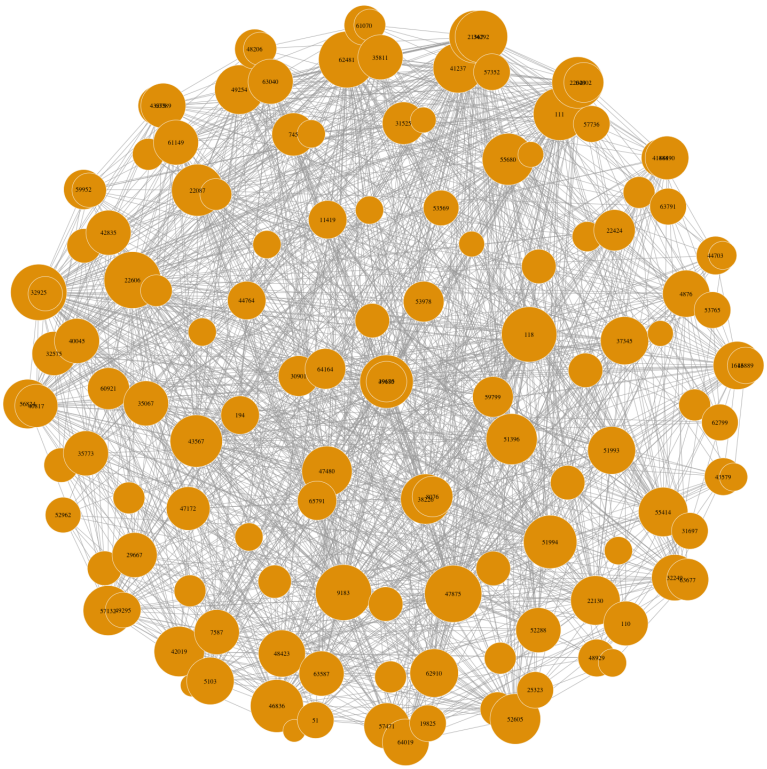

Figure 11: Social Network for March, April and May 2005

We also see that author 47875 remains an important player. However, we also notice that author 118 is also a top player during this time. Author 118 is the only author that appears as one of the top 6 players in each individual month and also when we built the network for March and April combined. This suggests that author 118 has consistently been a top player throughout March and April.

4 Conclusion

To summarise, activity from participants in this forum changes throughout. There are some periods where the forum experience increased activity and times where it does not but overall, the forum seems to have reached its peak in 2005 and has been decreasing in activity ever since. However, language, when measured by language variables, stays quite consistent in the forum throughout the years with the greatest proportions hailing from words that express analytical thinking, power and impact (clout), authenticity and emotional tone. There are some slight differences in the language used between different groups of threads but we generally see the same trend where language remains consistent through the duration of activity on the thread, suggesting that threads do not stray far from the sentiment they were originally expressing. Finally, from analysing the period from March 2005 to May 2005, we see typical attributes of a social network in this forum such as the fact that there are few people with very many connections while the majority have fewer connections. We were also able to point out a few key players during this time.

5 Appendix

5.1 Tidying Data

Importing the libraries:

```
library(readr)
library(ggplot2)
library(dplyr)
library(tidyverse)
library("tidyr")
library(reshape2)
```

Before running each code block, here is how I set up my data:

```
# clear working environment
rm(list=ls())

# create individual data
set.seed(30990971)
webforum <- read.csv("webforum.csv")
# create a sample of 20,000 rows
webforum <- webforum[sample(nrow(webforum), 20000), ]
# remove all anonymous authors.
webforum = webforum[webforum$AuthorID!=-1,]
# change Date column to date data type
webforum$Date = as.Date(webforum$Date, "%Y-%m-%d")
```

5.2 Additional Tables and Graphs

5.2.1 Part 1

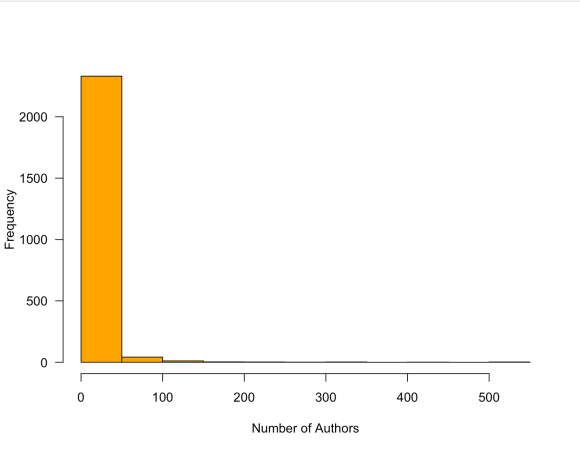

Figure 1: Histogram of the Number of Authors and the Total Number of Posts

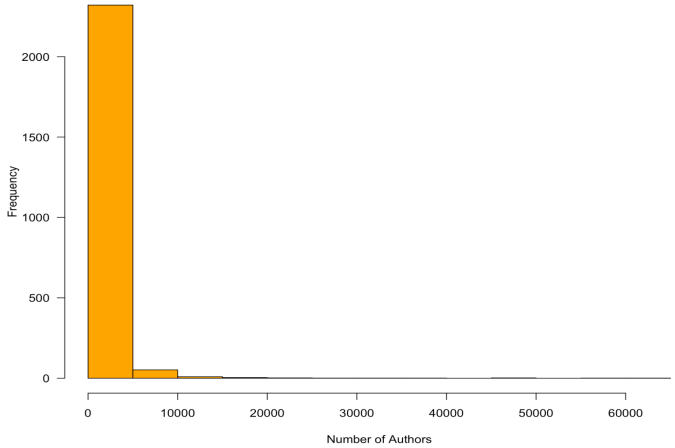

Figure 2: Histogram of the Number of Authors and the Total Word Count

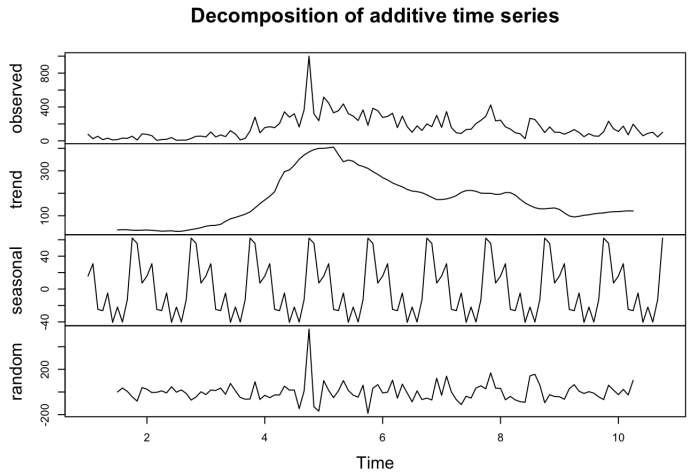

Figure 3: Decomposition of the Number of Monthly Posts Over Time

5.2.2 Part 2

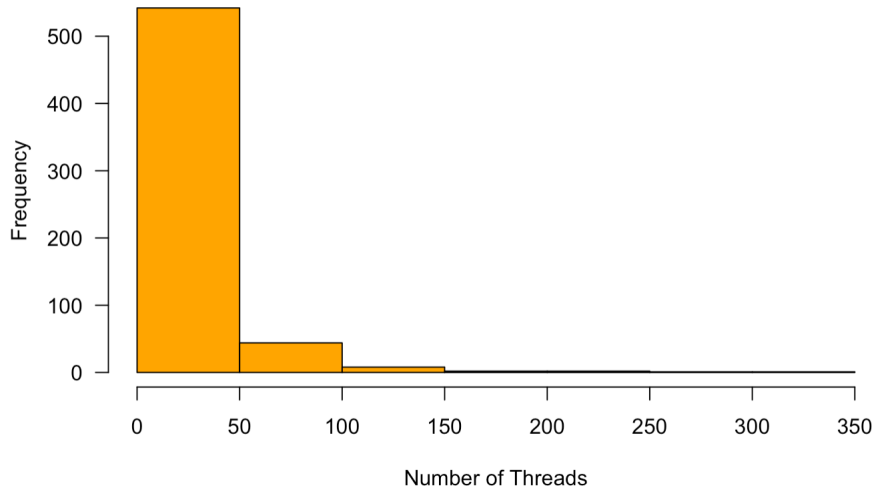

Figure 4: Histogram of the Number of Threads and Number of Posts

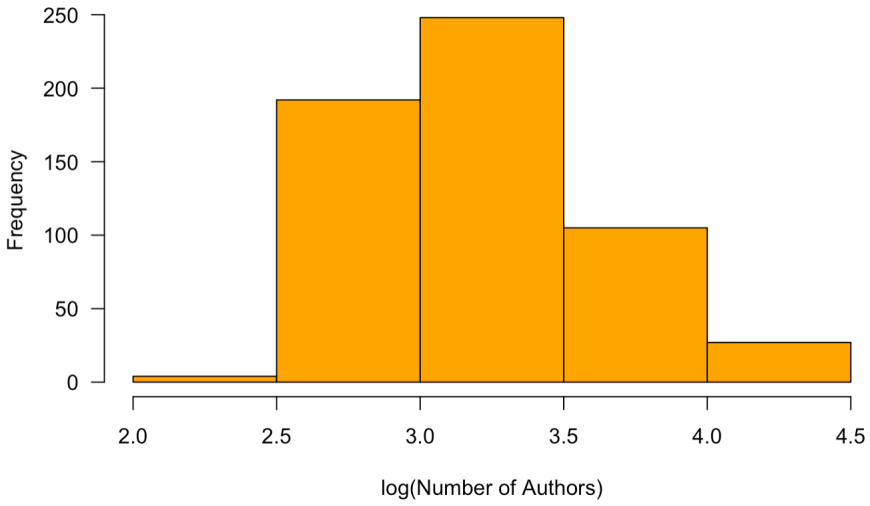

Figure 5: Histogram of the log(Number of Threads) and log(Number of Posts)

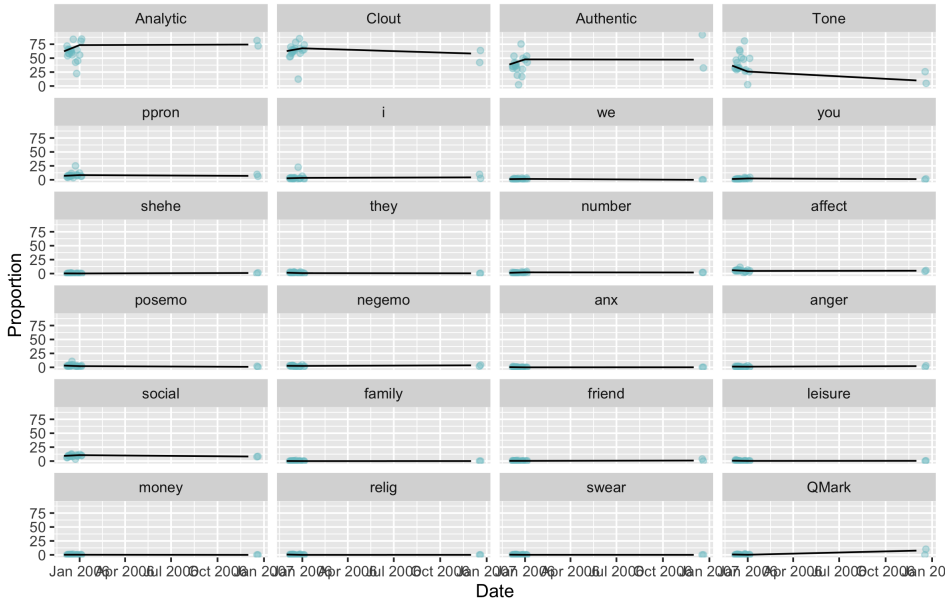

Figure 6: Comparing Language Variables for Thread 252620

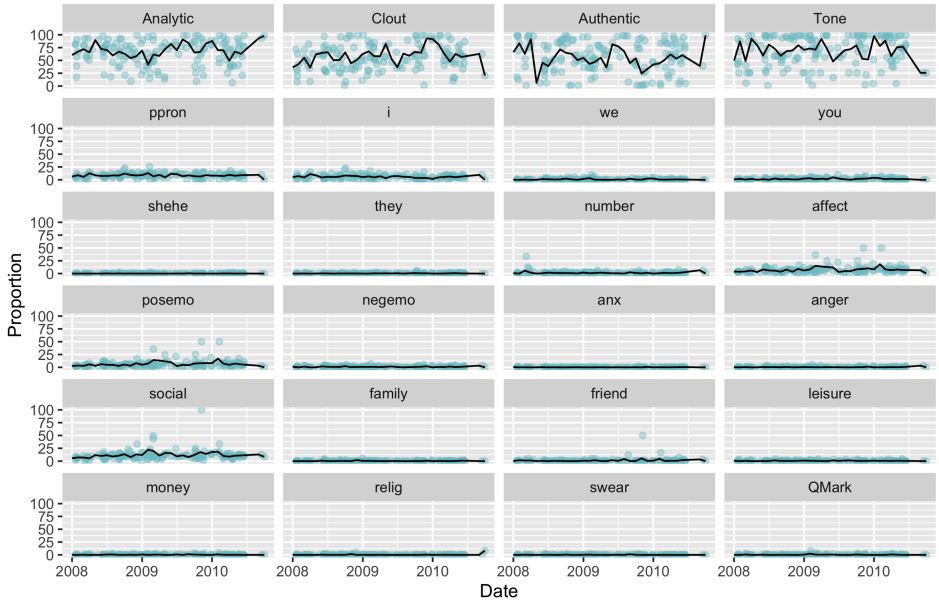

Figure 7: Comparing Language Variables for Thread 472752

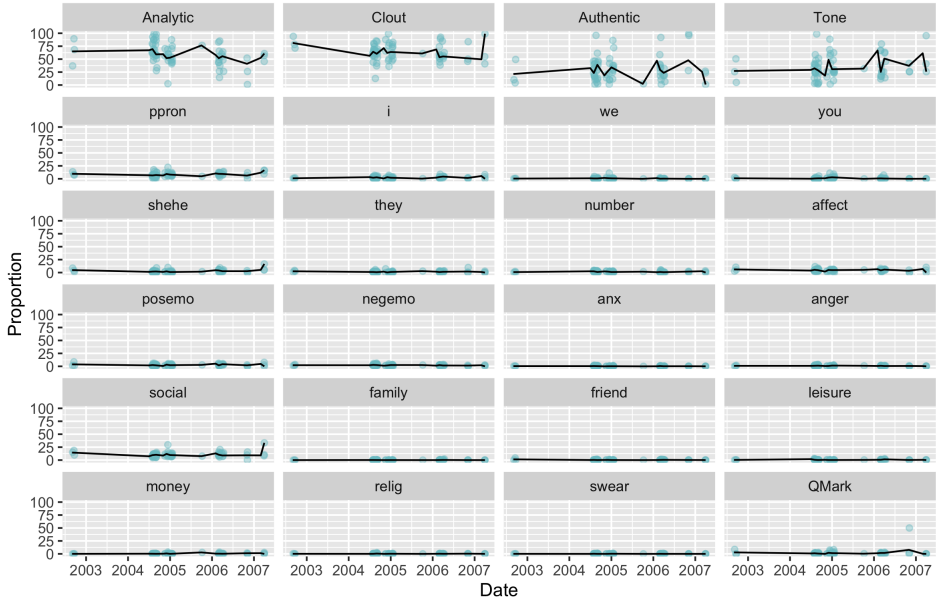

Figure 8: Comparing Language Variables for Thread 145223

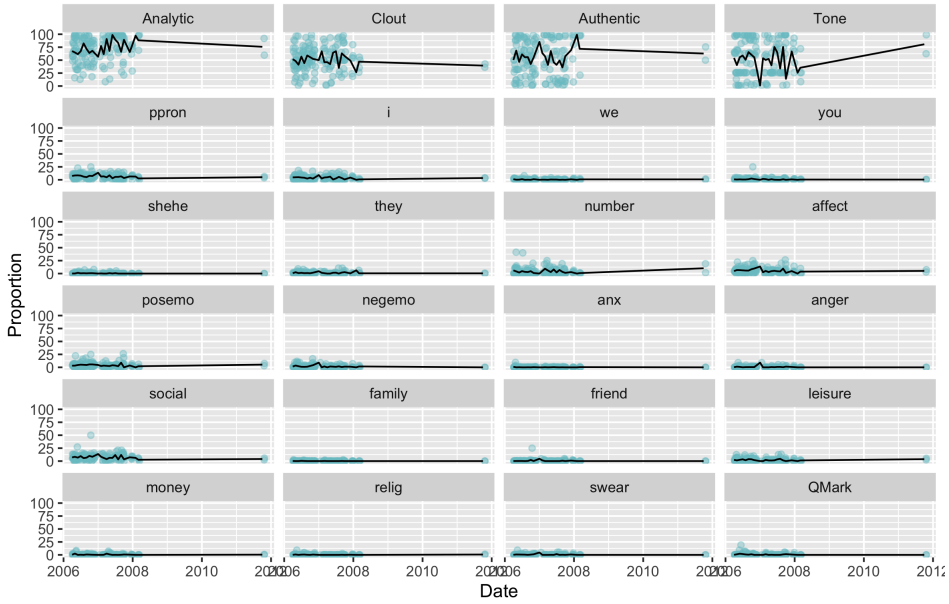

Figure 9: Comparing Language Variables for Thread 283958

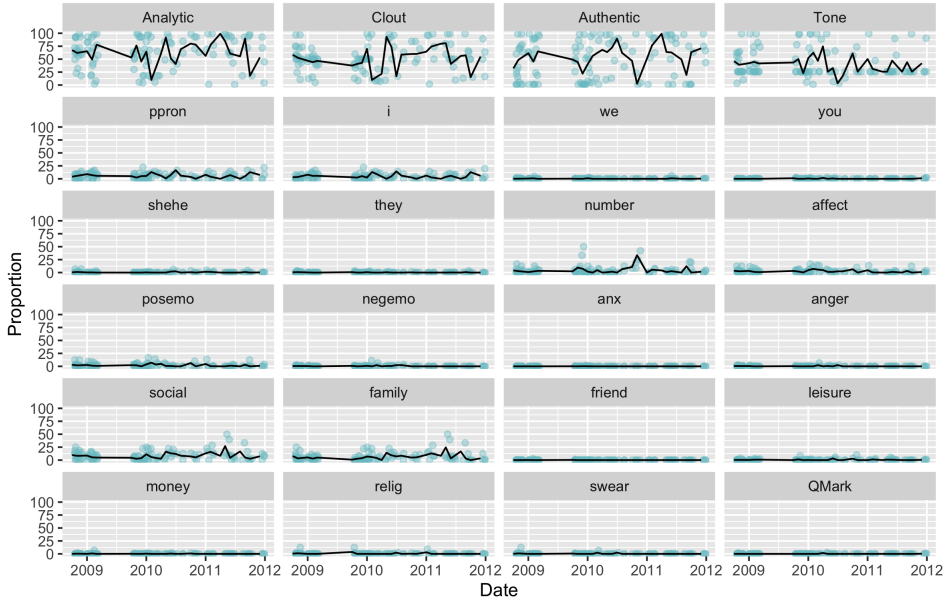

Figure 10: Comparing Language Variables for Thread 532649

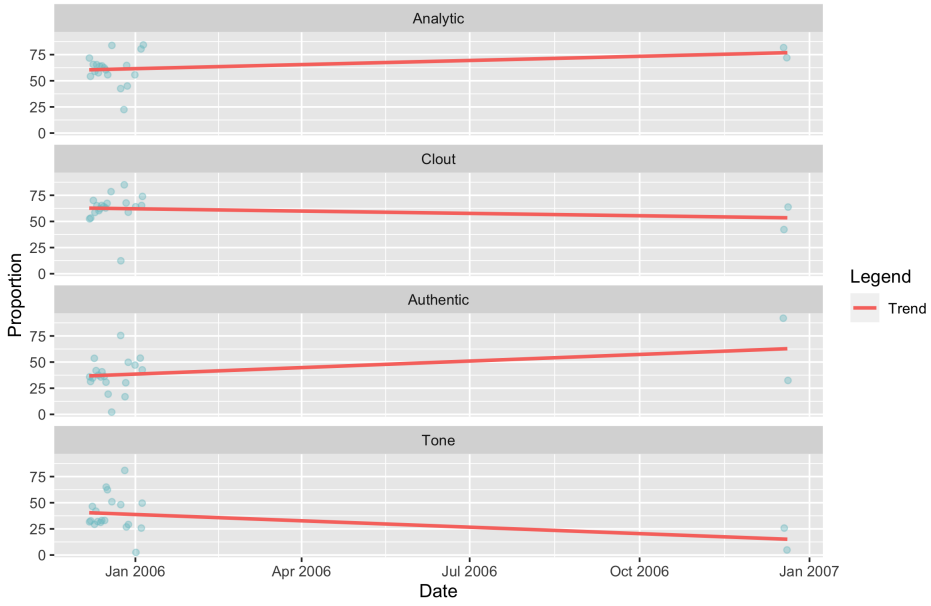

Figure 11: Comparing Language Variables for Thread 252620 with Trend Line

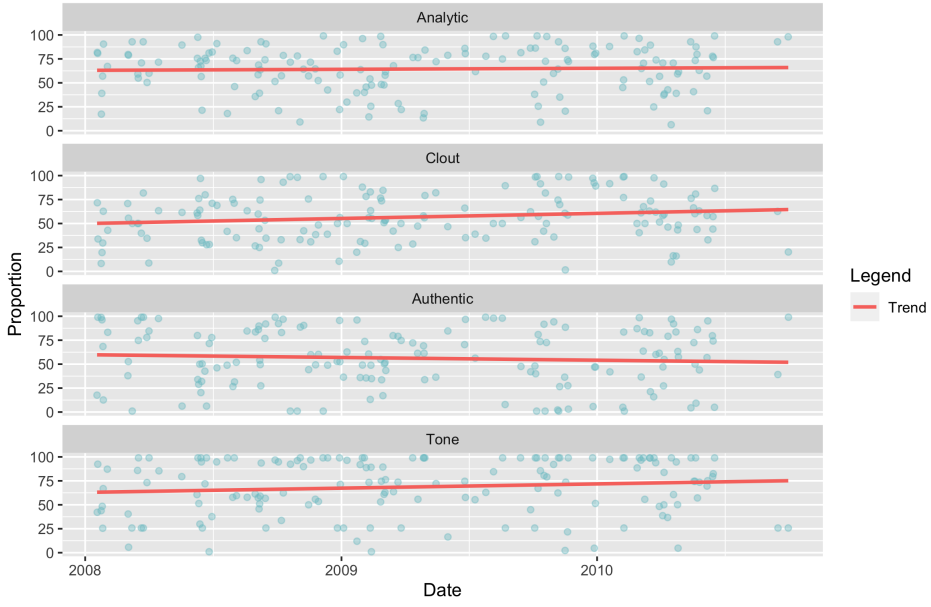

Figure 12: Comparing Language Variables for Thread 472752 with Trend Line

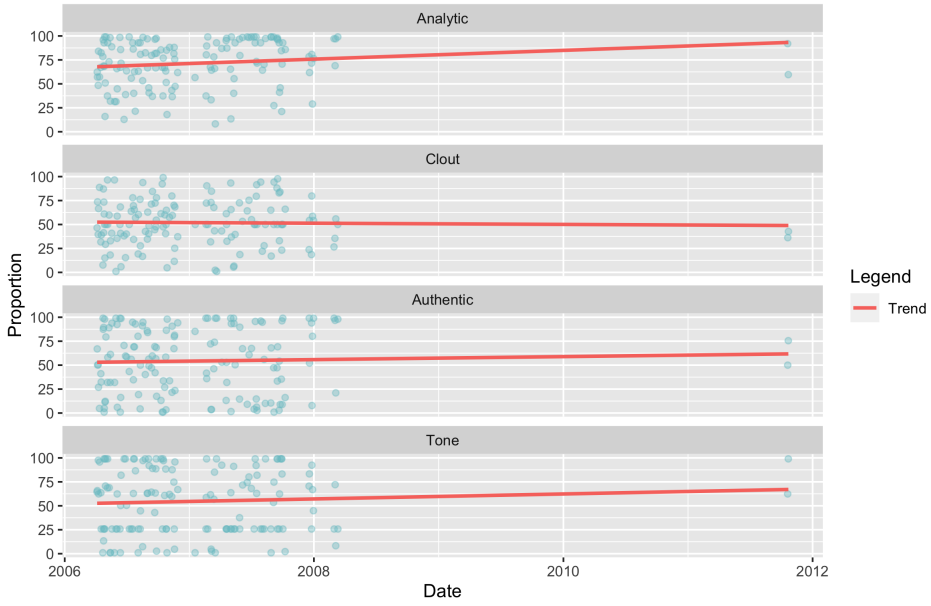

Figure 13: Comparing Language Variables for Thread 283958 with Trend Line

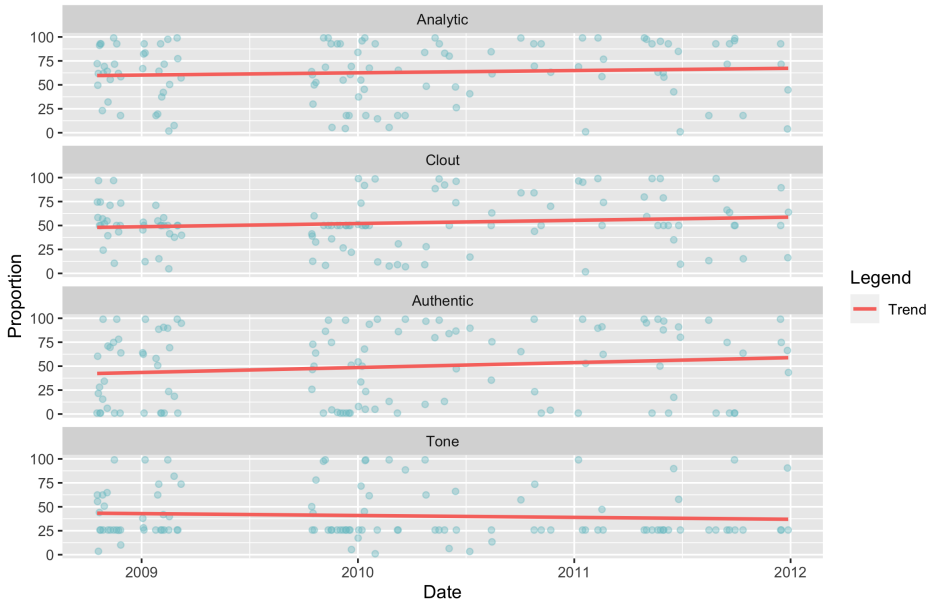

Figure 14: Comparing Language Variables for Thread 532649 with Trend Line

5.3 R Code

5.3.1 Part 1

Finding out the date range of the entire data:

```
# Find the minimum date.
min.date = webforum[which.min(as.Date(webforum$Date, "%Y-%m-%d")),]

# Find the maximum date.
max.date = webforum[which.max(as.Date(webforum$Date, "%Y-%m-%d")),]

min.date
max.date
```

Counting the number of threads:

```
# Calculate the number of threads and how often they appear in the data set.
num.threads = by(webforum, list(webforum$ThreadID), function(df) nrow(df))

# Combine output from by function into dataframe.
num.threads = as.data.frame(as.table(num.threads))
# Rename columns
colnames(num.threads) = c("ThreadID", "Count")

# Organise num.threads from least count to largest count.
num.threads = num.threads[order(num.threads$Count),]

dim(num.threads)
head(num.threads)
tail(num.threads)
```

Finding out how many authors are involved.

```
# Calculate the number of authors and how often they appear in the dataset.
num.authors <- by(webforum, list(webforum$AuthorID), function(df) nrow(df))

# Combine output from by function into dataframe.
num.authors <- as.data.frame(as.table(num.authors))
# Rename columns
colnames(num.authors) <- c("AuthorID", "Count")

# Orgnaise num.authors from least count to largest count.
num.authors <- num.authors[order(num.authors$Count),]

head(num.authors, 10)
tail(num.authors, 30)
dim(num.authors)
```

Statistics for the relationship between authors and the number of posts:

```
# Using data from before. Find statistics and the distribution of the number of posts
    per author.
summary(num.authors$Count)
getmode(num.authors$Count)
hist(num.authors$Count, main="", xlab="Number of Authors", border="black", col="orange",
     las=1, breaks=10)
# confidence interval
t.test(num.authors$Count)
```

Statistics for the relationship between authors and the total word count:

```
wc.authors = by(webforum$WC, list(webforum$AuthorID), sum)
```

```
wc.authors = as.data.frame(as.table(wc.authors))
colnames(wc.authors) = c("AuthorID", "TotalWC")
wc.authors = wc.authors[order(wc.authors$TotalWC),]


head(wc.authors)
tail(wc.authors)
summary(wc.authors$TotalWC)
getmode(wc.authors$TotalWC)
hist(wc.authors$TotalWC, main="", xlab="Number of Authors", border="black", col="orange"
    , las=1, breaks=10)
# confidence interval
t.test(wc.authors$TotalWC)
```

Fixing the right-skew distribution:

```
num.authors
dim(num.authors[num.authors$Count > 80,])
fix.skew = num.authors[num.authors$Count >= 60,]
fix.skew = fix.skew[fix.skew$Count < 110,]
hist(fix.skew$Count, main="", xlab="log(Number of Authors)", border="black", col="orange
    ", las=1, breaks=10)
hist(log(fix.skew$Count), main="", xlab="log(Number of Authors)", border="black", col="
    orange", las=1, breaks=5)
```

Finding the most active and least active authors, and performing t.tests against word count:

```
x = as.numeric(quantile(log(fix.skew$Count), prob=0.90))
y = as.numeric(quantile(log(fix.skew$Count), prob=0.10))
fix.skew$Count = log(fix.skew$Count)
fix.skew

# Define authors who post more as those who post in the top 10%.
most.posts = fix.skew[fix.skew$Count >= quantile(fix.skew$Count, prob=0.90),]
# Define authors who post less as those who post bottom 10%.
least.posts = fix.skew[fix.skew$Count <= quantile(fix.skew$Count, prob=0.10),]

most.posts.wc = wc.authors[wc.authors$AuthorID %in% most.posts$AuthorID,]
least.posts.wc = wc.authors[wc.authors$AuthorID %in% least.posts$AuthorID,]

# Perform t.test to conclude significance.
t.test(most.posts.wc$TotalWC, least.posts.wc$TotalWC, alternative="greater")

more.active = most.posts
less.active = least.posts
```

Statistics for the number of threads authors participate in:

```
webforum.temp = webforum

# Group by AuthorID
num.threads = as.data.frame(as.table(by(webforum.temp$WC, list(webforum.temp$ThreadID,
    webforum.temp$AuthorID), FUN=length)))
num.threads = na.omit(num.threads)
colnames(num.threads) = c("ThreadID", "AuthorID", "NumPosts")

sum.threads = num.threads %>% group_by(AuthorID) %>%
  arrange(ThreadID) %>%
  summarize(NumThreads = as.numeric(length(ThreadID)))

summary(sum.threads$NumThreads)
getmode(sum.threads$NumThreads)
# histograms
```

```
hist(sum.threads$NumThreads, main="", xlab="Days", border="black", col="orange", las=1,
    breaks=10)

# confidence intervals
t.test(sum.threads$NumThreads)
```

T.test comparing average thread count between more active and less active authors.

```
less.active.threads = sum.threads[(sum.threads$AuthorID %in% less.active$AuthorID),]
more.active.threads = sum.threads[(sum.threads$AuthorID %in% more.active$AuthorID),]

t.test(more.active.threads$NumThreads, less.active.threads$NumThreads, alternative="
    greater")
```

Statistics for the proportion of authors that remain active.

```
# Analyse activity by month.
webforum.temp = webforum
webforum.temp$Date = as.Date(webforum.temp$Date, "%Y-%m-%d")
webforum.temp$MonthYear = as.Date(format(webforum.temp$Date, "%Y-%m-01"))

# Group AuthorID by month to find total number of Authors that post per month.
new.authors = as.data.frame(as.table(by(webforum.temp$WC, list(webforum.temp$MonthYear,
    webforum.temp$AuthorID), FUN=sum)))
new.authors = na.omit(new.authors)
new.authors = new.authors[,-3]

# Rename Column Names.
colnames(new.authors) = c("Date", "AuthorID")

# Find total number of authors that post per month.
total.authors = as.data.frame(as.table(by(new.authors$AuthorID, list(new.authors$Date),
    FUN=length)))
colnames(total.authors) = c("MonthYear", "TotalAuthors")
# Change MonthYear Column to Date data type.
total.authors$MonthYear = as.Date(total.authors$MonthYear, "%Y-%m-%d")

# Plot total number of authors that post per month.
g = ggplot(total.authors, aes(MonthYear, TotalAuthors)) + geom_line() + xlab("") + ylab(
    "Number of Authors")
g = g + scale_x_date(date_labels = "%Y-%m", date_breaks = "6 month")
g = g + theme(axis.text.x = element_text(angle = 45, hjust = 1))
g

# remove duplicated rows and only preserve the first row which is the first time Author
    appeared in dataset to find only new authors who post per month.
new.authors = new.authors %>% distinct(AuthorID, .keep_all=TRUE)

# sort data frame by date
new.authors = new.authors[order(new.authors$Date),]

# find number of new authors permonth.
new.authors = as.data.frame(as.table(by(new.authors$AuthorID, list(new.authors$Date),
    FUN=length)))
# rename column names
new.authors[is.na(new.authors)] = 0
colnames(new.authors) = c("MonthYear", "NewAuthors")
# change MonthYear column to date data type
new.authors$MonthYear = as.Date(new.authors$MonthYear, "%Y-%m-%d")

# plot total number of new authors that post per month.
t = ggplot(new.authors, aes(MonthYear, NewAuthors)) + geom_line() + xlab("") + ylab("New
     Authors")
```

```
t = t + scale_x_date(date_labels = "%Y-%m", date_breaks = "6 month")
t = t + theme(axis.text.x = element_text(angle = 45, hjust = 1))
t

# find the proportion of the total number of authors that post per month out of the
    total number of authors who have ever posted in the forum until then
total.authors.percentages = total.authors
proportions = character()
for (i in 1:dim(total.authors)[1]) {
  proportions = c(proportions, total.authors.percentages[i, 2] / sum(new.authors[1:i, 2
      ]))
}
# column bind
total.authors.percentages = cbind(total.authors.percentages, proportions)
colnames(total.authors.percentages) = c("MonthYear", "TotalAuthors", "Proportion")

# change to numeric
total.authors.percentages$Proportion = as.numeric(as.character(total.authors.percentages
    $Proportion))
total.authors.percentages

# plot the proportion
p = ggplot(total.authors.percentages, aes(MonthYear, Proportion)) + geom_line() + xlab("
    ") + ylab("Proportion of Active Authors")
p = p + scale_x_date(date_labels = "%Y-%m", date_breaks = "6 month")
p = p + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p
```

Statistics for the duration of authors are active for on the forum and t.test comparing more active users against less active users.

```
webforum.temp = webforum
# convert to date
webforum.temp$Date = as.Date(webforum.temp$Date, "%Y-%m-%d")

# group by date and author
dur = as.data.frame(as.table(by(webforum.temp$WC, list(webforum.temp$Date, webforum.temp
    $AuthorID), FUN=length)))
dur = na.omit(dur)
dur = dur[,-3]
colnames(dur) = c("Date", "AuthorID")
# convert to date data type
dur$Date = as.Date(dur$Date, "%Y-%m-%d")

# convert to character
dur$AuthorID = as.character(dur$AuthorID)

# Find the date ranges, inclusive.
date.ranges = dur %>% group_by(AuthorID) %>%
  arrange(Date) %>%
  summarize(Duration = as.numeric(diff(range(Date)))+1)

date.ranges$AuthorID = as.numeric(date.ranges$AuthorID)
date.ranges = date.ranges[order(date.ranges$AuthorID),]
summary(date.ranges$Duration)
getmode(date.ranges$Duration)

hist(date.ranges$Duration, main="",
     xlab="Days",
     border="black",
     col="orange",
     las=1,
```

```
     breaks=20)

# Calculate confidence intervals for these means.
t.test(date.ranges$Duration)

# Can we conclude that more active authors are active for a longer period of time on avg. on the forum than less active authors?
less.active.daterange = date.ranges[(date.ranges$AuthorID %in% less.active$AuthorID),]
more.active.daterange = date.ranges[(date.ranges$AuthorID %in% more.active$AuthorID),]

t.test(more.active.daterange$Duration, less.active.daterange$Duration, alternative="greater")
```

Statistics for the responsiveness of authors and t.test:

```
webforum.temp = webforum
# convert to date
webforum.temp$Date = as.Date(webforum.temp$Date, "%Y-%m-%d")

# group by date and author
dur = as.data.frame(as.table(by(webforum.temp$WC, list(webforum.temp$Date, webforum.temp$AuthorID), FUN=length)))
dur = na.omit(dur)
dur = dur[,-3]
colnames(dur) = c("Date", "AuthorID")
# convert to date data type
dur$Date = as.Date(dur$Date, "%Y-%m-%d")

# convert to character
dur$AuthorID = as.numeric(as.character(dur$AuthorID))

# find frequency in dur
dur.freq = as.data.frame(as.table(by(dur$Date, list(dur$AuthorID), FUN=length)))
colnames(dur.freq) = c("AuthorID", "NumPosts")
dur.freq$AuthorID = as.numeric(as.character(dur.freq$AuthorID))
dur.freq = dur.freq[order(dur.freq$AuthorID),]

# Remove all authors that have posted only once as we cannot measure their responsiveness.
dur.freq = dur.freq[dur.freq$NumPosts!=1,]

# Remove all authors that have only posted once in dur.
dur = dur[(dur$AuthorID %in% dur.freq$AuthorID),]

# Find average duration between posts for each author.
avg.dur = dur %>% group_by(AuthorID) %>%
  arrange(Date) %>%
  summarize(avg = as.numeric(mean(diff(Date))))

summary(avg.dur$avg)
getmode(avg.dur$avg)
hist(avg.dur$avg, main="", xlab="Days", border="black", col="orange", las=1, breaks=10)

# Calculate confidence intervals.
t.test(avg.dur$avg)

less.active.avgdur = avg.dur[(avg.dur$AuthorID %in% less.active$AuthorID),]
more.active.avgdur = avg.dur[(avg.dur$AuthorID %in% more.active$AuthorID),]

t.test(more.active.avgdur$avg, less.active.avgdur$avg, alternative="less")
```

Plotting the number of monthly posts over time and decomposing the time series:

```r
# Consider month and year only
webforum$MonthYear = as.Date(format(webforum$Date, "%Y-%m-01"))

posts.per.month = as.data.frame(as.table(by(webforum$WC, list(webforum$MonthYear), FUN=
    length)))
colnames(posts.per.month) = c("Date", "NumPosts")
# convert date column to date data type
posts.per.month$Date = as.Date(posts.per.month$Date, "%Y-%m-%d")

# find the varaince
var(posts.per.month$NumPosts)

# plot
t = ggplot(posts.per.month, aes(Date, NumPosts)) + geom_line(color="red4") + xlab("Time"
    ) + ylab("Monthly Posts")
t = t + scale_x_date(date_labels = "%m-%Y", date_breaks = "6 month")
t = t + theme(axis.text.x = element_text(angle = 45, hjust = 1))
t = t
t

# save the time series ggplot graph
ggsave("~/Desktop/FIT3152_resources/posts_over_time.png", t, width = 20, height = 13,
    units = "cm")

posts.ts = ts(posts.per.month$NumPosts, frequency = 12)
posts.ts
# decomposition of time series
decomp = decompose(posts.ts)
plot(decomp)
```

Figure 3: Heat Map:

```r
# Descriptive statistics
summary(webforum)

# plot heat map of linguistic variables vs, year where shades represent average
webforum$Year = as.numeric(format(webforum$Date, "%Y"))
liwc.year.avg = aggregate(webforum[6:29], list(webforum$Year), FUN=mean)
colnames(liwc.year.avg)[1] = "Year"
liwc.year.long.avg = gather(data = liwc.year.avg, key = LIWC, value = Proportion, -1)

# heat map
a = ggplot(data=liwc.year.long.avg, aes(x=Year, y=LIWC))
a = a + geom_tile(aes(fill=Proportion))
a = a + scale_x_continuous(labels=liwc.year.long.avg$Year, breaks=liwc.year.long.avg$
    Year)
a
```

Figure 4: Correlation matrix as heat map:

```r
webforum$Date = as.Date(webforum$Date, "%Y-%m-%d")

# plot heat map of linguistic variables vs. year where shades represent sum
corr = round(cor(webforum[6:29]), digits=2)

# Get lower triangle of the correlation matrix
  get_lower_tri<-function(cormat){
    cormat[upper.tri(cormat)] <- NA
    return(cormat)
  }
  # Get upper triangle of the correlation matrix
  get_upper_tri <- function(cormat){
```

```
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }

# reorder matrix due to correlation coefficient
reorder_cormat <- function(cormat){
# Use correlation between variables as distance
dd <- as.dist((1-cormat)/2)
hc <- hclust(dd)
cormat <-cormat[hc$order, hc$order]
}

cormat = reorder_cormat(corr)
upper.tri = get_upper_tri(corr)
# melted
melted.cormat <- melt(upper.tri, na.rm = TRUE)
# Create a ggheatmap
g <- ggplot(melted.cormat, aes(Var2, Var1, fill = value)) +
    geom_tile(color = "white")+
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") +
    theme_minimal() + # minimal theme
    theme(axis.text.x = element_text(angle = 90, vjust = 0,
                                     size = 8, hjust = 1)) +
    coord_fixed()
g = g + geom_text(aes(label = value), color="black", size=1.5)
g = g + xlab("LIWC") + ylab("LIWC")
g
```

5.3.2 Part 2

Describing summary statistics for the number of posts per thread.

```
# Count the number of threads.
threads = as.data.frame(as.table(by(webforum$WC, list(webforum$ThreadID), FUN=length)))
colnames(threads) = c("ThreadID", "NumPosts")

# Order from least posts to the most posts.
threads = threads[order(threads$NumPosts),]
summary(threads$NumPosts)
boxplot(threads$NumPosts)
hist(threads$NumPosts, xlab="Number of Threads", border="black", col="orange", las=1,
    breaks=10)
getmode(threads$NumPosts)

# Fix the right skew.
fix.skew = threads[threads$NumPosts > 0,]
fix.skew = fix.skew[fix.skew$NumPosts < 80,]

fix.skew

hist(fix.skew$NumPosts, main="", xlab="log(Number of Authors)", border="black", col="
    orange", las=1, breaks=7)
hist(log(fix.skew$NumPosts), main="", xlab="log(Number of Authors)", border="black", col
    ="orange", las=1, breaks=7)

# Define threads who are more active as those who are in the top 10%.
top.threads = fix.skew[fix.skew$NumPosts >= quantile(fix.skew$NumPosts, prob=0.90),]
# Define threads who are less active as those are in the bottom 10%.
```

```
bottom.threads = fix.skew[fix.skew$NumPosts <= quantile(fix.skew$NumPosts, prob=0.10),]
top.threads
```

Threads by word count:

```
# Count the number of threads.
wc.threads = as.data.frame(as.table(by(webforum$WC, list(webforum$ThreadID), FUN=sum)))
colnames(wc.threads) = c("ThreadID", "TotalWC")

# Order
wc.threads = wc.threads[order(wc.threads$TotalWC),]

summary(wc.threads)

# Can we conclude that threads with a larger number of posts have a higher word count
    than threads with a lower number of posts?
top.posts.wc = wc.threads[wc.threads$ThreadID %in% top.threads$ThreadID,]
bottom.posts.wc = wc.threads[wc.threads$ThreadID %in% bottom.threads$ThreadID,]

t.test(top.posts.wc$TotalWC, bottom.posts.wc$TotalWC, alternative="greater")
```

Summary statistics for the number of authors per thread and perform t-test for the most popular and least popular threads.

```
# Group by number of threads and authors.
author.threads = as.data.frame(as.table(by(webforum$WC, list(webforum$ThreadID, webforum
    $AuthorID), FUN=length)))
colnames(author.threads) = c("ThreadID", "AuthorID", "NumPosts")
author.threads = na.omit(author.threads)
author.threads = author.threads[,-3]
author.threads = as.data.frame(as.table(by(author.threads$AuthorID, list(author.threads$
    ThreadID), FUN=length)))
colnames(author.threads) = c("ThreadID", "NumAuthors")
author.threads = author.threads[order(author.threads$NumAuthors),]

summary(author.threads$NumAuthors)
getmode(author.threads$NumAuthors)
boxplot(author.threads$NumAuthors)
hist(author.threads$NumAuthors)
t.test(author.threads$NumAuthors)

# Can we conclude that more popular threads (by post) have more authors posting in it
    than less popular threads? Yes, quite significant.
top.posts = author.threads[author.threads$ThreadID %in% top.threads$ThreadID,]
bottom.posts = author.threads[author.threads$ThreadID %in% bottom.threads$ThreadID,]
t.test(top.posts$NumAuthors, bottom.posts$NumAuthors)
```

Describe the summary statistics for how long threads are usually active for. Perform t-test for the most popular and least popular threads.

```
activity.period = as.data.frame(as.table(by(webforum$WC, list(webforum$Date, webforum$
    ThreadID), FUN=length)))
activity.period = na.omit(activity.period)
activity.period = activity.period[,-3]
colnames(activity.period) = c("Date", "ThreadID")

# Convert to date data type.
activity.period$Date = as.Date(activity.period$Date, "%Y-%m-%d")
activity.period$ThreadID = as.character(activity.period$ThreadID)

```

```
# Calculate date ranges.
date.ranges = activity.period %>% group_by(ThreadID) %>%
  arrange(Date) %>%
  summarize(Duration = as.numeric(diff(range(Date))))

date.ranges$ThreadID = as.numeric(date.ranges$ThreadID)
date.ranges = date.ranges[order(date.ranges$ThreadID),]

summary(date.ranges$Duration)
getmode(date.ranges$Duration)
boxplot(date.ranges$Duration)
hist(date.ranges$Duration)
t.test(date.ranges$Duration)

# Can we conclude that more popular threads (by post) are active for longer periods of
    time than less popular threads?
# No, we cannot conclude this at all. It's not significant.
top.posts = date.ranges[date.ranges$ThreadID %in% top.threads$ThreadID,]
bottom.posts = date.ranges[date.ranges$ThreadID %in% bottom.threads$ThreadID,]
t.test(top.posts$Duration, bottom.posts$Duration, alternative="greater")
```

Describe the summary statistics for the duration between new posts per thread. Perform t-test for the most popular and least popular threads.

```
thread.dur = as.data.frame(as.table(by(webforum$WC, list(webforum$Date, webforum$
    ThreadID), FUN=length)))
thread.dur = na.omit(thread.dur)
thread.dur = thread.dur[,-3]
colnames(thread.dur) = c("Date", "ThreadID")

# Convert to date data type.
thread.dur$Date = as.Date(thread.dur$Date, "%Y-%m-%d")
thread.dur$ThreadID = as.character(thread.dur$ThreadID)

avg.thread.dur = thread.dur %>% group_by(ThreadID) %>%
  arrange(Date) %>%
  summarize(Average = as.numeric(mean(diff(Date))))

avg.thread.dur = avg.thread.dur[order(avg.thread.dur$Average),]

summary(avg.thread.dur$Average)
getmode(avg.thread.dur$Average)
boxplot(avg.thread.dur$Average)
hist(avg.thread.dur$Average)
t.test(avg.thread.dur$Average)

# Can we conclude that more popular threads (by post) have shorter average durations
    between posts than less popular threads? No, not significant.
top.posts = avg.thread.dur[avg.thread.dur$ThreadID %in% top.threads$ThreadID,]
bottom.posts = avg.thread.dur[avg.thread.dur$ThreadID %in% bottom.threads$ThreadID,]
t.test(top.posts$Average, bottom.posts$Average, alternative="less")
```

Identify the main linguistic variables in the threads that have a lot of posts vs. threads that do not.

```
high.liwc = webforum[webforum$ThreadID %in% top.threads$ThreadID,]
low.liwc = webforum[webforum$ThreadID %in% bottom.threads$ThreadID,]

high.liwc = high.liwc[,6:29]
low.liwc = low.liwc[,6:29]

high.liwc = melt(high.liwc)
```

```
high.liwc.mean <- as.data.frame(as.table(by(high.liwc$value, list(high.liwc$variable), mean)))
colnames(high.liwc.mean) = c("LIWC", "Proportion")
high.liwc.mean = cbind(high.liwc.mean, Class="Threads with Most Number of Posts")

low.liwc = melt(low.liwc)
low.liwc.mean <- as.data.frame(as.table(by(low.liwc$value, list(low.liwc$variable), mean)))
colnames(low.liwc.mean) = c("LIWC", "Proportion")
low.liwc.mean = cbind(low.liwc.mean, Class="Threads with Lowest Number of Posts")

# Row bind
liwc.mean = rbind(high.liwc.mean, low.liwc.mean)

# Bar chart
g = ggplot(aes(fill=Class, y=Proportion, x=LIWC), data=liwc.mean)
g = g + geom_bar(position="dodge", stat="identity") + coord_flip()
g = g + theme_minimal()
g
ggsave("~/Desktop/FIT3152_resources/bar_chart.png", g, width = 20, height = 13, units = "cm")
liwc.vars = as.vector(unique(high.liwc$variable))

# higher word count liwc variable proportion greater than lower word count
for (x in liwc.vars) {
  print(x)
  print(t.test(high.liwc$value[high.liwc$variable==x], low.liwc$value[low.liwc$variable==x], alternative = "greater"))
}

# higher word count liwc variable proportion greater than lower word count
for (x in liwc.vars) {
  print(x)
  print(t.test(high.liwc$value[high.liwc$variable==x], low.liwc$value[low.liwc$variable==x], alternative = "less"))
}
```

Analyse Thread 252620. The same analysis is done for threads 532649, 283958, 145223, 472752, 127115.

```
thread.details = webforum[webforum$ThreadID=="252620",]
# Group by MonthYear.
thread.details$MonthYear = as.Date(format(thread.details$Date, "%Y-%m-01"))

liwc.thread = aggregate(thread.details[6:29], list(thread.details$Date), FUN=mean)
colnames(liwc.thread)[1] = "Date"

melted.thread = melt(liwc.thread, id.vars=c("Date"))
colnames(melted.thread)[2] = "LIWC"
colnames(melted.thread)[3] = "Proportion"

liwc.thread.month = aggregate(thread.details[6:29], list(thread.details$MonthYear), FUN=mean)
colnames(liwc.thread.month)[1] = "DateMonth"

melted.thread.month = melt(liwc.thread.month, id.vars=c("DateMonth"))
colnames(melted.thread.month)[2] = "LIWC"
colnames(melted.thread.month)[3] = "Proportion"

attach(melted.thread)
# plot faceted line graph
g = ggplot(melted.thread, aes(x=Date, y=Proportion)) + geom_jitter(colour="cadetblue3", alpha=0.4)
```

```
g = g + facet_wrap( ~LIWC, ncol=4)
g = g + geom_line(data=melted.thread.month, aes(x= DateMonth, y= Proportion), colour="
    black")
g

liwc.thread.analytic = aggregate(thread.details[6:9], list(thread.details$Date), FUN=
    mean)
colnames(liwc.thread.analytic)[1] = "Date"
melted.thread.analytic = melt(liwc.thread.analytic, id.vars=c("Date"))
colnames(melted.thread.analytic)[2] = "LIWC"
colnames(melted.thread.analytic)[3] = "Proportion"

# plot faceted line graph, analytic, clout, authentic and tone
g = ggplot(melted.thread.analytic, aes(x=Date, y=Proportion)) + geom_jitter(colour="
    cadetblue3", alpha=0.4)
g = g + facet_wrap( ~LIWC, nrow=4)
g = g + stat_smooth(method = 'lm', aes(colour = 'Trend'), se = FALSE)
g = g + labs(color="Legend")
g

ggsave("~/Desktop/FIT3152_resources/LIWC_over_time1pt2.png", g, width = 20, height = 13,
     units = "cm")
detach(melted.thread)
```

5.3.3 Part 3

Analyse for March- the same analysis is done for April and May.

```
webforum$Year = as.Date(format(webforum$Date, "%Y-%m-01"))
thread = webforum[webforum$Year=="2005-03-01",]

# Group by number of threads and authors.
author.threads = as.data.frame(as.table(by(thread$WC, list(thread$ThreadID, thread$
    AuthorID), FUN=length)))
colnames(author.threads) = c("ThreadID", "AuthorID", "NumPosts")
author.threads = na.omit(author.threads)
author.threads = author.threads[,-3]
author.threads = as.data.frame(as.table(by(author.threads$AuthorID, list(author.threads$
    ThreadID), FUN=length)))
colnames(author.threads) = c("ThreadID", "NumAuthors")
author.threads = author.threads[order(author.threads$NumAuthors),]

# There are 46 authors.
length(unique(thread$AuthorID))

# There are 7 threads in this month.
# Examine each of these threads.
thread.id = as.vector(author.threads$ThreadID)

threads = vector("list", length(thread.id))

for (i in seq_along(thread.id)) {
  threads[[i]] = webforum[webforum$ThreadID==thread.id[i] & webforum$Year==as.Date(
      format("2005-03-01")),]
}

author.id = vector()
# Examine the social network within each thread.
for (i in seq_along(thread.id)) {
  author.id = c(author.id, unique(threads[[i]][2]))
```

```
}

# graph adjacency matrices
matrices = vector("list", length(author.id))
for (i in seq_along(author.id)) {
  x = length(author.id[[i]])
  matrices[[i]] = matrix(runif(x*x, min=1, max=1), ncol=x)
  diag(matrices[[i]]) = 0
  colnames(matrices[[i]]) = author.id[[i]]
  rownames(matrices[[i]]) = author.id[[i]]
}

# graph social networks
graphs = vector("list", length(matrices))
par(mfrow=c(2, 3))
for (i in seq_along(matrices)) {
  graphs[[i]] = graph.adjacency(matrices[[i]], mode="undirected", weighted=NULL)
  if (i > 2) {
    par(mar=c(0,0,0,0)); plot(graphs[[i]], layout=layout.circle, vertex.color="red",
       vertex.label.cex = c(0.5),
       vertex.label.color = "black",
       vertex.frame.color = "white",
       vertex.size=20)
  }
}

par(mfrow=c(1, 1))
# merge graphs into huge social network
mar = graph_from_literal()
for (i in seq_along(graphs)) {
  mar = (mar %u% graphs[[i]])
}

V(mar)$size <- log(strength(mar)) * 3 + 3
V(mar)$label <- ifelse( strength(mar)>=10, V(mar)$name, NA )
par(mar=c(0,0,0,0))
png('march_graph.png', width = 2000, height = 1700)
plot(mar, layout=layout.fruchterman.reingold,
     vertex.frame.color="white",
     vertex.label.color="black",
     vertex.label.cex=c(1.5))
dev.off()

# Calculate statistics
mar

# Degree Distribution
hist(degree(mar), breaks = 30, col="grey")
deg = as.table(degree(mar))

# Betweeness
bet = (format(as.table(betweenness(mar)), digits=2))

# Closeness
clo = format(as.table(closeness(mar)), digits=2)

# Eigenvector
eig = as.table(evcent(mar)$vector)

# Put into data frame
summary = as.data.frame(cbind(deg, bet, clo, eig))
summary
```

```

# Diameter
diameter(mar)

# Average Path Length
average.path.length(mar)

# Cliques
# table(sapply(cliques(mar), length))

attach(summary)
deg = as.numeric(deg)
bet = as.numeric(bet)
clo = as.numeric(clo)
eig = as.numeric(eig)

hdeg = summary[order(-deg),]
head(hdeg)

hclo = summary[order(-clo),]
head(hclo)

hbet = summary[order(-bet),]
head(hbet)

heig = summary[order(-eig),]
head(heig)

detach(summary)
```

Analyse for March and April. The same analysis is done to combine March, April and May.

```
comb = (mar %u% apr)

V(comb)$size <- log(strength(comb)) * 3 + 3
V(comb)$label <- ifelse( strength(comb)>=10, V(comb)$name, NA )
par(mar=c(0,0,0,0))
png('march_april_graph.png', width = 2000, height = 1500)
plot(comb, layout=layout.fruchterman.reingold,
     vertex.frame.color="white",
     vertex.label.color="black",
     vertex.label.cex=c(1.3))
dev.off()

# Calculate statistics
comb

# Degree Distribution
hist(degree(comb), breaks = 30, col="grey")
deg = as.table(degree(comb))

# Betweeness
bet = (format(as.table(betweenness(comb)), digits=2))

# Closeness
clo = format(as.table(closeness(comb)), digits=2)

# Eigenvector
eig = as.table(evcent(comb)$vector)

# Put into data frame
```

```r
summary = as.data.frame(cbind(deg, bet, clo, eig))
summary

# Diameter
diameter(comb)

# Average Path Length
average.path.length(comb)

# Cliques
# table(sapply(cliques(comb), length))

attach(summary)
deg = as.numeric(deg)
bet = as.numeric(bet)
clo = as.numeric(clo)
eig = as.numeric(eig)

hdeg = summary[order(-deg),]
head(hdeg)

hclo = summary[order(-clo),]
head(hclo)

hbet = summary[order(-bet),]
head(hbet)

heig = summary[order(-eig),]
head(heig)

detach(summary)
```