Monash University Malaysia

BSc Computer Science

Assignment # 1

Social and linguistic dynamics of an online community

FIT3152 Data Analytics - Report

Prepared By:

Bazil Muzaffar Kotriwala

Siddharth Anil Shinde

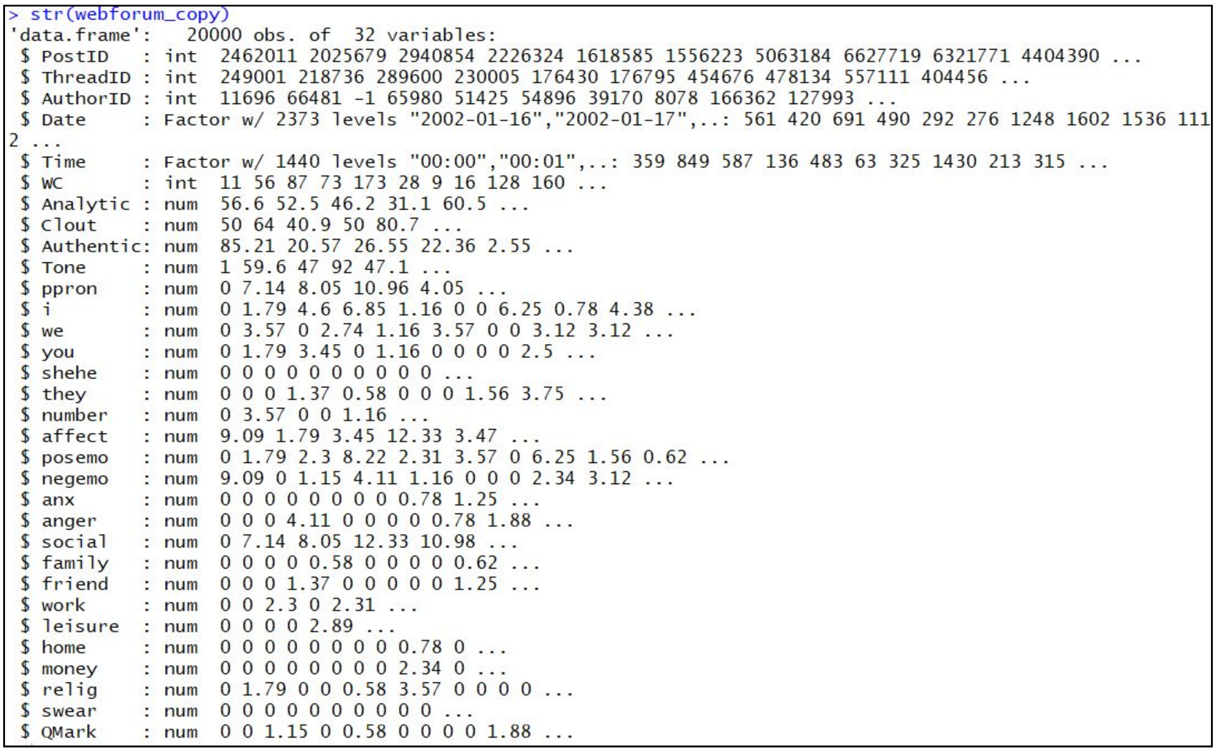
Table of Contents

Overview

This report investigates whether people adopt similar patterns of language use when they interact. The dataset (webforum.csv) being analysed contains data from an online forum where participants communicate with each other via conversations in a thread. The main focus is to find supporting evidence on whether members who are communicating directly on a specific thread with each other use similar language. In addition, whether the language used by the same authors changes over a period of time.

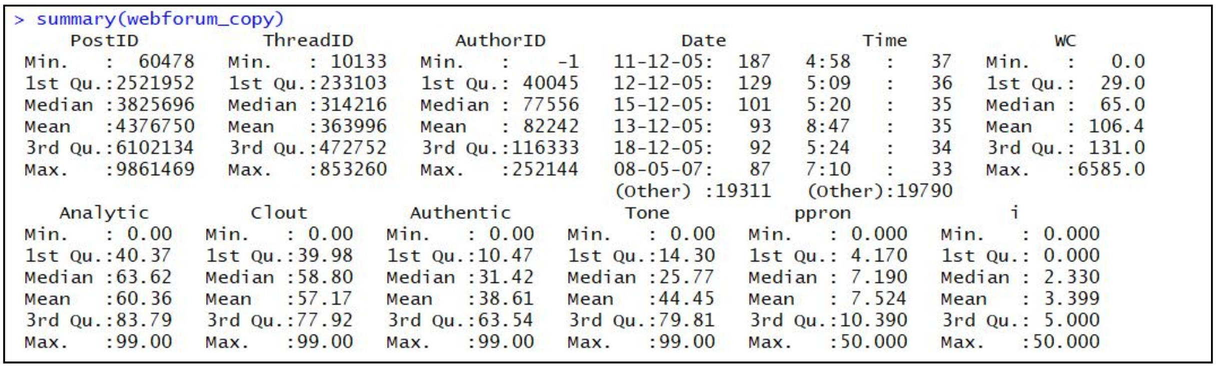
Introduction

The dataset (webforum.csv) contains a total of 20,000 observations. There are a total of 32 columns containing values for PostID, ThreadID, AuthorID, Date, Time, Word Count etc. The large proportion of column variables contain the percentage of linguistic words and categories which have occurred on the post such as 'we', 'you' or the amount of words written in the post belonging to a certain category such as 'Analytic', 'Authentic' etc. The datatypes used for the column values are all integers and numbers respectively.



```
> str(webforum_copy)
'data.frame':	20000 obs. of  32 variables:
 $ PostID   : int  2462011 2025679 2940854 2226324 1618585 1556223 5063184 6627719 6321771 4404390 ...
 $ ThreadID : int  249001 218736 289600 230005 176430 176795 454676 478134 557111 404456 ...
 $ AuthorID : int  11696 66481 -1 65980 51425 54896 39170 8078 166362 127993 ...
 $ Date     : Factor w/ 2373 levels "2002-01-16","2002-01-17",..: 561 420 691 490 292 276 1248 1602 1536 111
2 ...
 $ Time     : Factor w/ 1440 levels "00:00","00:01",..: 359 849 587 136 483 63 325 1430 213 315 ...
 $ WC       : int  11 56 87 73 173 28 9 16 128 160 ...
 $ Analytic : num  56.6 52.5 46.2 31.1 60.5 ...
 $ Clout    : num  50 64 40.9 50 80.7 ...
 $ Authentic: num  85.21 20.57 26.55 22.36 2.55 ...
 $ Tone     : num  1 59.6 47 92 47.1 ...
 $ ppron    : num  0 7.14 8.05 10.96 4.05 ...
 $ i        : num  0 1.79 4.6 6.85 1.16 0 0 6.25 0.78 4.38 ...
 $ we       : num  0 3.57 0 2.74 1.16 3.57 0 0 3.12 3.12 ...
 $ you      : num  0 1.79 3.45 0 1.16 0 0 0 0 2.5 ...
 $ shehe    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ they     : num  0 0 0 1.37 0.58 0 0 0 1.56 3.75 ...
 $ number   : num  0 3.57 0 0 1.16 ...
 $ affect   : num  9.09 1.79 3.45 12.33 3.47 ...
 $ posemo   : num  0 1.79 2.3 8.22 2.31 3.57 0 6.25 1.56 0.62 ...
 $ negemo   : num  9.09 0 1.15 4.11 1.16 0 0 0 2.34 3.12 ...
 $ anx      : num  0 0 0 0 0 0 0 0 0.78 1.25 ...
 $ anger    : num  0 0 0 4.11 0 0 0 0 0.78 1.88 ...
 $ social   : num  0 7.14 8.05 12.33 10.98 ...
 $ family   : num  0 0 0 0 0.58 0 0 0 0 0.62 ...
 $ friend   : num  0 0 0 1.37 0 0 0 0 0 1.25 ...
 $ work     : num  0 0 2.3 0 2.31 ...
 $ leisure  : num  0 0 0 0 2.89 ...
 $ home     : num  0 0 0 0 0 0 0 0 0.78 0 ...
 $ money    : num  0 0 0 0 0 0 0 0 2.34 0 ...
 $ relig    : num  0 1.79 0 0 0.58 3.57 0 0 0 0 ...
 $ swear    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ QMark    : num  0 0 1.15 0 0.58 0 0 0 0 1.88 ...
```

Fig 1.0: Structure of Webforum Dataset



```
> summary(webforum_copy)
     PostID           ThreadID          AuthorID              Date             Time             WC        
 Min.   :  60478   Min.   : 10133   Min.   :    -1   11-12-05:  187   4:58   :   37   Min.   :   0.0  
 1st Qu.:2521952   1st Qu.:233103   1st Qu.: 40045   12-12-05:  129   5:09   :   36   1st Qu.:  29.0  
 Median :3825696   Median :314216   Median : 77556   15-12-05:  101   5:20   :   35   Median :  65.0  
 Mean   :4376750   Mean   :363996   Mean   : 82242   13-12-05:   93   8:47   :   35   Mean   : 106.4  
 3rd Qu.:6102134   3rd Qu.:472752   3rd Qu.:116333   18-12-05:   92   5:24   :   34   3rd Qu.: 131.0  
 Max.   :9861469   Max.   :853260   Max.   :252144   08-05-07:   87   7:10   :   33   Max.   :6585.0  
                                                     (Other) :19311   (Other):19790                   
    Analytic         Clout         Authentic          Tone           ppron              i         
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:40.37   1st Qu.:39.98   1st Qu.:10.47   1st Qu.:14.30   1st Qu.: 4.170   1st Qu.: 0.000  
 Median :63.62   Median :58.80   Median :31.42   Median :25.77   Median : 7.190   Median : 2.330  
 Mean   :60.36   Mean   :57.17   Mean   :38.61   Mean   :44.45   Mean   : 7.524   Mean   : 3.399  
 3rd Qu.:83.79   3rd Qu.:77.92   3rd Qu.:63.54   3rd Qu.:79.81   3rd Qu.:10.390   3rd Qu.: 5.000  
 Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :50.000   Max.   :50.000  
```

Fig 1.1 (Excerpt): Summary of Webforum Dataset

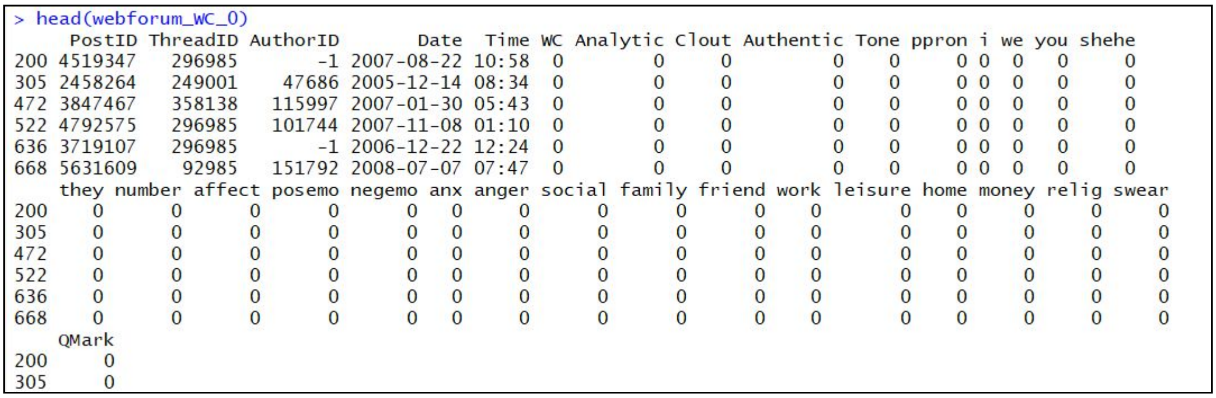
The summary statistics of the linguistic variables in the dataset are retrieved to get an overview of the data. These summary statistics give us an insight into the linguistic variables such as ‘Clout’, ‘Analytic’ which can be used to check the impact of them over time.

Pre-Processing

Data Cleaning

A copy of the original dataset was made so that the original dataset remains unaffected and intact from any sort of manipulation. All the manipulation of data was done on the copy. New subsets were created and written to csv files depending on the data manipulated on the copy.

The posts containing a word count of 0 (i.e. the posts which were either images or videos) were removed from the dataset since they were not useful in our analysis as they did not give us any insight on how people were communicating on a thread. The images and videos could not be categorised into a type in the data collection. Therefore, this data was redundant and of no use to the analysis being done.

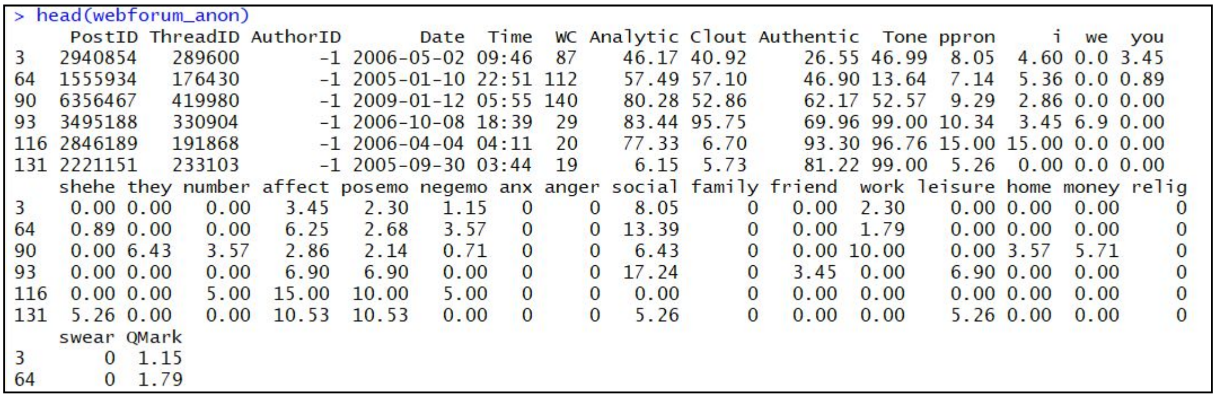


```
> head(webforum_WC_0)
     PostID ThreadID AuthorID       Date  Time WC Analytic Clout Authentic Tone ppron i we you shehe
200 4519347   296985       -1 2007-08-22 10:58  0        0     0         0    0     0 0  0   0     0
305 2458264   249001    47686 2005-12-14 08:34  0        0     0         0    0     0 0  0   0     0
472 3847467   358138   115997 2007-01-30 05:43  0        0     0         0    0     0 0  0   0     0
522 4792575   296985   101744 2007-11-08 01:10  0        0     0         0    0     0 0  0   0     0
636 3719107   296985       -1 2006-12-22 12:24  0        0     0         0    0     0 0  0   0     0
668 5631609    92985   151792 2008-07-07 07:47  0        0     0         0    0     0 0  0   0     0
    they number affect posemo negemo anx anger social family friend work leisure home money relig swear
200    0      0      0      0      0   0     0      0      0      0    0       0    0     0     0     0
305    0      0      0      0      0   0     0      0      0      0    0       0    0     0     0     0
472    0      0      0      0      0   0     0      0      0      0    0       0    0     0     0     0
522    0      0      0      0      0   0     0      0      0      0    0       0    0     0     0     0
636    0      0      0      0      0   0     0      0      0      0    0       0    0     0     0     0
668    0      0      0      0      0   0     0      0      0      0    0       0    0     0     0     0
    QMark
200     0
305     0
```

Fig 2.0: Rows removed (WC = 0)

After removing all WC = 0 rows, the total number of observations changed from 20,000 to 19,922. Therefore, a total of 78 rows were removed.

Furthermore, the dataset contains data collected from posts by anonymous authors (AuthorID = -1). This data does not give us any information on which authors are communicating directly with each other. Therefore, we cannot identify the trends and similarities in language used by these anonymous authors, making this data redundant hence, it has been removed from the dataset. This changed the total number of observations from 19,922 to 18,808. Therefore, a total of 1114 rows were removed.



```
> head(webforum_anon)
    PostID ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic  Tone ppron     i  we  you
3   2940854   289600       -1 2006-05-02 09:46  87    46.17 40.92     26.55 46.99  8.05  4.60 0.0 3.45
64  1555934   176430       -1 2005-01-10 22:51 112    57.49 57.10     46.90 13.64  7.14  5.36 0.0 0.89
90  6356467   419980       -1 2009-01-12 05:55 140    80.28 52.86     62.17 52.57  9.29  2.86 0.0 0.00
93  3495188   330904       -1 2006-10-08 18:39  29    83.44 95.75     69.96 99.00 10.34  3.45 6.9 0.00
116 2846189   191868       -1 2006-04-04 04:11  20    77.33  6.70     93.30 96.76 15.00 15.00 0.0 0.00
131 2221151   233103       -1 2005-09-30 03:44  19     6.15  5.73     81.22 99.00  5.26  0.00 0.0 0.00
    shehe they number affect posemo negemo anx anger social family friend  work leisure home money relig
3    0.00 0.00   0.00   3.45   2.30   1.15   0     0   8.05      0   0.00  2.30    0.00 0.00  0.00     0
64   0.89 0.00   0.00   6.25   2.68   3.57   0     0  13.39      0   0.00  1.79    0.00 0.00  0.00     0
90   0.00 6.43   3.57   2.86   2.14   0.71   0     0   6.43      0   0.00 10.00    0.00 3.57  5.71     0
93   0.00 0.00   0.00   6.90   6.90   0.00   0     0  17.24      0   3.45  0.00    6.90 0.00  0.00     0
116  0.00 0.00   5.00  15.00  10.00   5.00   0     0   0.00      0   0.00  0.00    0.00 0.00  0.00     0
131  5.26 0.00   0.00  10.53  10.53   0.00   0     0   5.26      0   0.00  0.00    5.26 0.00  0.00     0
    swear QMark
3       0  1.15
64      0  1.79
```

Fig 2.1: Rows removed (AuthorID = -1)

The data was checked for any duplicate values, however, no duplicate values were found. Therefore, the number of rows in the dataset remained unaffected.

The column for ‘Time’ can be omitted from our analysis since the time at which each post was made is of no real significance. Instead, to analyse any variables over a period of time, we can use the ‘Date’ column provided which shows the date at which each post was made.

There are a few variables which could be grouped together for the analysis such as ‘ThreadID’ and ‘AuthorID’ along with any linguistic variable to be analysed such as ‘Clout’. Similarly, ‘affect’, ‘posemo’ and ‘negemo’ are variables which can be grouped together since the sum of ‘posemo’ and ‘negemo’ equal to the ‘affect’ of each respective post.

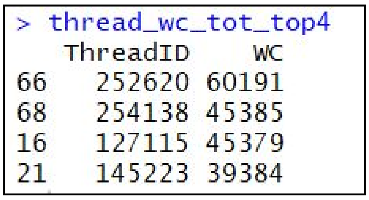
Once the data cleaning was complete, it was now manipulated to create different subsets depending on the criteria of our analysis.

Creating Subsets

The grouping variable chosen for our analysis is *‘ThreadID’*. Since there are a large number of unique threads in the dataset, we are going to choose the top four threads which have the highest word count. The idea behind choosing the top four threads is that we will be analysing the threads with the highest word count, which means that these respective threads will have a large amount of data showing us different trends and similarities which we are looking for. In each of these top four threads, we will be comparing the language factors against each of the author's posts and change over time in that respective thread. This would show whether the author's posting in that respective thread are using similar language as to the author's on that same thread. To achieve this goal, we need to create subsets. The following procedure was followed to create the respective subsets for our use:

1. The total sum of the word count of each *'ThreadID'* was found and was created into a subset along with its respective *'ThreadID'*.

2. These ThreadID's were sorted in order of highest word count to lowest word count. The top 4 highest word count ThreadID's were chosen and stored in a new subset which was written to a csv file.

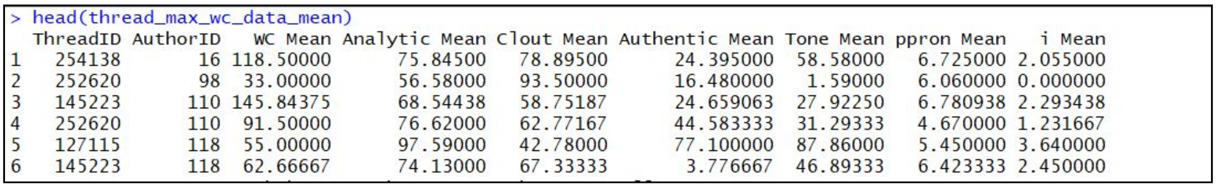


```
> thread_wc_tot_top4
   ThreadID    WC
66   252620 60191
68   254138 45385
16   127115 45379
21   145223 39384
```

Fig 3.0: Top 4 Max WC ThreadID's

3. All the data of these four specific threads was found and was binded together into one subset.

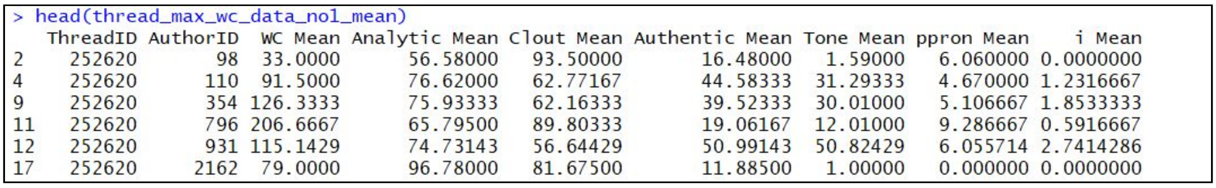
4. The mean of all the linguistic factors was taken out for each author in that respective thread thereby shrinking the repeated posts by the same author to just the mean linguistic values of all the posts that author has done in that specific thread.



```
> head(thread_max_wc_data_mean)
  ThreadID AuthorID   WC Mean Analytic Mean Clout Mean Authentic Mean Tone Mean ppron Mean   i Mean
1   254138       16 118.50000      75.84500   78.89500      24.395000  58.58000   6.725000 2.055000
2   252620       98  33.00000      56.58000   93.50000      16.480000   1.59000   6.060000 0.000000
3   145223      110 145.84375      68.54438   58.75187      24.659063  27.92250   6.780938 2.293438
4   252620      110  91.50000      76.62000   62.77167      44.583333  31.29333   4.670000 1.231667
5   127115      118  55.00000      97.59000   42.78000      77.100000  87.86000   5.450000 3.640000
6   145223      118  62.66667      74.13000   67.33333       3.776667  46.89333   6.423333 2.450000
```

Fig 3.1 (Excerpt): Top 4 Max WC Thread's Linguistic Means

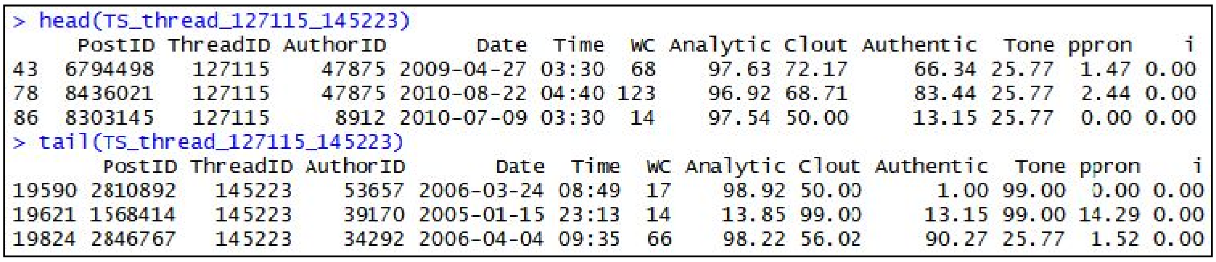
5. Similarly, another subset was created for all the data of the thread with the max word count which was '*ThreadID = 252620'*. This subset was created from the "Top 4 Thread's Linguistic Means" displayed in Fig 3.1. This subset contained the mean of all the linguistic data of each author involved in the '*ThreadID 252620'*.



```
> head(thread_max_wc_data_no1_mean)
   ThreadID AuthorID  WC Mean Analytic Mean Clout Mean Authentic Mean Tone Mean ppron Mean    i Mean
2    252620       98  33.0000      56.58000   93.50000       16.48000   1.59000   6.060000 0.0000000
4    252620      110  91.5000      76.62000   62.77167       44.58333  31.29333   4.670000 1.2316667
9    252620      354 126.3333      75.93333   62.16333       39.52333  30.01000   5.106667 1.8533333
11   252620      796 206.6667      65.79500   89.80333       19.06167  12.01000   9.286667 0.5916667
12   252620      931 115.1429      74.73143   56.64429       50.99143  50.82429   6.055714 2.7414286
17   252620     2162  79.0000      96.78000   81.67500       11.88500   1.00000   0.000000 0.0000000
```

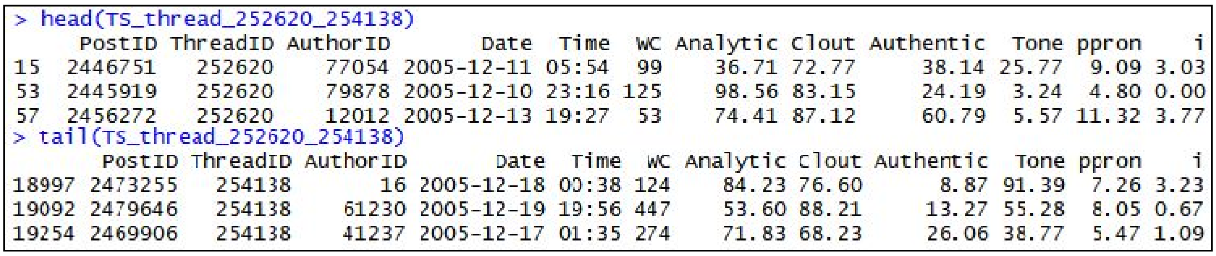
Fig 3.2 (Excerpt): Max WC Thread's Linguistic Means

6. Another subset was created with the dates of each post being retained. The dates were converted into the date format by using *as.Date()* to make plotting a time series more meaningful. These dates can be used to plot across a timeline. The top 4 maximum word count threads were further divided into 2 subsets containing two threads each respectively. The paired threads were '*ThreadID 252620 and 254138'* and *'ThreadID 127115 and 145223'*. These pairs were created based on the fact that their timelines coincided with one another. *Fig 3.3* and *Fig 3.4* displays the excerpt of what the two subsets contain.



```
> head(TS_thread_127115_145223)
     PostID ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic  Tone ppron    i
43  6794498   127115    47875 2009-04-27 03:30  68    97.63 72.17     66.34 25.77  1.47 0.00
78  8436021   127115    47875 2010-08-22 04:40 123    96.92 68.71     83.44 25.77  2.44 0.00
86  8303145   127115     8912 2010-07-09 03:30  14    97.54 50.00     13.15 25.77  0.00 0.00
> tail(TS_thread_127115_145223)
       PostID ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic  Tone ppron    i
19590 2810892   145223    53657 2006-03-24 08:49  17    98.92 50.00      1.00 99.00  0.00 0.00
19621 1568414   145223    39170 2005-01-15 23:13  14    13.85 99.00     13.15 99.00 14.29 0.00
19824 2846767   145223    34292 2006-04-04 09:35  66    98.22 56.02     90.27 25.77  1.52 0.00
```

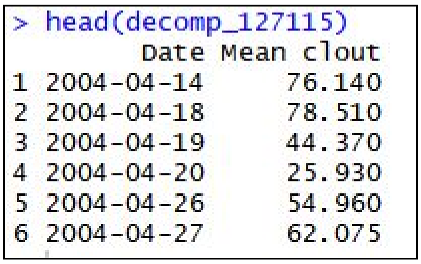
Fig 3.3 (Excerpt): Subset of ThreadID 127115 and ThreadID 145223



```
> head(TS_thread_252620_254138)
     PostID ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic  Tone ppron    i
15  2446751   252620    77054 2005-12-11 05:54  99    36.71 72.77     38.14 25.77  9.09 3.03
53  2445919   252620    79878 2005-12-10 23:16 125    98.56 83.15     24.19  3.24  4.80 0.00
57  2456272   252620    12012 2005-12-13 19:27  53    74.41 87.12     60.79  5.57 11.32 3.77
> tail(TS_thread_252620_254138)
       PostID ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic  Tone ppron    i
18997 2473255   254138       16 2005-12-18 00:38 124    84.23 76.60      8.87 91.39  7.26 3.23
19092 2479646   254138    61230 2005-12-19 19:56 447    53.60 88.21     13.27 55.28  8.05 0.67
19254 2469906   254138    41237 2005-12-17 01:35 274    71.83 68.23     26.06 38.77  5.47 1.09
```

Fig 3.4 (Excerpt): Subset of ThreadID 252620 and ThreadID 254138

7. For the decomposition of a time series, a subset was created for the '*ThreadID 127115'* which spanned over a period of more than three years. The aggregate mean of the linguistic variable *'Clout'* was found for all the posts on each specific date. An excerpt of the subset can be seen in *Fig 3.5.*



```
> head(decomp_127115)
        Date Mean Clout
1 2004-04-14     76.140
2 2004-04-18     78.510
3 2004-04-19     44.370
4 2004-04-20     25.930
5 2004-04-26     54.960
6 2004-04-27     62.075
```

Fig 3.5 (Excerpt): Mean Clout of ThreadID 127115

Multivariate Graph

The multivariate graphs were constructed using two packages ‘plotly’ and ‘ggplot2’.

Max Word Count Thread

The *‘plotly’* package was used on the aforementioned subset in *Fig 3.2* (subset of the thread with the maximum word count). This subset contained the mean values for all the linguistic variables for each unique author on this thread. The variables considered for our analysis are *‘affect’, ‘posemo’, ‘negemo’, ‘anger’ and ‘anx’*. Two separate graphs were created, one highlighting whether the authors use a similar amount of positive and negative language leading to the total effect of the post by that author. The other graph focused on whether the authors use a similar amount of words relating to anger and anxiety.

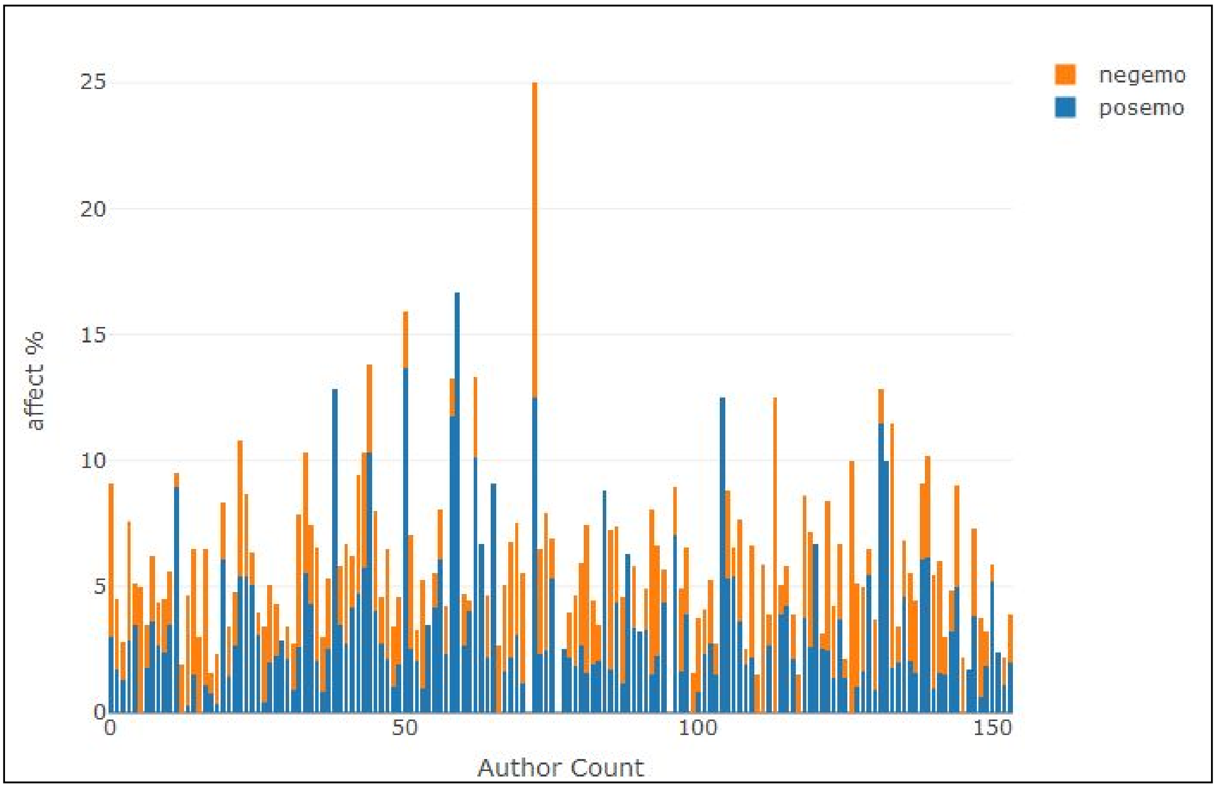


Fig 4.0: Max WC Thread’s Affect

Fig 4.0 shows the mean affect of each author on the specific max word count thread. We can see from the data that majority of the authors using positive language contribute to *0-6%* affect. Similarly, the same majority of the authors also using negative language which is also contributing roughly *0-5%* with the exception of a few outliers for both positive and negative language. Each author is using positive and negative language within the range of *0-6%* and *0-5%* respectively, shows us that all the authors on the thread use similar type of words.

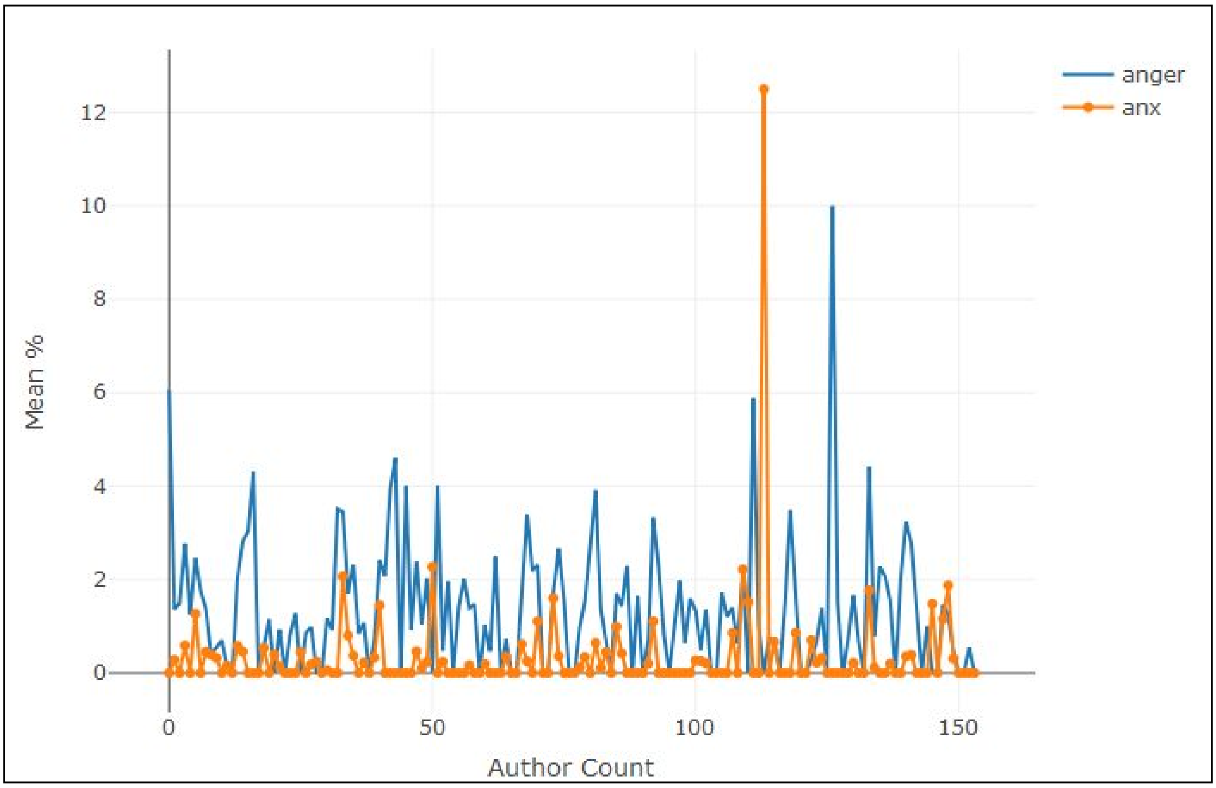


Fig 4.1: Max WC Thread's Mean (Anger & Anxiety)

Fig 4.1 shows the mean percentage of the words relating to anger and anxiety used by the authors in the thread with the maximum word count. We can see that majority of the words related to anxiety only contribute to *0-2%* on average by each author respectively with the exception of two outliers in the graph. Each author using words related to anxiety within the range of *0-2%* shows us that all the authors on the thread express a similar level of anxiety which is close to negligible.

Similarly, anger is another variable considered for each author on the thread. The majority of words used by each author relating to anger are also consistent within the range of *0-4%* with the exception of a few outliers.

The two aforementioned graphs *Fig 4.0* and *Fig 4.1* indicate to us that author's interacting on the same thread use similar language whether it is positive or negative language or words relating to anger or anxiety since they are roughly contributing to the same average for all of them.

However, having said that, it could be the case that this is the scenario for this specific thread. It may be the case, that the authors interacting on another thread may not be using similar language. Nonetheless, using a thread with the maximum word count gives us a better chance of confirming our conclusion as opposed to other threads since we have more data to support it. Applying this technique on a variety of other thread's would give us a clearer and more constructive conclusion.

Top 4 Max Word Count Threads

The *'ggplot2'* package was used on the aforementioned subset in *Fig 3.1*, the subset of the threads with the top four maximum word counts. This subset contained the mean values for all the linguistic variables for each unique author on all the top 4 max word count threads. The variables considered for our analysis are '*ThreadID', 'AuthorID'* and *'Analytic Mean'*.

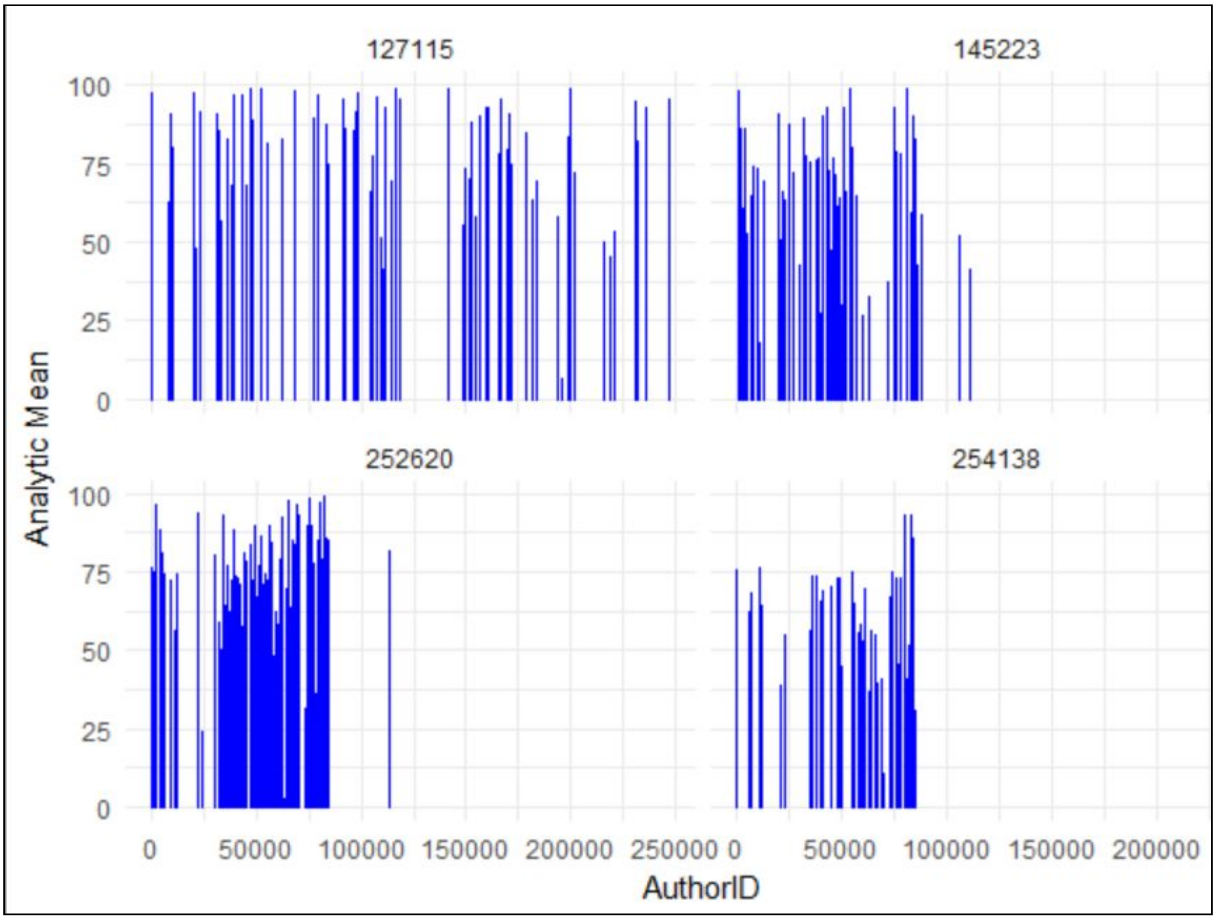


Fig 4.2: Top 4 Max WC Thread's Analytic Mean

In *Fig 4.2*, the graphs displayed are grouped by *'ThreadID'*. Each graph shows the mean analytical thinking value of each author in that respective thread. In respect to the previous graphs constructed, we are now analysing four different threads as opposed to one thread. The dark blue area in the graphs indicate the majority of the authors. Each thread shows that the majority of the authors have the mean analytical thinking value ranging between *50% - 75%* with the exception of a few outliers.

Since all the graphs are showing that on average the majority of the authors have the analytical thinking value within the same range of *50-75%*, therefore, it seems that the majority of the authors use similar language on each respective thread which contributes to their similar level of mean analytical thinking.

Using four different threads with the maximum word counts further affirms any conclusion we make as we are analysing across four different threads respectively as opposed to one thread. Furthermore, we are analysing the top 4 threads with the highest word counts giving us substantial data to make a conclusion.

In light of the above analysis, there still may be room for improvement since these are just four different threads. It is a possibility that there may be a selection bias involved amongst the threads or the authors on these threads may be talking similarly by coincidence. To tackle this, we may want to apply the same technique to a larger number of threads.

Time Series Analysis

The time series visualisations were constructed using *'ggplot2'*, '*ggseas*' and '*seasonal*' packages. The two previously mentioned subsets, shown in *Fig 3.3* and *Fig 3.4*, were used to create the time series plots.

Yearly Paired Threads

The first subset used in *Fig 3.3* was to create a yearly plot for two different threads for the linguistic variable *'Clout'*.

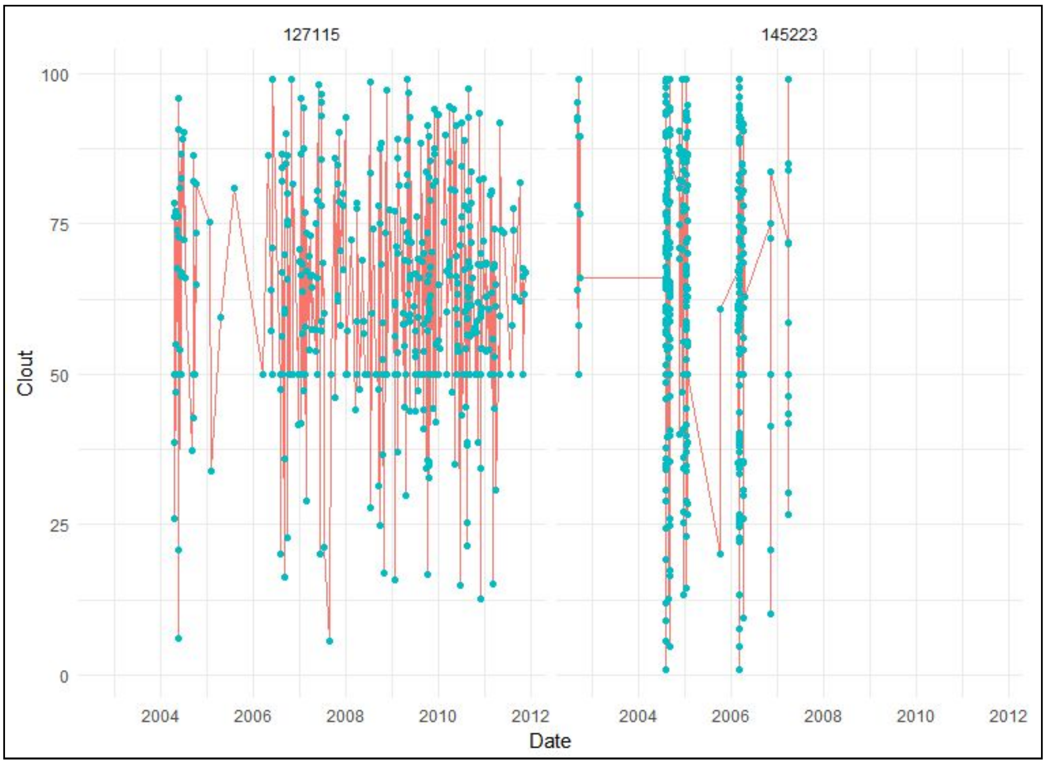


Fig 5.0: Clout Yearly Time Series (ThreadID 127115 and 145223)

Fig 5.0 was created using *'ggplot2'*. The yearly time series was grouped by *'ThreadID'* in the subset shown in *Fig 3.3.*

The yearly plot for '*ThreadID 127115'* spans over a period of 7 years from 2004 to 2011. Over the years, fluctuation can be seen in the graph going as low as 5 and as high as 99. However, the average observation for *'Clout'* ranges approximately between *50% - 80%*. This indicates that the fluctuations may be due to some erratic posts which affects the overall tone of the thread. The majority of the peaks and troughs lie above 50 showing that every author is highly invested and interested in the posts that he/she is writing.

Similarly, the yearly plot for *'ThreadID 145223'* spans over a period of 5 years from 2002 to 2007. Over the years, fluctuation can be seen in the graph going as low as 1 and as high as 99. However, the average observations for *'Clout'* ranges approximately between 50% to 99%. The majority of the peaks and troughs lie above 50 since the graph indicates that the post garners a lot more interest near the end of 2004 until the start of 2006.

Overall for the two threads analysed, we can conclude that pertaining to the high average *'Clout'* value, the authors are highly interested and invested in the ongoing posts in the threads. This shows that their interest in the thread remains at a consistent level with a few minor fluctuations over the years.

Having said that, we cannot conclusively adhere to this analysis since we are only observing two threads which may have some sort of bias to the overall dataset. In addition, this may just be a coincidence as the thread may be a debateable topic. However, to further affirm our analysis, we can apply this technique to a larger number of threads.

Yearly Decomposed Thread

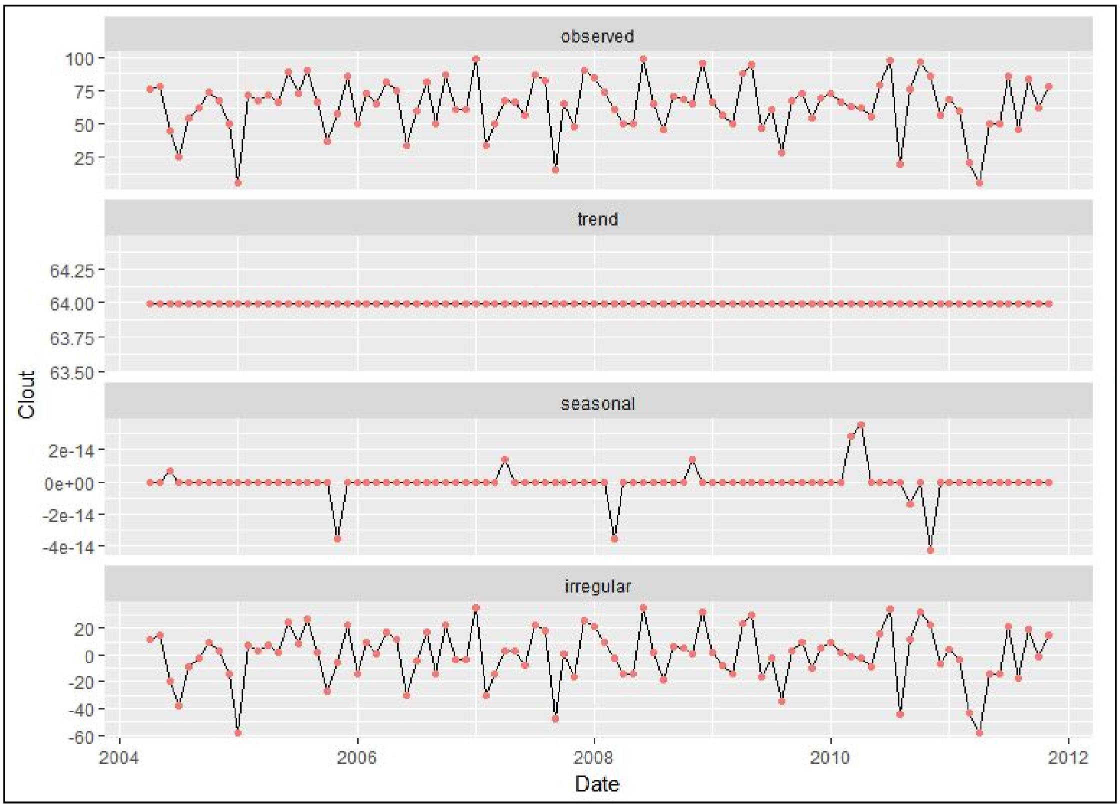


Fig 5.1: Clout Decomposed Yearly Time Series (ThreadID 127115)

Fig 5.1 was created used *'ggseas'*, namely the '*ggsdc' and 'tsdf'* commands.

The trend exists when there is a long term increase or decrease in the data. Since the trend line is a horizontal line, this indicates there is no long term increase or decrease. This means the clout value over the years remains on a consistent level which is in line with our previous analysis that over the years on average the authors are highly invested and interested in the post.

The seasonal trend is an inconclusive as it does not reoccur in a periodic fashion each year indicating that the season has no effect on the interest of the authors on posting on the thread.

Since our irregularity averages between 20 and -20, there is no masking of the trend and seasonal plots. The irregularity results from short term fluctuations in the series which are neither systematic nor predictable as previously mentioned in our analysis. Since the irregularities are short term, they have no effect on the trend and seasonal plots.

Weekly Paired Threads

The second subset used in *Fig 3.4* was to create a weekly plot for two different threads for the linguistic variable *'Clout'*.

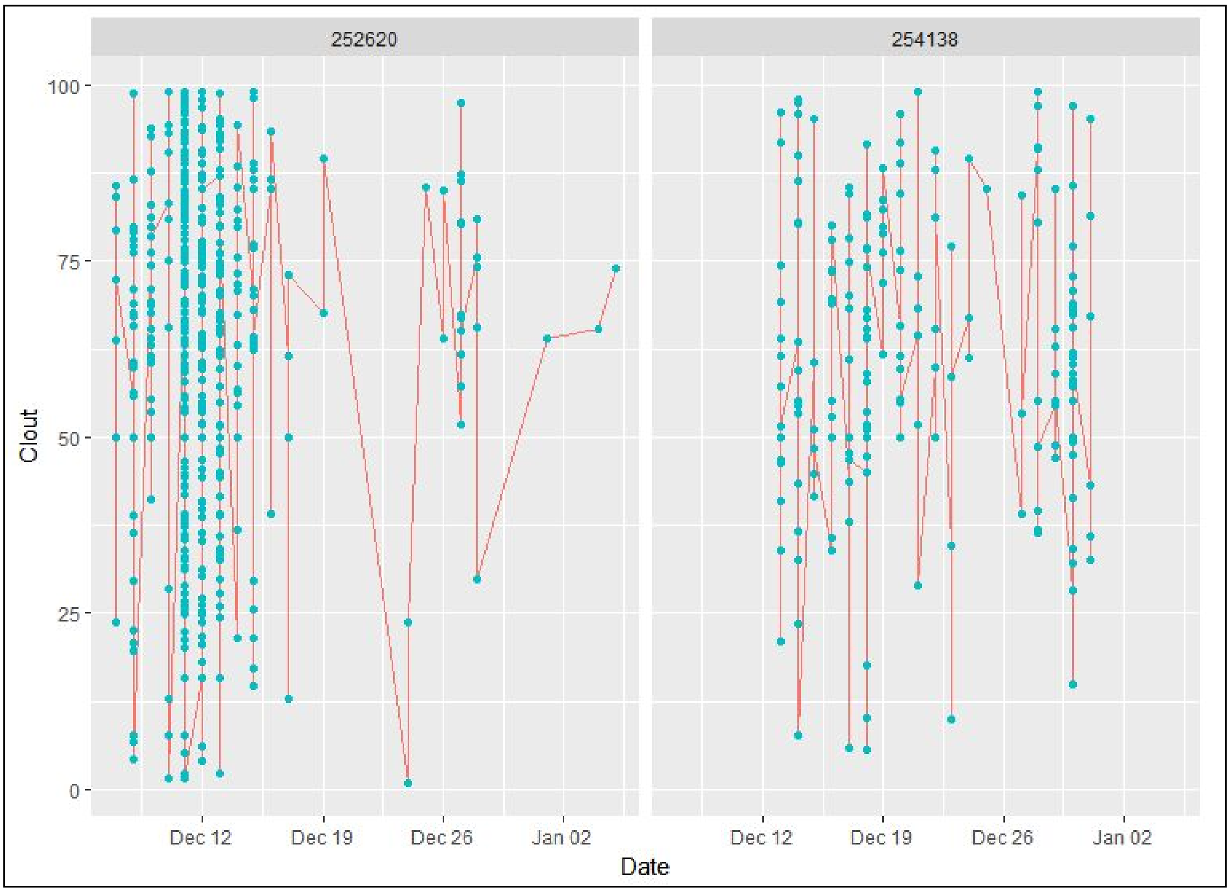


Fig 5.2: Clout Weekly Time Series (ThreadID 252620 and 254138)

Fig 5.2 was created using *'ggplot2'*. The weekly time series was grouped by *'ThreadID'* in the subset shown in *Fig 3.4.*

The weekly plot for '*ThreadID 252620'* spans over a period of 4 weeks from Dec 7 2005 to Jan 02 2006. Over the observed four week span, fluctuation can be seen in the graph going as low as 7 and as high as 99. The fluctuation was of a substantial degree indicating that the interest of the authors was mixed for this specific thread. The peaks and troughs lied throughout the graph below 50 and above 50. Therefore, we cannot conclusively predict the impact of the *'Clout'* data collected for this thread.

Similarly, the weekly plot for '*ThreadID 254138'* spans over a period of approximately 3 weeks from Dec 13 2005 to Dec 31 2005. Over the weeks, fluctuation can be seen in the graph going as low as 1 and as high as 99. The fluctuations are consistently occurring throughout the 3 weeks with data being spread evenly over time. This indicates that some of the authors may be deeply interested in the thread whilst others not so much. Thereby, this gives us an average of 50% *'Clout'* resulting in divided use of strong language. Therefore, we cannot conclusively predict the impact of the *'Clout'* data collected for this thread since the authors have evenly divided share of using strong language.

Having said that, we cannot conclusively adhere to this analysis since we are only observing two threads which may have some sort of bias to the overall dataset. In addition, this may just be a coincidence as the thread may not be an interesting or debateable topic for all authors involved. However, to further affirm our analysis, we can apply this technique to a larger number of threads.

Analysis Insights

From our constructive analysis of the web forum data, we have concluded a few key points:

1. On average, the authors interacting on a specific thread tend to use similar language.
2. To further affirm this observation, we need to test this technique using different linguistic variables on multiple threads.
3. Some threads observed show that, over time, the strong language (*Clout*) used by the authors remains consistent on a high level, indicating that the author is deeply invested and interested in the post.
4. However, some threads observed show that, over time, this is not the case as the fluctuations are substantial i.e. the use of strong language *(Clout)* by the authors is evenly spread across ranging from high to low.
5. Therefore, to successfully conclude whether the effect of strong language remains consistent, increases, or decreases over time, we need to apply this technique on multiple threads.

Appendix

R Code

R code used for the analysis is in the appendix folder.

Member Contribution

Member / Task	Bazil M. Kotriwala	Siddharth A. Shinde	Total
Preliminary Analysis	50%	50%	100%
R research and coding	50%	50%	100%
Preparation of graphics	50%	50%	100%
Analysis of results	50%	50%	100%
Writing up the report	50%	50%	100%