Online Forum Data Analysis

Team Members:
Suu Le Min, Li Jie Lin, David Elcheikh, Jess Nalder

Introduction

There is a theory in social science that people adopt similar patterns of language use when they interact. In this report, we will investigate if this concept is evident using data from an online forum. Some preliminary analysis is first done to determine the predictors of which we would use to analyse and justify our data. After which we'll discuss the following topics. Firstly, we will find out if the sentiments expressed in language are different between threads. Secondly, this report will investigate if the language used by the most active members are different to that used by other members who aren't as active.

Data tidying

Tidying the data involves taking out irrelevant, inaccurate and incomplete data that may impinge our analysis. It also involves arranging our data to be more perceivable for our analysis. Steps taken to tidy our data includes:

1. Swapping 2 columns: PostID and ThreadID
2. Sort according to ascending ThreadID
3. Remove rows that have -1 as author ID
4. Remove posts with less than 20 word counts

Analysing general trends of LIWC predictors over the years

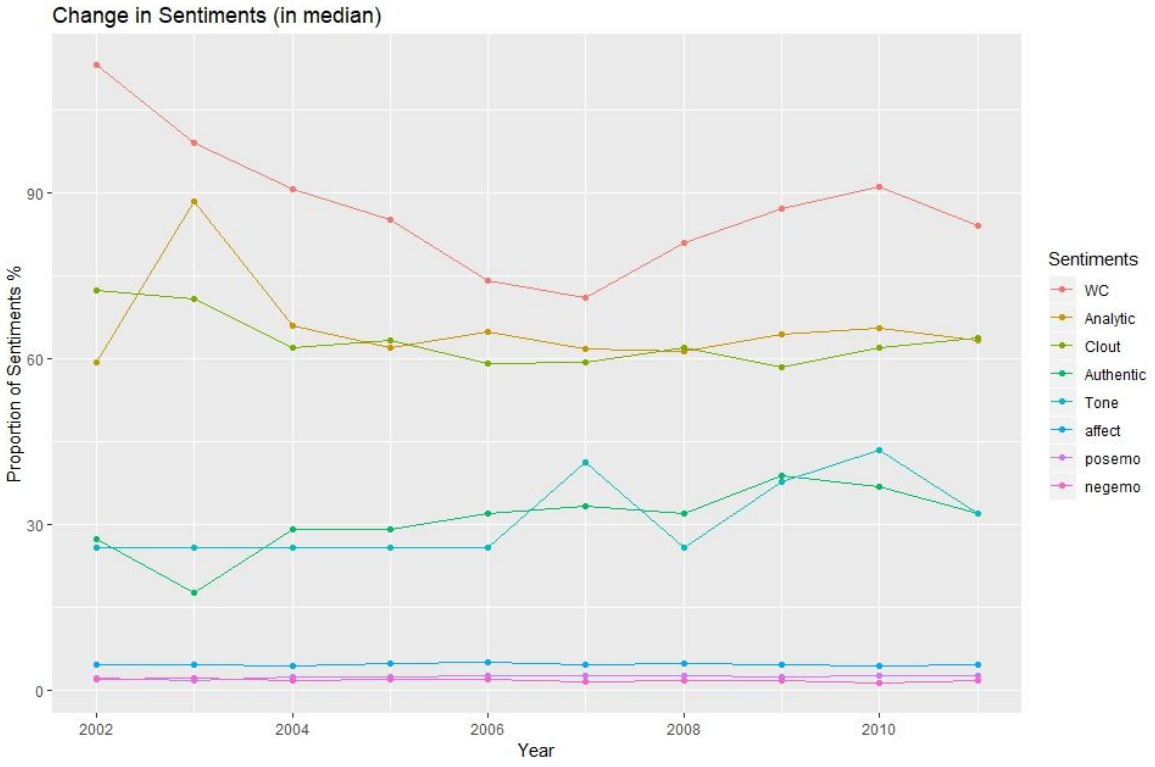

Fig 1. Overall trends of some predictors

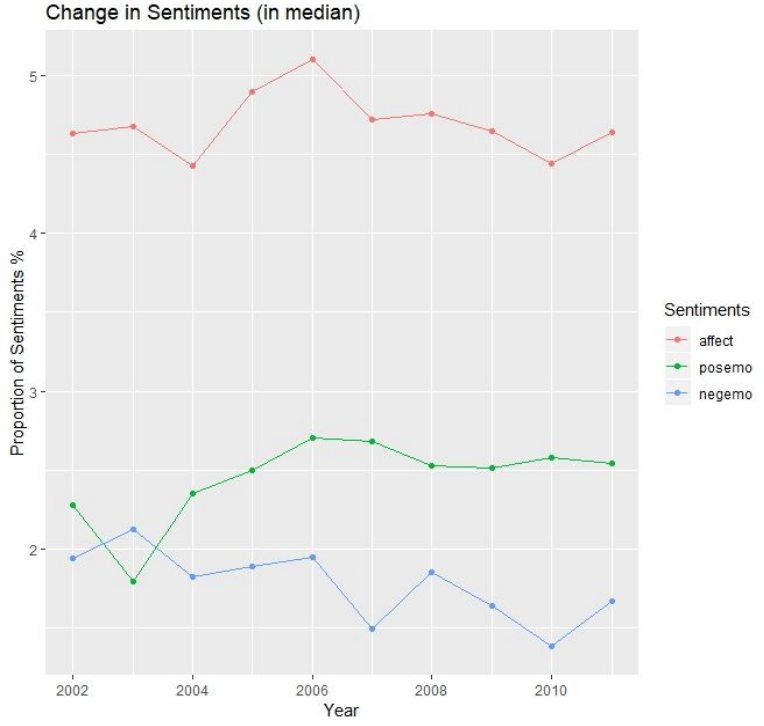

Fig 2. Closer overview of trends for 3 predictors

Fig 1 and Fig 2 display the change in median values of some key indicators affecting sentiment and language between threads over the years.

The dominant indicators of language in general are analytic (analytical thinking) and clout (power/force). Authentic and Tone have less of an impact as the general percentage of words within each post which have either emotional or authentic tone is lower when compared to analytical thinking or powerful language.

As clearly displayed in the graph, over time these indicators are very stable, as the average variance between years for each indicator would be around 5%. Although there are occasional rare dips or spikes in these percentages for each indicator, which could indicate that posts between years are starting to use a more emotional and authentic tone. We also see a spike in the use of analytic language over the period 2002-2004, where it jumped to a maximum average of just below 90% in 2003.

Because we are observing median of these indicators between the years, although small, this would give an indication that different amounts of tone, power and analytic language are being used in groups. So we might expect to see different sentiments being expressed in the language of these threads when we compared them over time.

Are the sentiments expressed in language different between groups, for example, the proportion of language expressing optimism? Does this change over time?

Approach to analysis

Group variables by:

- o ThreadID
- o Year

Correlation matrix of LIWC

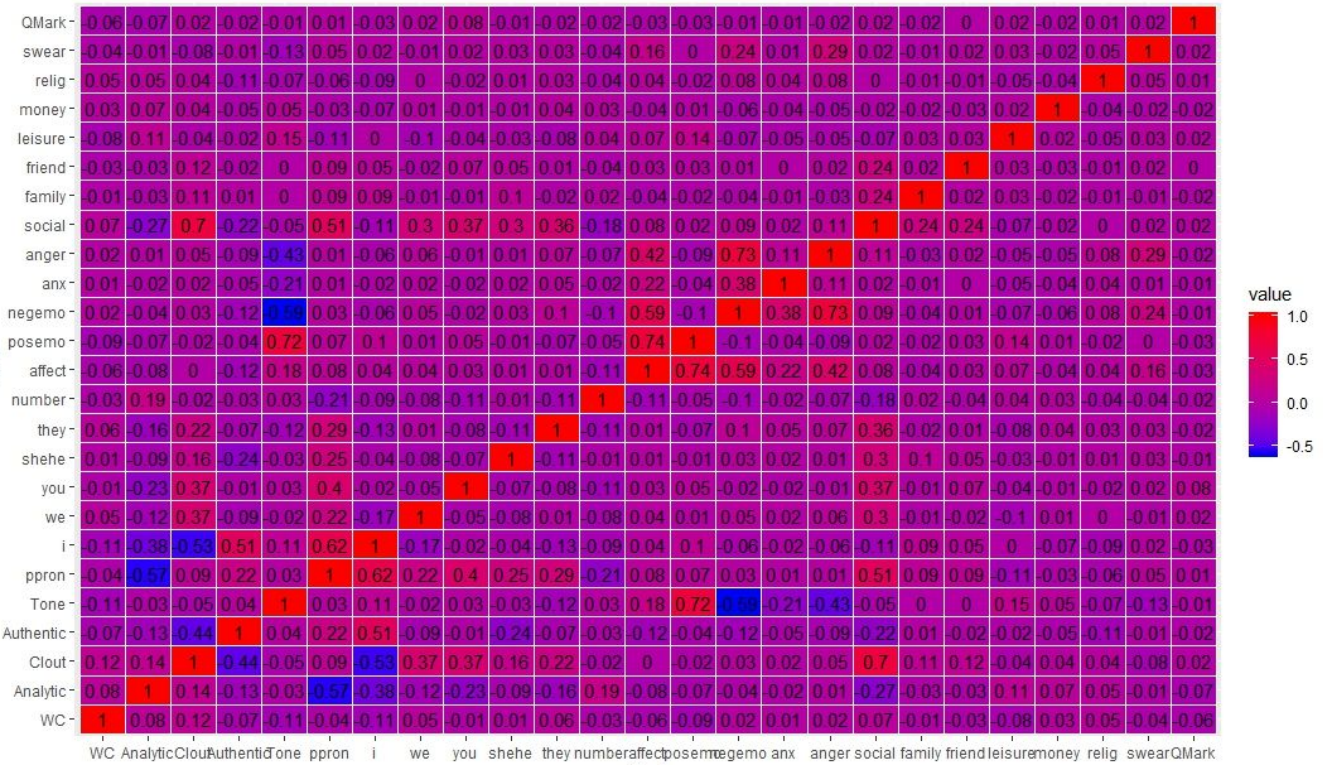

Fig 3. Correlation Matrix

From the above graph, we can see the the liner line of red blocks indicate the there is a strong relationship between the the attributes.

The scattered dark blue blocks indicate that there is also a strong relationship between the attributes. For example, Tone and negemo (negative emotion) have a strong relationship) as these attributes are directly correlated within the threads.

Since we want to investigate if sentiments have changed over the years, we have decided choose predictors that have high correlation coefficient with the predictor 'affect'. Its correlation coefficient with posemo is 0.74, with negemo is 0.59 and with anger is 0.42. To ensure accuracy of the predictors affecting sentiments, we perform 2 regressions on all the threads to check the significance of those predictors. The first regression model on affect against posemo, negemo and anger, and the second regression model on affect with all the predictors except Time. The results show that its p-values of 3 predictors(posemo, negemo, anger) is highly significant as shown in Fig 2, but its R-squared value(0.9911) is slightly smaller than the model with more predictors, shown in Fig 3. The second model has a R-squared value of 0.9912, and a few predictors such as social, family and authentic have significant p-values. Hence, we will use the following predictors, affect, posemo, negemo, authentic, social and family for the first part of our analysis.

```
Call:
lm(formula = webforum$affect ~ webforum$posemo + webforum$negemo +
    webforum$anger)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1652 -0.0624 -0.0562 -0.0524  8.6345

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)     0.0508693  0.0046603   10.916   <2e-16 ***
webforum$posemo 1.0010563  0.0009402 1064.743   <2e-16 ***
webforum$negemo 1.0002368  0.0016297  613.759   <2e-16 ***
webforum$anger  0.0045642  0.0024113    1.893   0.0584 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.315 on 15794 degrees of freedom
Multiple R-squared:  0.9911,    Adjusted R-squared:  0.9911
F-statistic: 5.866e+05 on 3 and 15794 DF,  p-value: < 2.2e-16
```

Fig 4. Linear model with 3 predictors

```
Call:
lm(formula = affect ~ . - Time, data = webforum)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1985 -0.0739 -0.0569 -0.0374  8.6000

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.992e+00  1.624e+01   0.123 0.902376
ThreadID      -4.641e-08  2.743e-08  -1.692 0.090707 .
PostID        -1.185e-08  7.115e-09  -1.665 0.095950 .
AuthorID      -3.127e-08  6.088e-08  -0.514 0.607487
Date           4.534e-05  3.019e-05   1.502 0.133121
WC             1.977e-05  1.634e-05   1.211 0.226105
Analytic      -8.405e-05  1.330e-04  -0.632 0.527298
Clout         -2.072e-04  2.294e-04  -0.903 0.366315
Authentic     -2.634e-04  1.076e-04  -2.447 0.014419 *
Tone          -3.285e-06  1.607e-04  -0.020 0.983694
ppron          2.435e-01  4.246e-01   0.573 0.566405
i             -2.424e-01  4.246e-01  -0.571 0.568062
we            -2.506e-01  4.246e-01  -0.590 0.555118
you           -2.488e-01  4.246e-01  -0.586 0.557848
shehe         -2.492e-01  4.246e-01  -0.587 0.557217
they          -2.465e-01  4.246e-01  -0.581 0.561545
number        -3.331e-03  8.618e-04  -3.865 0.000111 ***
posemo         1.001e+00  1.683e-03 594.573  < 2e-16 ***
negemo         9.987e-01  2.272e-03 439.675  < 2e-16 ***
anx            2.392e-03  4.050e-03   0.591 0.554801
anger          4.513e-03  2.550e-03   1.770 0.076821 .
social         4.967e-03  1.087e-03   4.571  4.9e-06 ***
family        -9.164e-03  2.528e-03  -3.625 0.000290 ***
friend         1.251e-03  3.088e-03   0.405 0.685319
leisure       -3.396e-03  1.260e-03  -2.695 0.007046 **
money          1.899e-03  2.104e-03   0.903 0.366580
relig          1.791e-03  1.831e-03   0.978 0.328077
swear         -5.963e-03  3.587e-03  -1.663 0.096428 .
QMark          5.435e-05  1.018e-03   0.053 0.957408
Year          -1.231e-03  8.229e-03  -0.150 0.881133
TimeCategory2 -5.580e-03  1.061e-02  -0.526 0.599042
TimeCategory3 -1.200e-02  9.643e-03  -1.245 0.213296
TimeCategory4  1.254e-03  9.640e-03   0.130 0.896526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3142 on 15765 degrees of freedom
Multiple R-squared:  0.9912,    Adjusted R-squared:  0.9911
F-statistic: 5.526e+04 on 32 and 15765 DF,  p-value: < 2.2e-16
```

Fig 5. Linear model with all predictors except time

Are the sentiments expressed in language different between threads?

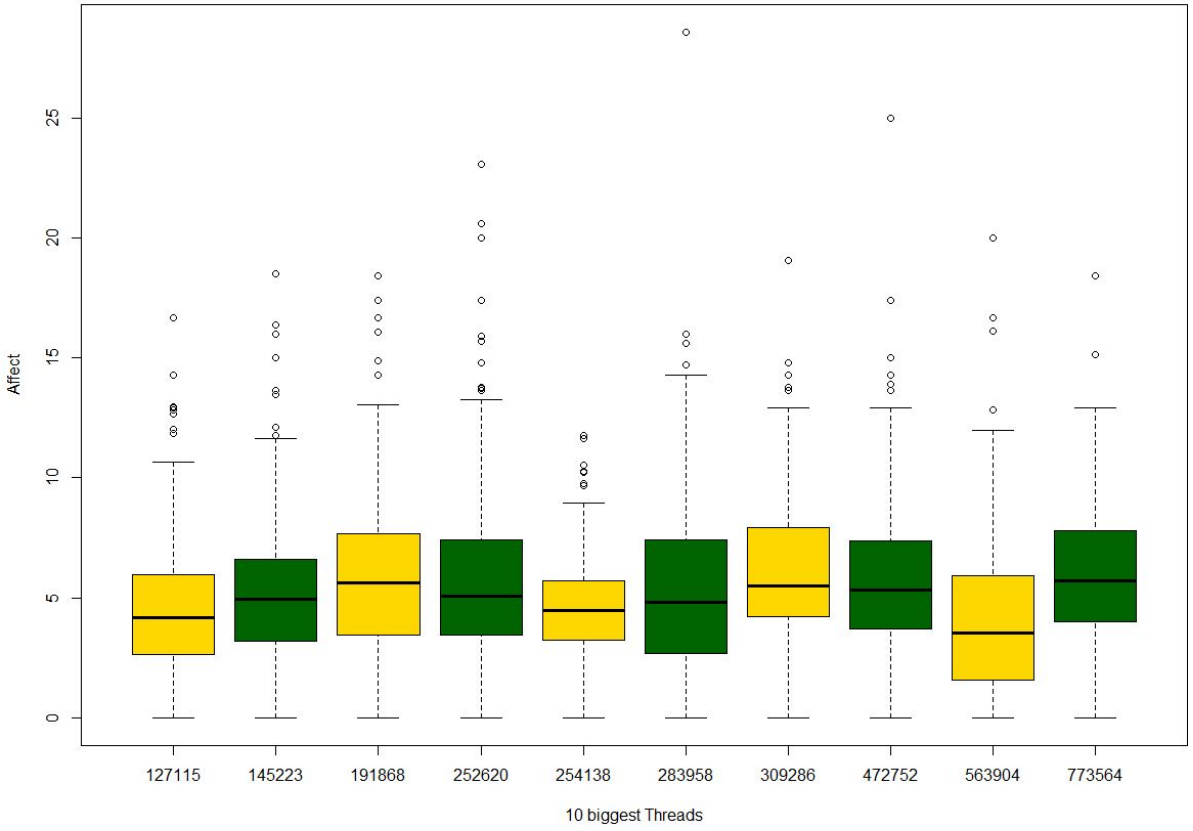

Fig 6. Boxplot of 10 biggest threads

We extracted a sample of 10 threads with the most number of posts and created a box plot(Fig 6) plotting against the 'affect' predictor. Results are as shown below.

- The distribution of values are similarly clustered around the same range.
- All threads have the same minimum value of 0 and similar maximum values.
- every thread contains outliers.
- Median values are similar, suggesting that sentiments expressed in language are similar between the top 10 threads with the most number of posts.
- We can also see that the average amongst the 10 biggest threads would correlate amongst the other remaining threads with analysis of the threads showing that the 'affect' predictor has a similar proportion of sentiments regardless of thread.

Top 10 Biggest Threads vs Remaining Threads

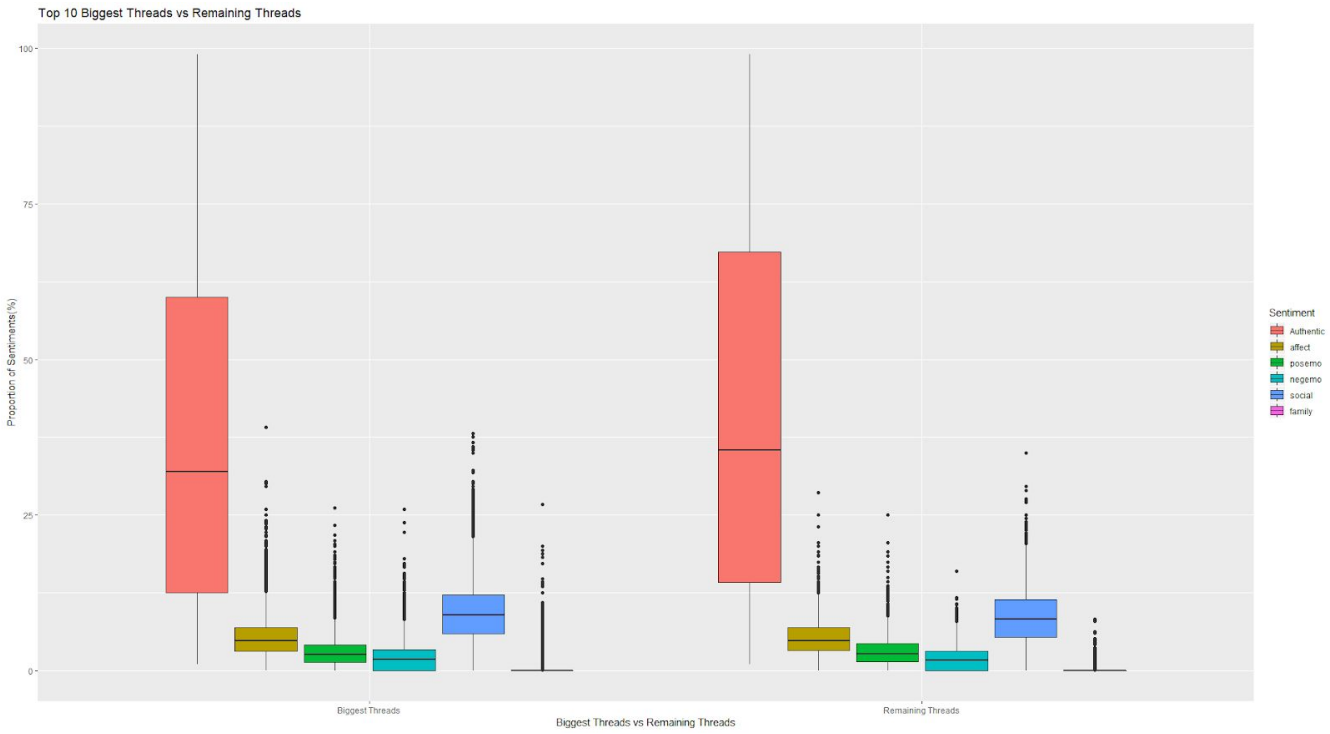

Fig 7. Boxplot of 10 biggest threads vs remaining threads

The above graph(Fig 7) displays the percentages of sentiments between the top 10 biggest threads vs the remaining threads from the webforum. As demonstrated, the sentiments relating to tone i.e. positive and negative emotions remain consistent between these two sub set groups.

The remaining threads would more likely result in a higher percentage of authentic sentiments as these are more or less based on a word count. Based on the analytics of the data, we can see through many other forms of graphs that emotional sentiment remains consistent no matter what the subset rules may be.

Authentic and Tone differ very minimally between the subsets, with the top ten biggest threads incurring an approximate ~ ± 20% difference in Authentic sentiment between the threads, since Affect has no outliers.

As mentioned above, Affect and Tone don't have the impact on threads as much as analytical and powerful languages as these express views and feelings that may not specifically relate to a thread (i.e. relate to the specific user)

As mentioned above, Affect and Tone do not have any impact on threads as much as analytical and powerful languages as these express views and feelings that may not

specifically relate to a thread (i.e. relate to the specific user). Amongst the biggest threads, we can see through the outliers that affect has a higher proportion of sentiment outside the median value. This results due to the fact the bigger threads have a disparity of different users within them, which can then cause outliers as shown. This trend mirrors onto the remaining threads with fewer outliers, suggesting the smaller the thread, the less disparity in sentiment.

With regards to positive and negative emotions, we can see that the data from the top 10 threads compared to the remaining threads are almost identical. Similar to Affect, tone results in more outliers amongst the bigger threads as they incur higher word counts with more users. This increases the chances of more emotional sentiment (although still a low proportion) amongst these threads. Comparing this to the remaining threads, we can also see that given any subset of the main data, tone remains a consistent proportion of sentiment throughout the threads.

Trend of positive emotions grouped by the 10 biggest threads over the years

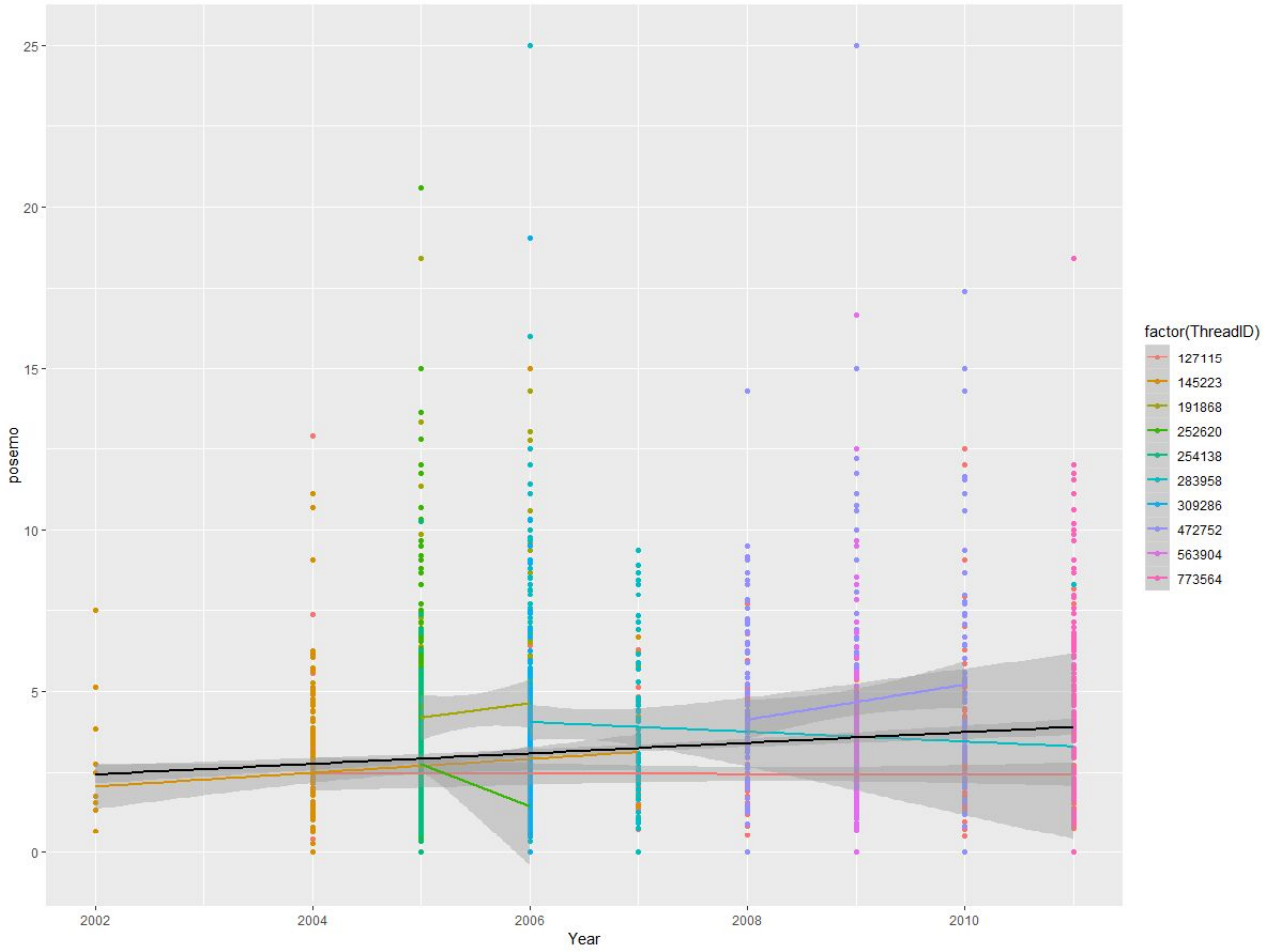

Fig 8.

Based on our preliminary analysis, we identified that the indicator posemo has a significant influence on that of the indicator affect (which is what we're trying to observe). Therefore, for the above graph we compared the amount of positively emotive language per post for the 10 biggest threads as the years progressed. We wanted to spot any trends in the change of sentiments as time progresses.

Clearly displayed in the graph is the distinct color of each column. This is because when each of these main threads were created/active, a large majority of their posts were in that given year. This helps to show us the difference in amounts of positive emotion shown for each post within the thread.

For each of the threads we can see that some have posts with a significantly higher use of positively emotive language than that of other threads. The orange thread (ID:

145233) for example has clusters its data points towards the bottom of the graph, whereas the green and blue threads (ID's: 252620 & 309286) seem to have a cluster of posts with a higher use of the language.

Another key indicator of a trend in the average use of positively emotive language are the regression lines for each thread. The fact that each of these lines are in general quite different to one another would imply that the use of positively emotive language tends to differ quite obviously between threads. The black line indicates the regression and overall trend of all the 10 threads combined, despite the fact that some threads have a decline in positive emotion over time, there is a slow but constant increase in trend throughout the years.

In conclusion, this graph would indicate that the expression of sentiment are not only visibly different between threads, but also are also seeing different trends relating to the expression of sentiment as well. This means that users are posting more positive posts as time passes.

Is the language used by the most active and/or socially connected members of the forum different to that used by the other members? Does this change over time?

Approach to analysis

Group A - The most active and socially connected members of the forum

We first took the subset of the top 20 biggest threads by postID to determine the most active threads, of which we then found the 10 biggest threads by authorID in that subset to determine the most active and socially connected members of the forum

Group B - Remaining members

Hypothesis Testing

We will perform t.test with 95% confidence interval.
Rejection criteria: p-value <= α (level of significance, 0.05)

Analytic -
Null hypothesis: Group A has higher analytically values than group B.
Alternative hypothesis: Group A does not a have higher analytically values than group B.

Clout -
Null hypothesis: Group A has higher influential power than group B.
Alternative Hypothesis: Group A does not have higher influential power than group B.

Authentic -
Null hypothesis: Group A has higher authentic tone of voice than group B.
Alternative hypothesis: Group A does not have higher authentic tone of voice than group B.

Tone -
Null hypothesis: Group A has higher emotional tone than group B.
Alternative hypothesis: Group A does not have higher emotional tone than group B.

Attribute	p-Value	Conclusion
Analytic	4.034×10^{-11}	Do not reject
Clout	0.2667	Reject
Authentic	0.9684	Reject
Tone	3.14×10^{-6}	Do not reject

Proportion of language predictors over time

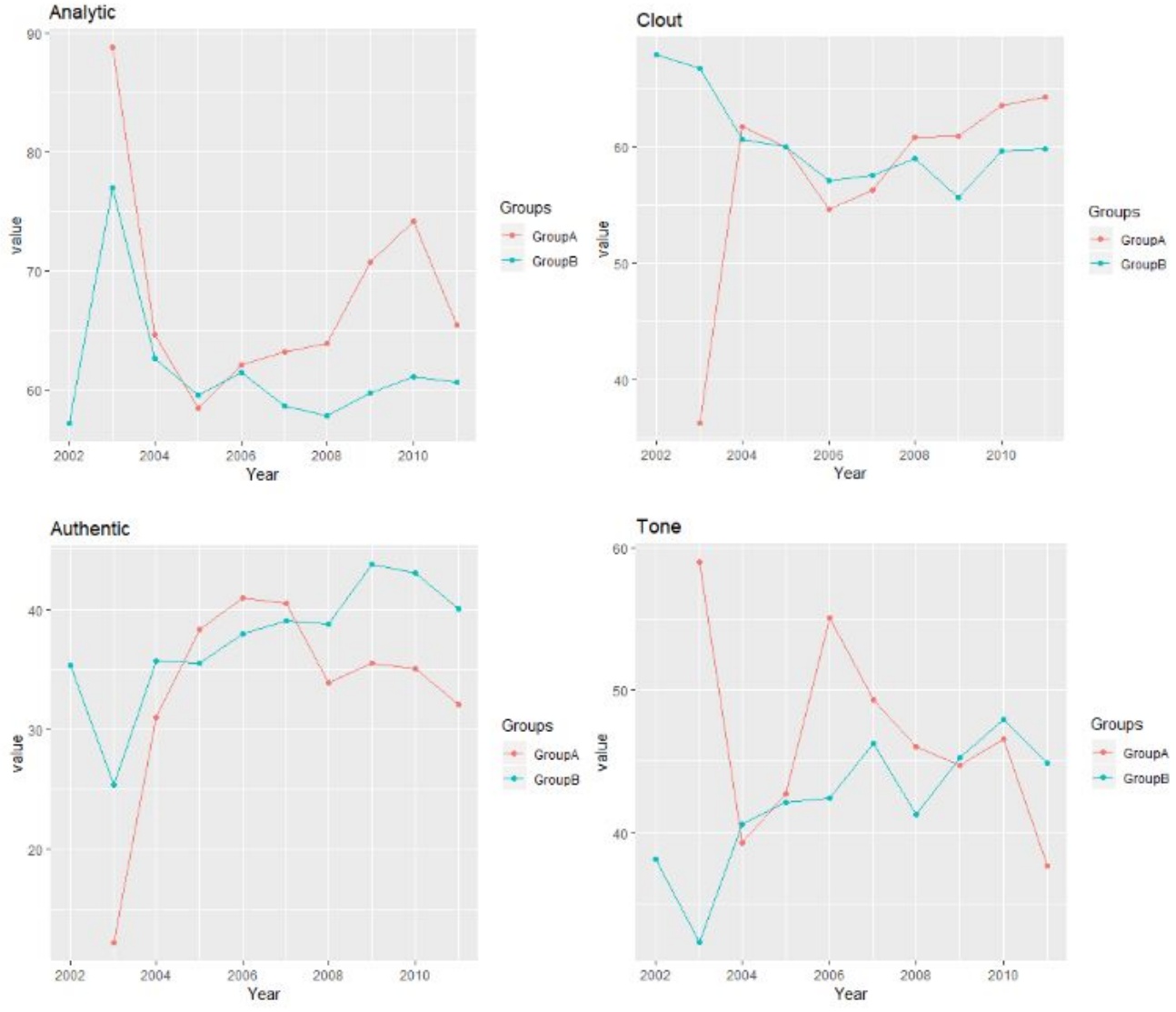

Fig 9. Change in mean values over the years

The multifaceted graphs above show the progression of the indicators; Tone, Authentic, Clout and Analytic respectively over a timeframe of a decade with yearly intervals. Using the multiple facets to display the different indicators, rather than displaying all indicators in the one graph makes it easier to observe the trend for that indicator and the difference between groups.

Analytic: The analytic graph at first shows a very similar trend between the groups, but a clear difference in trend between the groups emerges after 2006.
Clout: For clout, we initially see a quite different trend between the groups which quickly synchronises. But as time progresses we begin to see a growing difference between the groups.
Authentic: Similar to that of the analytic graph, at first the groups show a very similar trend to one another, but at 2007 for 1-2 years we see them separate before synchronising again at the end on different levels.
Tone: The overall trend of the use of emotional tone is quite different between the groups. Although they pass one another a number of times, the trend of either group never aligns, and both tell different stories.

Group A is visibly more erratic in relation to Group B because it is based off of a smaller dataset. So the average use of each of the indicators is likely to see a much larger variance between the years than that of Group B, as Group B's outliers will have less of an impact compared to Group A's outliers. This means that although level at which each indicator is used is important, when comparing the two groups, trend will provide a better insight into any signs of difference between the groups.

Although the trends of Group A & B are very similar for all four of the analysed indicators. Because at one point for each of the graphs we can see different patterns between the groups, it indicates that the use of that indicator for either group is quite different. Because of our preliminary analysis of the indicators, we've found that these indicators have the highest influence on the language of either group. So the results we have identified from the graphs above would indicate that there is a difference of language between the two groups.