FIT3152: Data Analytics

Assignment 1

“Exploring the Social and Linguistic Dynamics of an Online Community”

Ya Elaine Gao - 29808448

Loi Chai Lam - 28136179

James Leslie Akerman - 20301693

Huu Loc Le - 25187198

Tables of Contents

Introduction

The purpose of this report is to analyse the language in an online forum, with specific focus on different groups within this forum and how the use of language changes over time. The investigation was split into two parts to ensure that both questions were answered thoroughly the two questions provided in the assignment brief.

Pre-processing of the data

A major issue encountered in the early stages of exploratory analysis on the data was that the noise from uninformative data. There were a large number of observations where the word count was very low or even zero in some cases. Though it was obvious that the rows with zero values for word count should be removed from the data, it was difficult to decide on an acceptable word-count; we did not want to jeopardize the data's statistical integrity by removing an excessive amount of rows. A minimum word count of 25 was eventually decided on after consulting a study found in the LIWC2015 Development Manual. This study only used corpora with a minimum of 25 words.

Part One

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

Approach

The approach taken to analysing the sentiments expressed in language in the data was as follows included the following steps:

1. Identifying a broad subset of the data.
2. Identifying narrower subsets of data.
3. Sentiment Analysis according to Year and Time of Day.
4. Sentiment Analysis according to Year; Time of Day; and ThreadID.

1. Identifying a broad subset of the data

In order to make analysis easier, we created a subset of the data. We selected threads which had >=35 rows. This gave us 213 threads and 13295 rows. We chose 35 because this was the mode number of rows for any thread (5% of all threads) in our cleaned data. This number was also judged to be a sufficient minimum row-count for us to conduct meaningful analysis on threads. Through the use of boxplots, we found that the most active threads were the threads from 2005 until 2011 (201 out of 213).

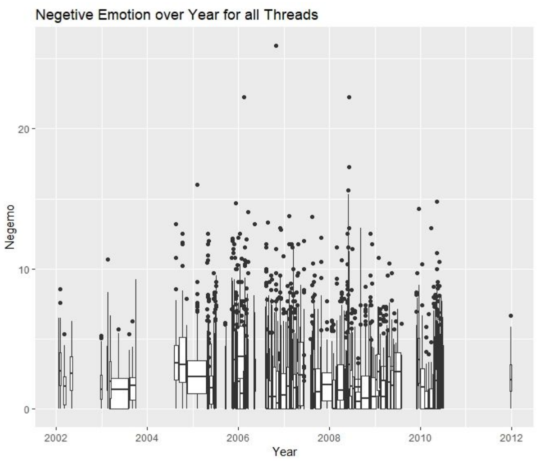


Figure 1.1

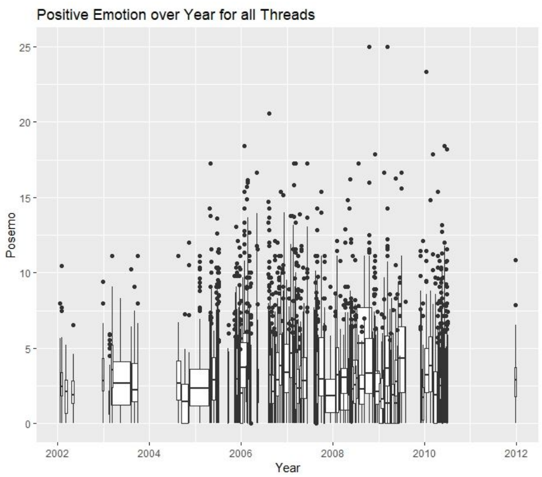


Figure 1.2

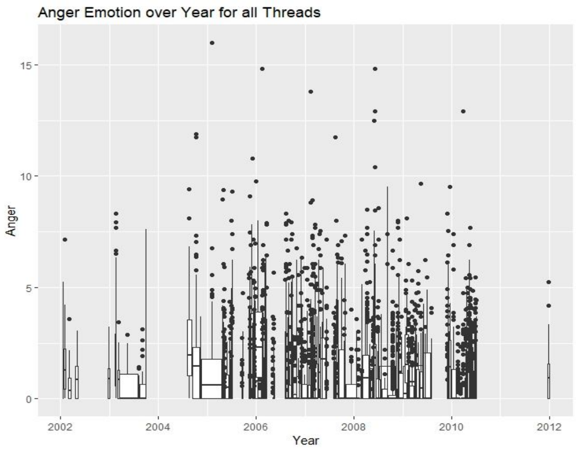


Figure 1.3

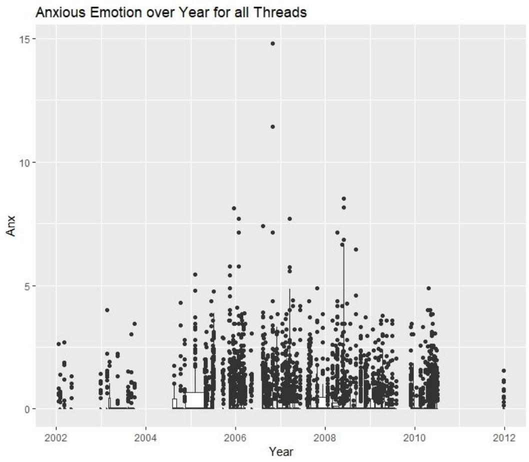


Figure 1.4

2. Identifying narrower subsets of data

In our approach to analysing our subset of data, we decided to further break down our subset according to different years; different times of day; as well as different threads.

Dividing the data into different years allowed us to compare sentiment across large periods of time. We wanted to see if sentiment varied between years, such as if some years showed a greater proportion of positive sentiment than other years.

Further dividing the data into different time blocks allowed us to see if the time of day affected the sentiment of the data in anyway. For the sake of simplicity, we choose to analyse two major time-blocks: 12am to 12pm; and 12pm to midnight.

3. Sentiment Analysis according to Year and Time of Day.

The two main statistical tools that we used in order to analyse the data were the medians and correlations.

Median

When comparing sentiment proportion across years and across times-blocks, we chose to use the median. We did this for two main reasons.

1. Our data subsets were not of equal sizes; for example, some years and threads had more posts than others.
2. To reduce the influence of sentiment proportion outliers on our analysis results.

Correlation

We used correlation in order to test if there were relationships between different types of sentiment. We would usually only accept correlations that were statistically significant ($p < 0.05$). However, the degree to which the p-value fell under the 0.05 limit was also taken into consideration.

Positive Emotional Sentiment/posemo (See Appendix 1)

Our data showed mild posemo variation across different years as well as different times of the day.

The 2006 year showed the highest posemo proportion (2.68%) on a year to year basis. The 2006 year was also the same year when morning posts (12am to 12pm) showed the highest posemo proportion (2.65%), and when afternoon posts (12pm to midnight) showed the highest posemo proportion as well (2.73%).

Overall, afternoon posts were slightly more positive than morning posts: six out of the seven years analysed showed higher posemo proportions in afternoon posts than in morning posts. Analysis of all data from the 2005-2011 time period showed the same pattern: afternoon posts had a 2.61% posemo proportion, and morning posts had a 2.53% posemo proportion: a difference of 0.08%.

Negative Emotional Sentiment/negemo (See Appendix 2)

The data also showed mild variation in negemo between both years as well as different times of the day.

The 2005 year showed the highest negemo proportion (1.97%) on a year to year basis. The 2005 year was also the year when morning posts were the most negative (2.02%), and when afternoon posts were the most negative (1.92%).

Overall, morning posts were slightly more negative than afternoon posts: six out of the seven years analysed showing higher negemo proportions in morning posts than in afternoon posts. Analysis of all data from the 2005-2011 time period showed the same pattern: morning posts had a 1.85% negemo proportion, and afternoon posts had a 1.72% negemo proportion: a difference of 0.13%.

4. Sentiment Analysis according to Year; Time of Day; and ThreadIDs.

Positive Emotional Sentiment/posemo (see Appendix 3)

While the 2006 year showed the highest posemo proportion on a year by year basis, the 2007 year actually contained the most overall positive thread: thread 296985 with 7.69%. Interestingly, thread 296985 was also the thread with the highest posemo proportion in the afternoon. In six out of seven years the threads with the highest overall posemo proportion, were also threads with the highest posemo proportion in the afternoon.

Negative Emotional Sentiment/negemo (see Appendix 4)

While the 2005 year showed the highest negemo proportion on a year by year basis, the 2010 year actually contained the most overall negative thread: thread 539505 with 9.68%. Interestingly, thread 539505 was also the thread with the highest negemo proportion in the morning. In six out of seven years, the threads with the highest overall negemo proportion were also threads with the highest negemo proportion in the morning.

Another interesting observation is that thread 179689 showed the highest negemo proportion in the morning in the 2007 year, and also the highest posemo proportion in morning in the 2006 year.

Correlation between Positive and Negative Emotional Sentiment

We wanted to see if there was any relationship between the posemo and negemo proportions in the data. Our null hypothesis was that would be no correlation between these two sentiments (a correlation of zero). Our findings showed that there was in fact a statistically significant negative correlation between posemo and negemo proportions in the data, albeit a very very weak one: -0.09920716.

Table 1.1 shows that this correlation changed very slightly depending on the year: the 2005 and 2009 years showed the strongest correlation between these two sentiments. This correlation is also slightly stronger in morning posts (-0.1034317) than in afternoon posts (-0.0968531).

Table 1.1

Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Correlation	-0.1166316	-0.1016225	-0.09644692	-0.08743787	-0.1147673	Not statistically significant	-0.07680528

Comparison of the Positive Emotion, Negative Emotion, Anxious and Anger between Threads.

To compare the positive emotion (posemo), negative emotion (negemo), anxiety (anx) and anger (anger) between threads, we plotted the boxplots of these sentiments (group by thread).

The thread which have the highest negemo is also the thread which have the highest anx or anger (see Appendix 5). For example, in year 2011, 2010 and 2009, thread 426705, 120790 and 558874 have the highest negemo and highest anger. In year 2008 and 2007, the thread 461357, 179689 and 361963 have the highest negemo and highest anx.

To prove that there is a relationship between the negemo with anger and anx, we conducted two correlation tests for negemo with anger, and negemo with anx (see Appendices 6 and 7).

From the results, we proved that there is a relationship between negemo with anger and anx. The higher the negemo, the higher the anger; the higher the negemo, the higher the anx. To further investigate that the relationship between anger and anx, we conducted another correlation test, this time between anger and anx (see Appendix 8). It showed that the higher the anx, the higher the anger. Thus, there is a relation between anger and anx.

We found that thread 120790 has the highest negemo and anger in year 2010, and highest anger in the 2007 year (see Appendix 6). In the 2008 year, thread 120790 has the highest posemo and anger. We were interested in the relationship between posemo and anger for that thread in the 2008 year. Thus, we conducted a correlation test for the data (see Appendix 9).

The test shows that the correlation between anger and positive emotion (posemo) is -0.8895592, which is the higher the positive emotion, the lower the anger. However, the p-value for the test is 0.04332, which is barely less than 0.05. This gives us the reason for not rejecting the null hypothesis. Therefore, in this case, there is no relationship between the positive emotion and anger for thread 120790 in year 2008.
The boxplots in the Appendix 11 till Appendix 15 shows the results for thread 120790 stated above.

Correlation between Personal Pronoun and Negative Emotion, Positive Emotion, Anger and Anxiety

We investigated the correlation between personal pronoun (we, heshe, they, i and you) and the sentiments (negemo, posemo, anger and anx). We filtered the table and printed the data with the correlation larger than absolute value of 0.5 (see Appendix 10). We observed that for thread 229152, the higher the usage of "i", the higher the posemo; the higher the usage of "we", the higher the anx in that thread. Similarly, for thread 274018, the higher the usage of "we", the higher the posemo; the higher usage of "you", the higher the anger and negemo; the higher the usage of "they", the higher the negemo. Also, for thread 330904, the higher the usage of"i", the higher the posemo; the higher the usage of "shehe", the higher then anx; the higher the usage of "they", the higher the anger. From the above results, we found that there is a relationship between personal pronoun and sentiments between each group. Figures 1.5, 1.6, and 1.7 below show the graph of the correlation for thread 229152, 274018 and 330904.

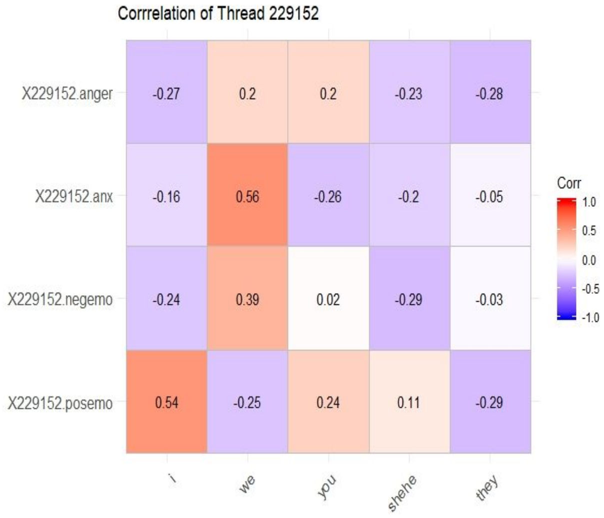


Figure 1.5

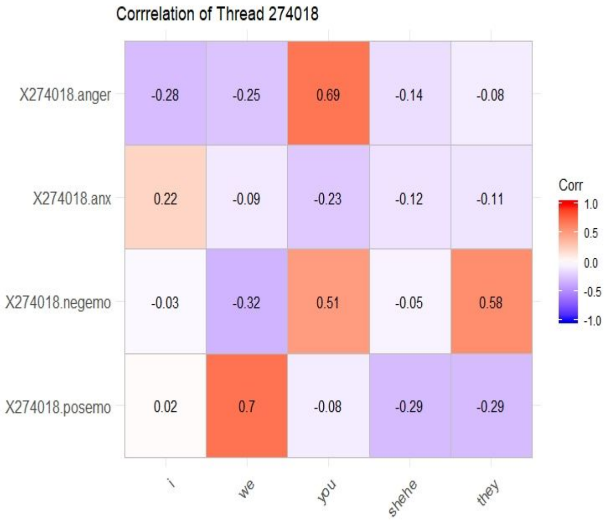


Figure 1.6

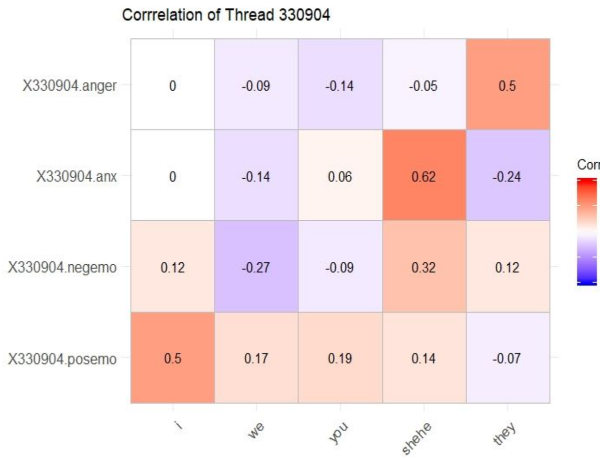


Figure 1.7

Part Two

Analysing the question: "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

The approach taken to analysing the language differences between the most active members of the forum and the other users was as follows:

1. Identifying the most active members
2. Identifying the most important/significant attributes
3. Comparing the groups overall
4. Comparing the groups over time

1. Identifying the Most Active Members

There are 2335 unique users in total (2334 if the anonymous user is discounted). It was decided that the top 1% of users (23 users) would be to classified as the 'most active' group.

There are a number of ways which we can identify the 'most active' users, including taking into account the word count or the number of threads a user is involved in and selecting the most active users from this information. However, a more holistic approach was taken, and the way that the 'most active' users on the forum were identified in this methodology was through finding the number of occurrences (i.e. posts) belonging to each unique AuthorID and selecting the 23 users with the highest count of posts.

2. Identifying the Most Important Attributes

In order to identify the most important or significant attributes, a correlation matrix was plotted (shown in Appendix 2.1), and from this, the top 6 attributes by sum of (absolute) correlation values were extracted and are shown in Table 2.1.

Table 2.1: Key variables based on sum of correlation coefficient

Variable	**Sum of Correlation Coefficient**
Social	4.438
Ppron	4.147
Clout	3.832
Tone	3.261
Analytic	3.051
Authentic	2.747

3. Comparing the Groups

The approach taken to initially compare the language used by the most active members of the forum and that of other members is to compare the ‘average’ values for each attribute for the group we have defined as the most active users and the rest of the users.

The measure of centre used in this case was the arithmetic mean. Though the median is generally preferable to the mean due to decreased sensitivity to outliers, in this case, use of the median returned a large number of 0 values for attributes, making it difficult to compare values from the different groups.

After calculating the mean for each attribute from the two groups, the difference in the values was calculated, then transformed into a percentage to more accurately reflect the differences (as some of the attributes had fairly small values to begin with, which would make them difficult to compare overall)

$$100 * (\mu(\textit{most active users}) - \mu(\textit{other users}))/\mu(\textit{most active users})$$

The results were plotted in the diverging bar chart in Figure 2.1.

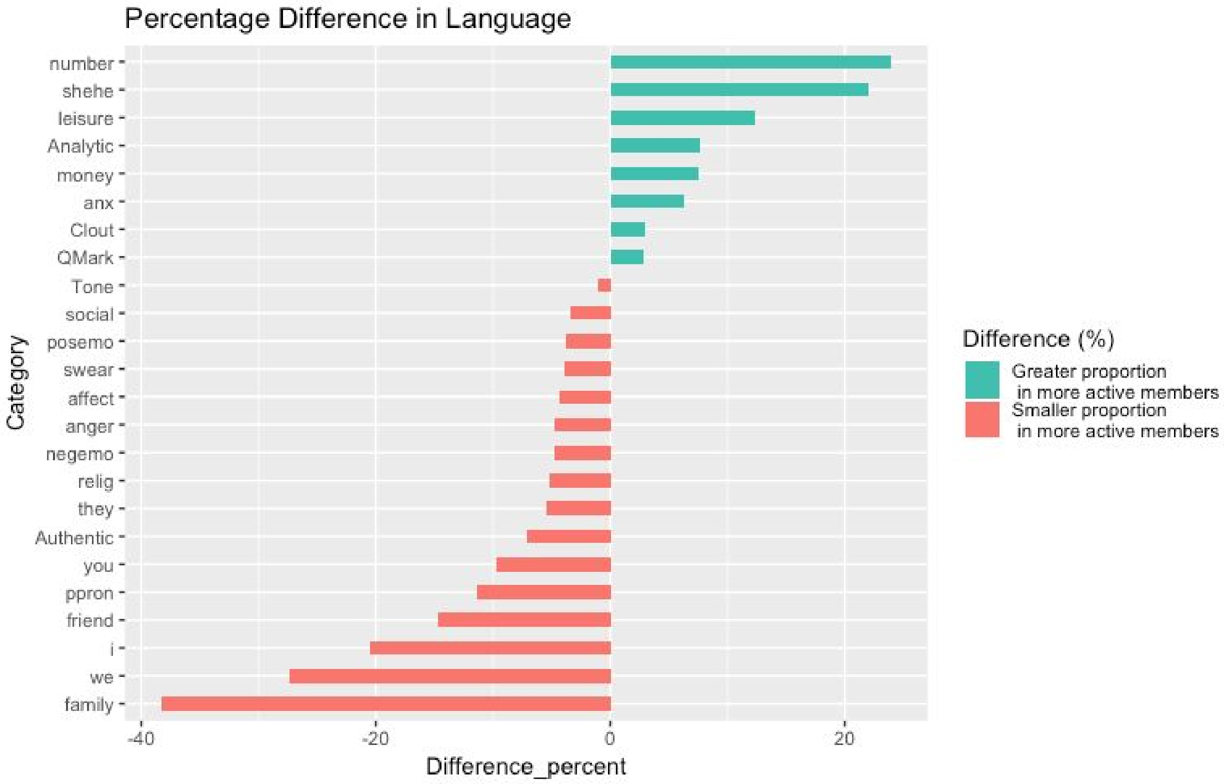


Figure 2.1: diverging bar chart showing the percentage language difference between the most active users in the forum and the rest of the members

We can see from Figure 2.1 that there seem to be some notable differences in the types of language expressed by the two groups. In order to support the analysis, t-tests have been carried out for each of the categories, and this information is included in Appendix 17. First considering the most significant language attributes:

- Social: the percentage of language reflecting on social processes is slightly lower in the most active group compared to the rest of the users (this is supported by a low p-value in the t-test performed). Interestingly, the attributes of family and friend (where language may be similar or overlap with social) are also significantly lower in the most active group.
- Ppron: the proportion of personal pronouns used is also lower in the most active group, and in addition to this, the individual attributes (i, we, you and but not heshe), are also lower (evidence for these in the performed t-tests). The proportion of they is also lower, but not supported by the statistical testing.
- Clout: it seems that the most active members have a slightly higher amount of power or impact compared to the rest of the group (supported by testing).
- Tone: from the chart, the emotional tone seems to be slightly lower in the more active group, however, the information from t-tests performed leads us to reject this claim, so the difference for this attribute is inconclusive.
- Analytic: the most active users are seen to be more analytic, with support from the t-tests. They also use a higher proportion of numbers/quantifiers, suggesting that they are more inquisitive than the rest of the users
- Authentic: the proportion of authentic tone used is lower in the most active users, which is supported by the t-tests

4. Comparing the groups over time

In terms of comparing over time the language used by the most active members and the other members of the forum, we began the exploration of this by plotting the overall percentage difference between the two groups over time (Appendix 26), but the produced plot does not give much insight as to how language changes over time, as language expressed is subject to dramatic changes between threads due to the topic of the thread. So instead of analysing the overall trends, we have identified specific threads to investigate.

We have created a table (Appendix 18 & Appendix 19) to contrast the mean and its correlations between three different years of each attribute over the span from 2002 (or 2004) to 2011. We have chosen 2002 as the start, 2006 (or 2007) as the middle and 2011 end of the year of the dataset so that we are able to see how each of the distinct years changed through time.

In the table of the correlation coefficients between years (Appendix 18), over the nine years, we are able to see that there are generally high correlations between the 2002/2011 period. Before categorising the data by threadIDs, we wanted to discern how the language had generally evolved over time and we expected that there would be some attributes in which there would be some low correlations as norms in society changed. The use of analytical, clout, authentic, emotional tone, personal pronoun and social language showed a high correlation of 0.94, 0.89, 0.89, 0.88, 0.89 and 0.87 respectively. Through this, a hypothesis could be established whereby authors' online use of analytical, clout, authentic, emotional tone and social languages had not changed. On the other hand, leisure and religious language had very low correlations, of 0.51 and 0.30, signifying more change between the two periods (2004/2007 and 2007/2011).

In the correlation coefficient of ThreadID 127115 (Appendix 19(, which represented eight years from 2004 to 2011, we are able to also see similar correlations in language use. The use of analytical, clout, emotional tone, personal pronoun and social language showed a high correlation of 0.99, 0.93, 0.96, 0.92, 0.84 respectively. We are able to see that the authentic language from this thread has lower correlation between 2004 and 2011. Furthermore, there is no almost no (0.06) correlation in swear language between 2004 and 2011.

Time series Graph of LIWC Attributes - aggregated by mean

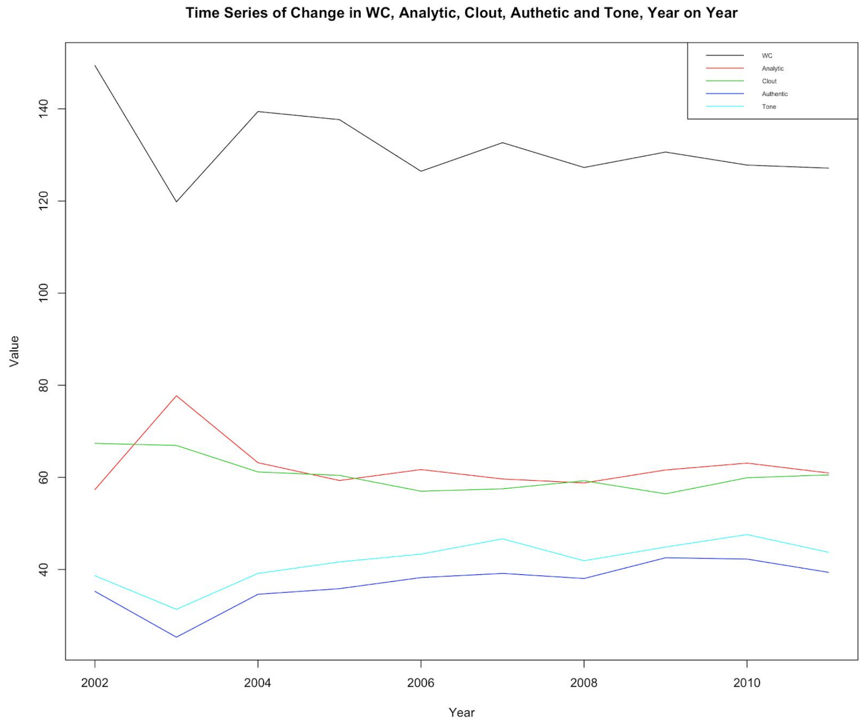


Figure 2.2 Time series of WC, Analytic, Clout, Authentic, Tone, Year-on-Year

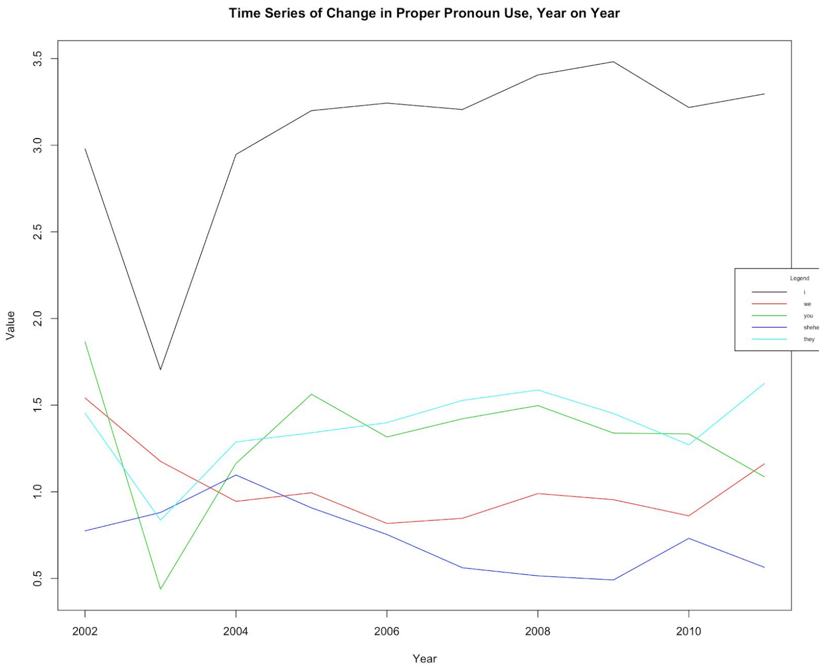


Figure 2.3: Time series of Personal Pronouns Use, Year-on-year

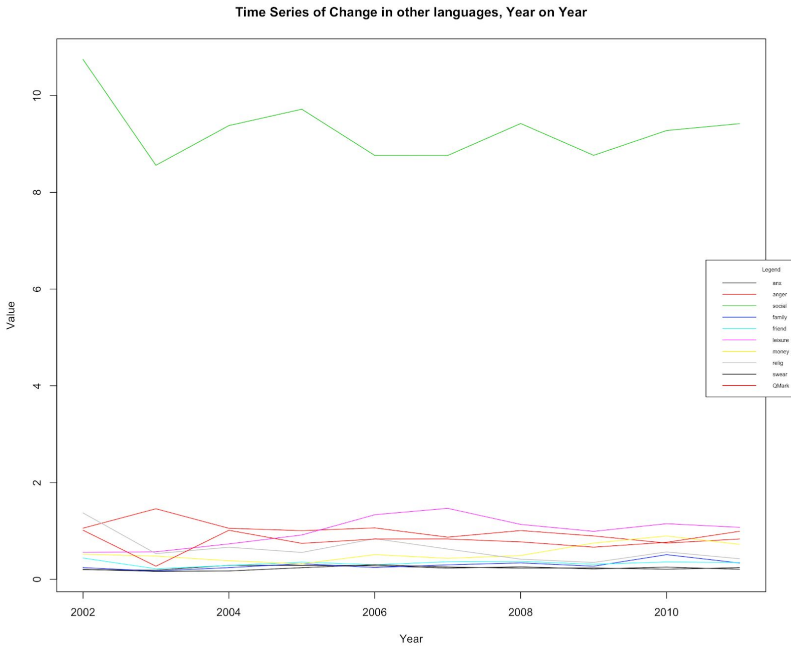


Figure 2.4: Time series of other language attributes, Year-on-year

As mentioned during the analysis of the correlation coefficients between periods, we are able to consistently see that language doesn't generally change over the period in the above three time series graphs of the overall landscape of every thread aggregated together.

Grouping by ThreadID

Table 2.2: Finding the top threads by the number of posts

Top 10 ThreadIDs by Number of Posts	
252620	400
127115	360
145223	306
472752	240
283958	186
254138	161
309286	154
191868	134
773564	123
563904	122

We investigated the threads by the number of posts so that we were able to focus on the analysis of just one thread to see whether there were any changes over time. These are as followed in the table above with a mean of 4.1 post by each AuthorID.

Furthermore, the top posters (with more than 20 posts) in significant ThreadID were:

- T252620 - AuthorID 76174 with 35 post.
- T127115 - AuthorID 47875 with 152 posts and AuthorID 8912 with 26 posts
- T145223 - AuthorID 1582 with 40 posts, AuthorID 39170 with 39 posts, AuthorID 32249 with 28 posts, AuthorID 110 with 27 posts,
- T283958 - AuthorID 79334 with 74 posts and AuthorID 98061 with 21 posts
- T191868 - AuthorID 47875 with 44 posts,

Bar graph of pronoun usage for largest 10 threads

We investigated the percentage of pronoun utilisation within each thread in Figure 2.7. Generally, the 'i' pronoun was used to a greater extent relative to other pronouns, except for 127115 and 14522, whereby 'i' and 'shehe' pronouns are closely of almost 25% in the threads. We are able to see that the TreadID

472752 and 773564 have the greatest proportion of ‘i’ usage and smallest proportion of ‘shehe’ usage. ThreadID 127115 had the second largest number of posts and had a relative equal use of each pronoun compared to the other threads. We can make the assumption that posters online are possibly more likely to talk about oneselves, which is quite evident in the bar graph below. Furthermore, the pronoun ‘i’ and words relating to social language was most commonly used.
Another interesting insight was that there were relatively low proportions of the use of ‘we’ in this threads, signifying lower collective/group support of ideas.

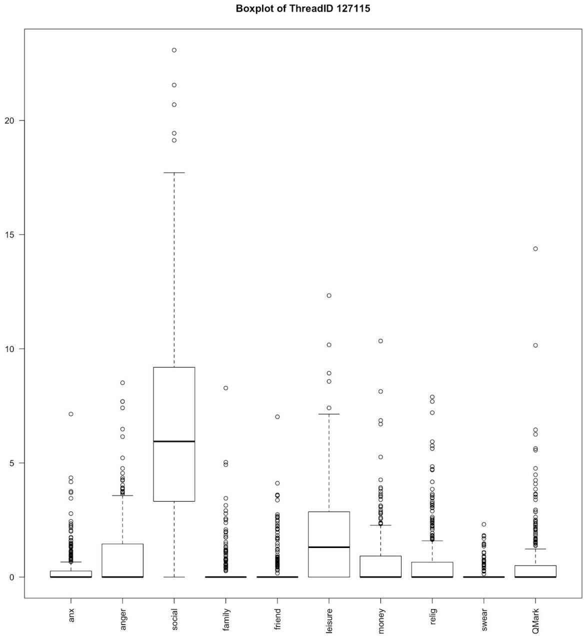


Figure 2.5: Boxplot of languages

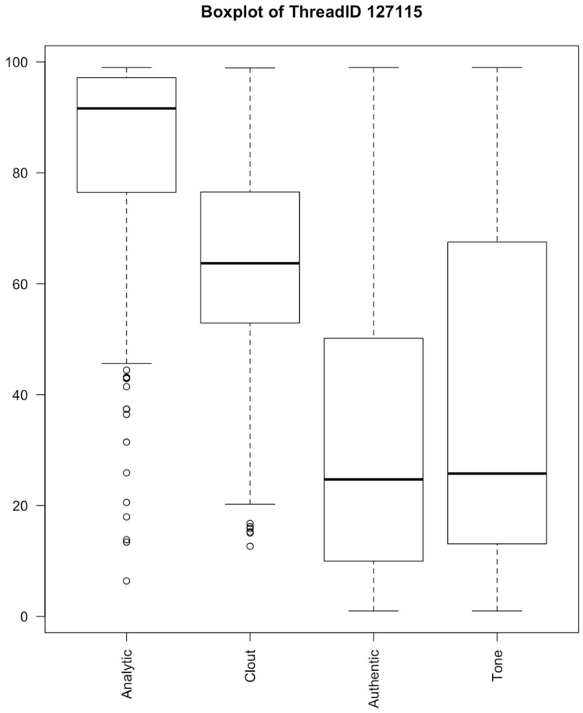


Figure 2.6: Boxplot of LIWC summary

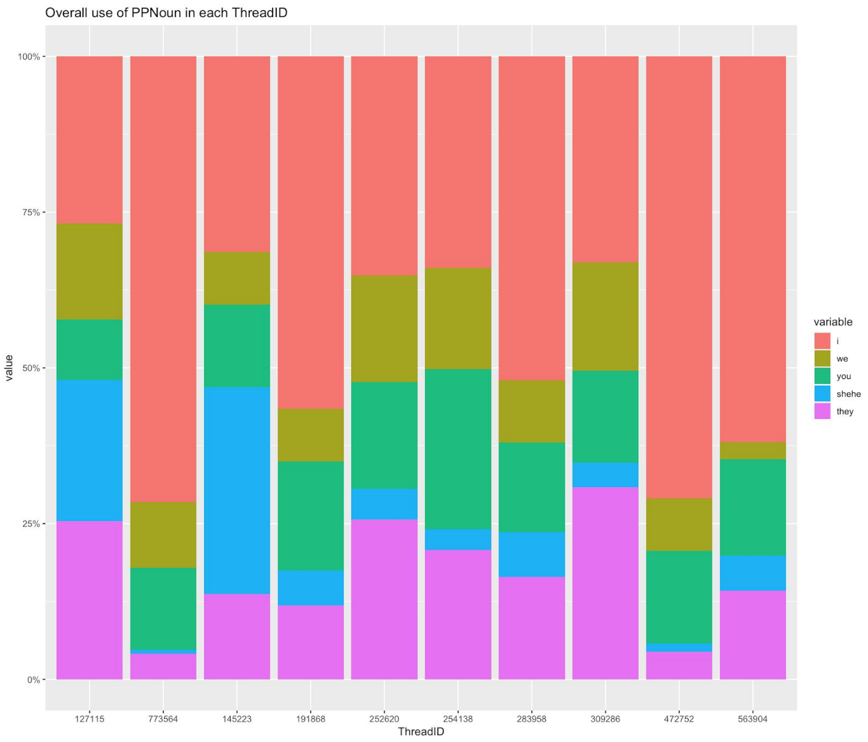


Figure 2.7: barplot of the proportional use of personal pronouns in each thread

We have decided to look into threads 127115 because of the larger number of posts and occurrence of its top author (AuthorID 47875), which had 152 posts ranging from August-2006 to February-2011.

From the LIWC Summary boxplot, we are able to see that in this thread, there are extremely high levels of analytical thinking compared to other attributes, with a median of 91. We are also able to see that the authentic and emotional tone had a very wide range, with similar means of 32.64 and 40.43 respectively, which is contrary to the relatively high clout value. This could suggest that there could be some form of strong debate or opinion. This is further emphasised in the relatively high clout median score of above 60 (in Figure 2.6), which suggests that posts are coming from someone's perspective.

From the time series graphs in Appendix 20-25, which displays an analysis of the threadID 127115 of analytic, clout, authentic, tone, social and proper pronouns, we are not able to see any significant changes in the authorID 47875 compared to the rest of the authors in the post (excluding the authorID 8912 due to low data count). Interestingly, in particular to the time series of the authentic language in Appendix 22, we see that authorID 47875's authentic tone deviates from the overall thread posters to a lower value, which indicates a more distanced form of discourse than the rest of the thread. In conjunction with the clout attribute in Appendix 21, authorID 47875 sees a minute uptrend from the overall clout attribute over time, which could suggest a slightly more domineering stance on the topic of discussion in the overall thread.

In conclusion, from the data that we had been given and after some processing, we have found from the data analysis that the language doesn’t change much over time. Looking at both the overall language used from every single thread and from threadID 127115 (with the most active user (AuthorID 47875)) in both the time series graph and correlation coefficient table, language has not changed significantly.

Conclusion

We found that there are differents in the proportion of sentiment within the data over different years and different times of day. The years which had either the highest positive or negative sentiment also showed the highest levels of these sentiments at different times of the day. Threads which showed the highest proportion of positive or negative sentiment in a given year, usually showed the highest proportion of that sentiment at given time of day during that year as well. Posts tended to show a higher proportion of negative sentiment before 12pm (morning), and higher proportion of positive sentiment after 12pm (afternoon). However, the overall the correlation between positive and negative sentiment was found to be very weak.

Conversely, the correlation between negative sentiment and anger and anxiety, appears to be quite strong. This is further supported by several threads having both the highest negative sentiment and either anger or anxiety in a given year. Different threads also showed different relationships between different sentiments and the use of certain personal pronouns.

We also found that there are significant differences in the language expressed by the most active users in the forum to the rest of the users; most notably, the group of most active users used less language relating to social processes, family and friends, used a lower proportion of personal pronouns and expressed less emotional tone but were more analytic and had greater impact (Clout).

In terms of the threadID, in threadID 127115, where its most active user had the most data points across the four or so years, there aren’t much change in terms of the attributes which were analysed, mainly analytic, authentic, clout, tone, ppron and social.

Appendices

Appendix 1

Median Posemo Proportion (Year and Time)							
Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
In general	2.43%	2.68%	2.59%	2.56%	2.49%	2.52%	2.56%
12am to 12pm	2.39%	2.65%	2.54%	2.50%	2.53%	2.50%	2.50%
12pm to midnight	2.50%	2.73%	2.70%	2.62%	2.38%	2.54%	2.69%

Median posemo proportion according to year and time of day

Appendix 2

Median Negemo Proportion (Year and Time)							
Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
In general	1.97%	1.96%	1.69%	1.71%	1.69%	1.38%	1.69%
12am to 12pm	2.02%	2.00%	1.67%	1.75%	1.75%	1.50%	1.79%
12pm to midnight	1.92%	1.88%	1.74%	1.66%	1.54%	1.23%	1.53%

Median negemo proportion according to year and time of day

Appendix 3

Highest Posemo Proportion by ThreadID (Year and Time)								
	Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
In general	**ThreadID**	246014	249001	296985	120790	246014	482758	283958
	%	7.41	4.23	7.69	7.41	5.00	6.45	6.56
12am to 12pm	**ThreadID**	126241	179689	296985	467590	92985	651844	283958
	%	11.10	4.55	7.69	6.67	5.45	11.40	6.31
12pm to midnight	**ThreadID**	246014	191377	296985	120790	246014	482758	283958
	%	8.55	4.81	9.90	9.62	5.92	6.98	8.33

Shows the ThreadID with the highest posemo proportion in a given year, also well in the morning and afternoon of that year.

Appendix 4

Highest Negemo Proportion by ThreadID (Year and Time)								
	Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
In general	**ThreadID**	231935	231935	179689	461357	558874	539505	426705
	%	4.84	3.51	4.17	4.63	3.92	9.68	4.29
12am to 12pm	**ThreadID**	231935	301325	179689	461357	558874	539505	426705
	%	5.06	3.75	5.62	4.50	3.95	9.68	4.79
12pm to midnight	**ThreadID**	233124	325032	361963	461357	140099	767538	426705
	%	4.29	5.14	5.62	5.07	4.92	3.68	4.16

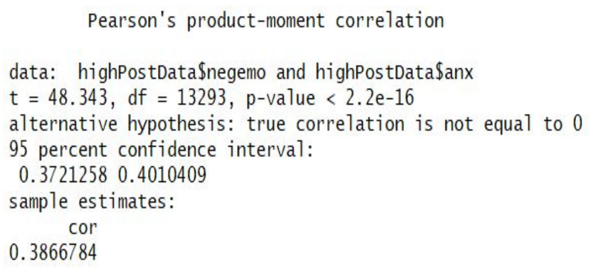
Shows the ThreadID with the highest negemo proportion in a given year, also well in the morning and afternoon of that year.

Appendix 5

	negemo	**posemo**	**anx**	**anger**
2011	426705	283958	127115	426705
2010	120790	482758, 651844	530558	120790
2009	558874	246014	578694	558874
2008	461357	120790	461357	120790
2007	179689, 361963	296985	179689, 361963	345383, 120790
2006	231935	249001	325032	209991
2005	231935, 233124	246014, 249001	222887	126241

Shows the ThreadID with the highest sentiment for years between 2005 till 2011

Appendix 6

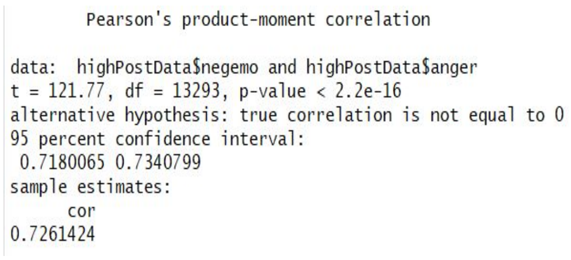


```
        Pearson's product-moment correlation

data:  highPostData$negemo and highPostData$anx
t = 48.343, df = 13293, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3721258 0.4010409
sample estimates:
      cor
0.3866784
```

Correlation between negative emotion/negemo and anxiety/anx

Appendix 7

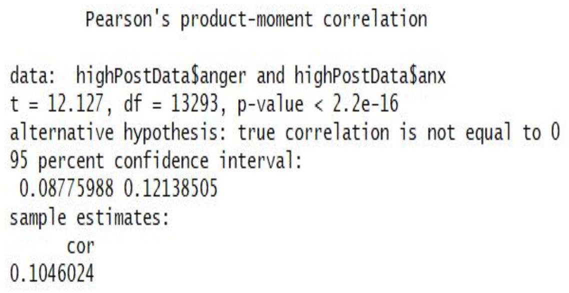


```
        Pearson's product-moment correlation

data:  highPostData$negemo and highPostData$anger
t = 121.77, df = 13293, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7180065 0.7340799
sample estimates:
      cor
0.7261424
```

Correlation between negative emotion/negemo and anger

Appendix 8

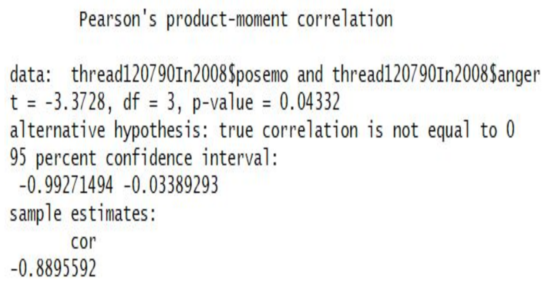


```
        Pearson's product-moment correlation

data:  highPostData$anger and highPostData$anx
t = 12.127, df = 13293, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08775988 0.12138505
sample estimates:
      cor
0.1046024
```

Correlation between anger and anxiety/anx

Appendix 9

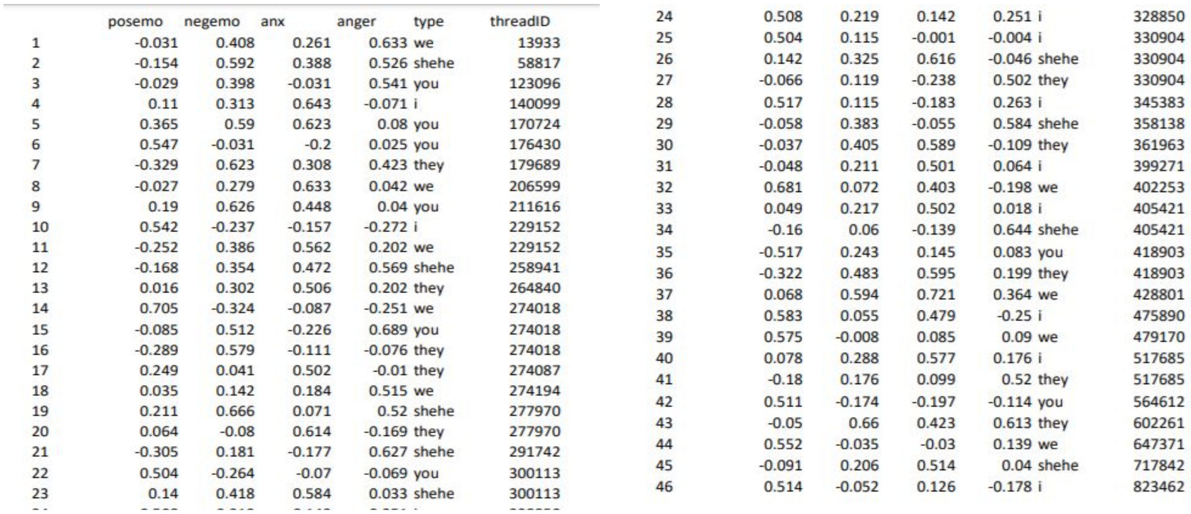


```
        Pearson's product-moment correlation

data:  thread120790In2008$posemo and thread120790In2008$anger
t = -3.3728, df = 3, p-value = 0.04332
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.99271494 -0.03389293
sample estimates:
      cor
-0.8895592
```

Correlation between positive emotion/posemo and anger for Thread 120790 in year 2008

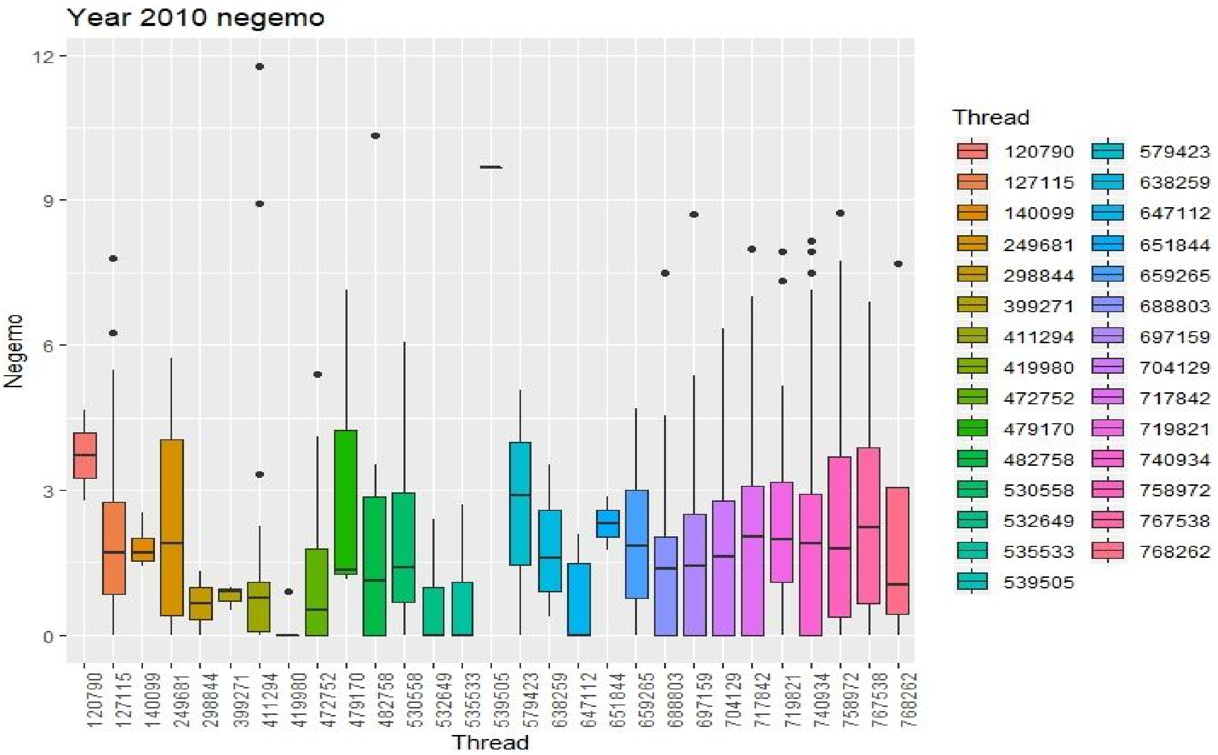
Appendix 10



	posemo	negemo	anx	anger	type	threadID
1	-0.031	0.408	0.261	0.633	we	13933
2	-0.154	0.592	0.388	0.526	shehe	58817
3	-0.029	0.398	-0.031	0.541	you	123096
4	0.11	0.313	0.643	-0.071	i	140099
5	0.365	0.59	0.623	0.08	you	170724
6	0.547	-0.031	-0.2	0.025	you	176430
7	-0.329	0.623	0.308	0.423	they	179689
8	-0.027	0.279	0.633	0.042	we	206599
9	0.19	0.626	0.448	0.04	you	211616
10	0.542	-0.237	-0.157	-0.272	i	229152
11	-0.252	0.386	0.562	0.202	we	229152
12	-0.168	0.354	0.472	0.569	shehe	258941
13	0.016	0.302	0.506	0.202	they	264840
14	0.705	-0.324	-0.087	-0.251	we	274018
15	-0.085	0.512	-0.226	0.689	you	274018
16	-0.289	0.579	-0.111	-0.076	they	274018
17	0.249	0.041	0.502	-0.01	they	274087
18	0.035	0.142	0.184	0.515	we	274194
19	0.211	0.666	0.071	0.52	shehe	277970
20	0.064	-0.08	0.614	-0.169	they	277970
21	-0.305	0.181	-0.177	0.627	shehe	291742
22	0.504	-0.264	-0.07	-0.069	you	300113
23	0.14	0.418	0.584	0.033	shehe	300113
24	0.508	0.219	0.142	0.251	i	328850
25	0.504	0.115	-0.001	-0.004	i	330904
26	0.142	0.325	0.616	-0.046	shehe	330904
27	-0.066	0.119	-0.238	0.502	they	330904
28	0.517	0.115	-0.183	0.263	i	345383
29	-0.058	0.383	-0.055	0.584	shehe	358138
30	-0.037	0.405	0.589	-0.109	they	361963
31	-0.048	0.211	0.501	0.064	i	399271
32	0.681	0.072	0.403	-0.198	we	402253
33	0.049	0.217	0.502	0.018	i	405421
34	-0.16	0.06	-0.139	0.644	shehe	405421
35	-0.517	0.243	0.145	0.083	you	418903
36	-0.322	0.483	0.595	0.199	they	418903
37	0.068	0.594	0.721	0.364	we	428801
38	0.583	0.055	0.479	-0.25	i	475890
39	0.575	-0.008	0.085	0.09	we	479170
40	0.078	0.288	0.577	0.176	i	517685
41	-0.18	0.176	0.099	0.52	they	517685
42	0.511	-0.174	-0.197	-0.114	you	564612
43	-0.05	0.66	0.423	0.613	they	602261
44	0.552	-0.035	-0.03	0.139	we	647371
45	-0.091	0.206	0.514	0.04	shehe	717842
46	0.514	-0.052	0.126	-0.178	i	823462

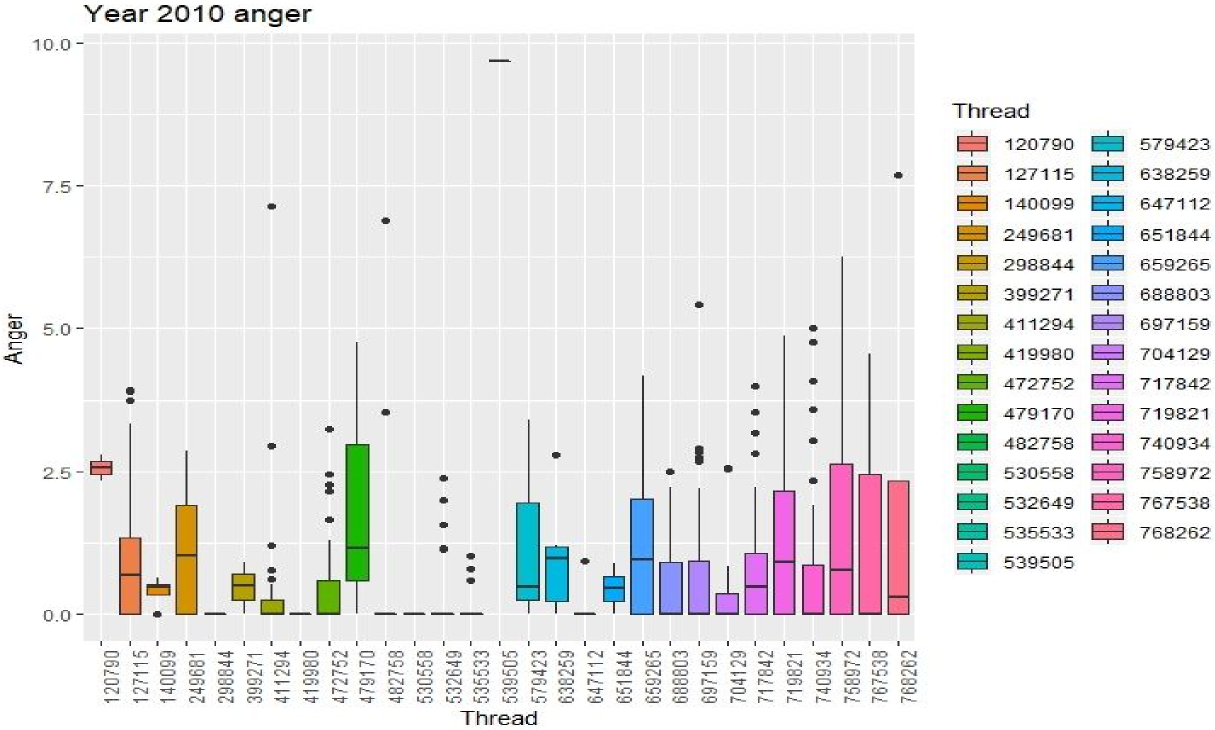
The correlation between personal pronoun and sentiments

Appendix 11



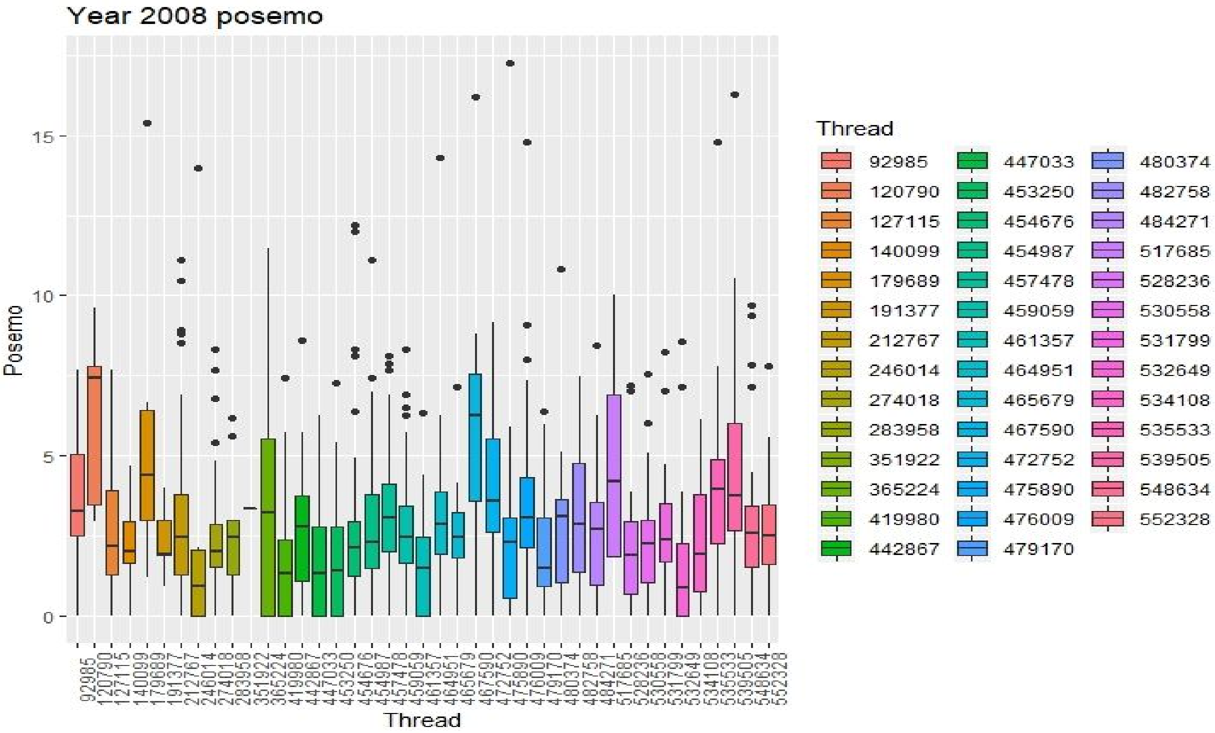
Boxplot of negative emotion / negemo in year 2010 group by threadID

Appendix 12



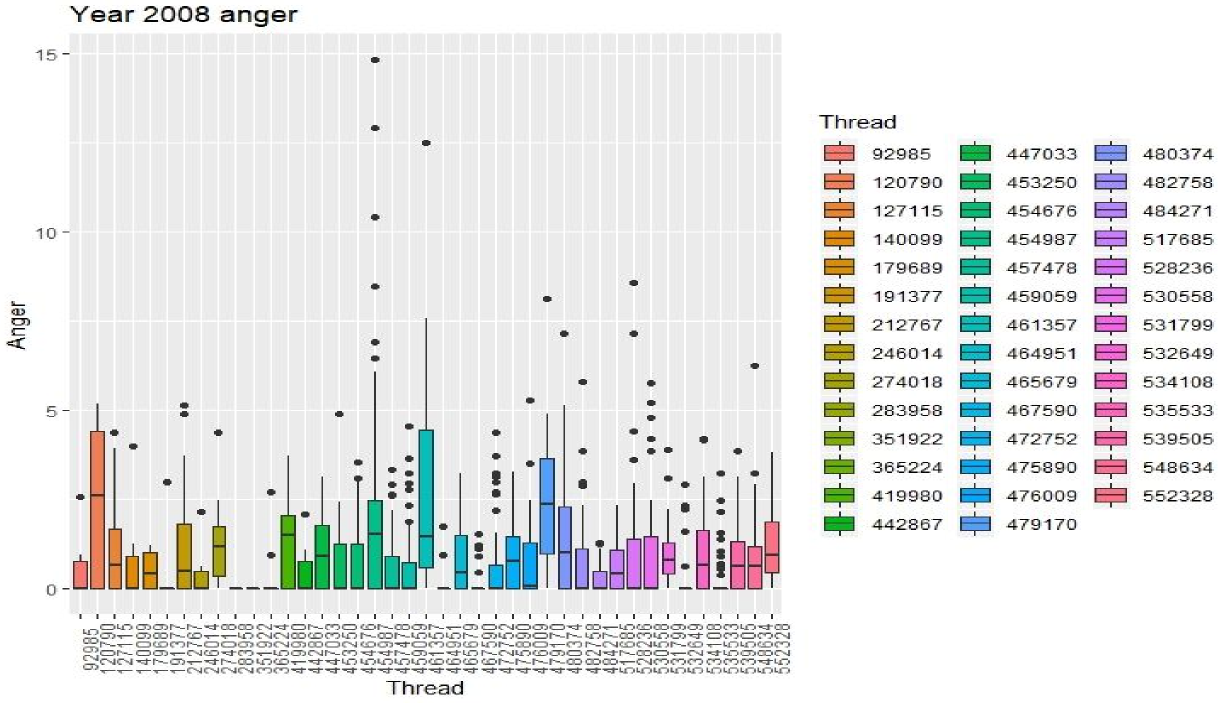
Boxplot of anger in year 2010 group by threadID

Appendix 13



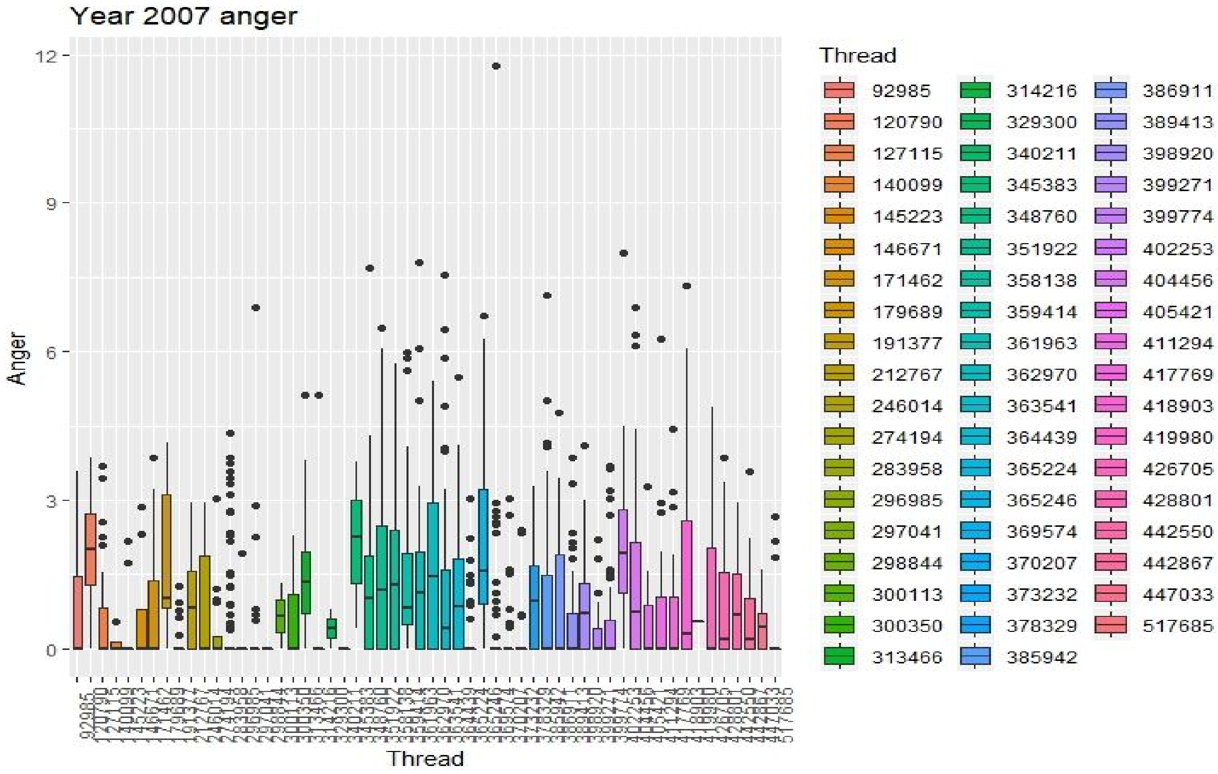
Boxplot of positive emotion / posemo in year 2008 group by threadID

Appendix 14



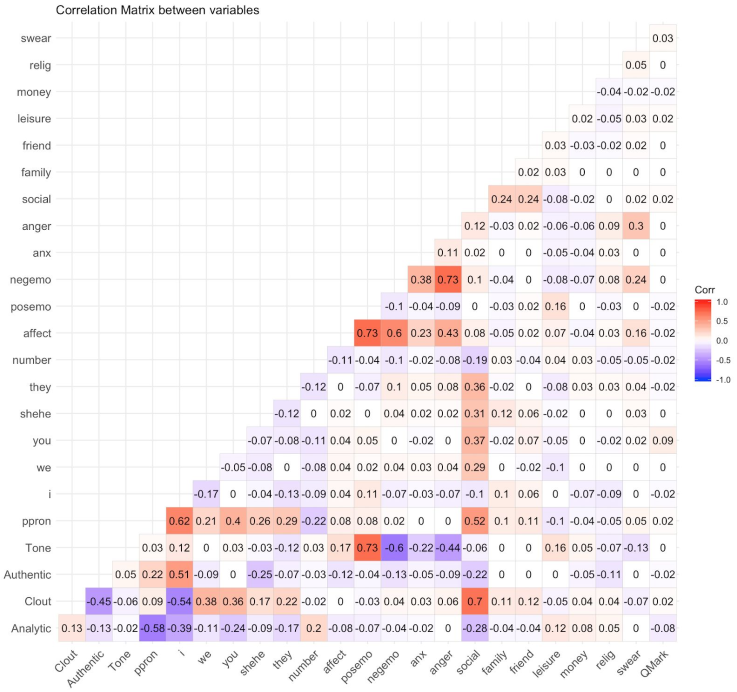
Boxplot of anger in year 2008 group by threadID

Appendix 15

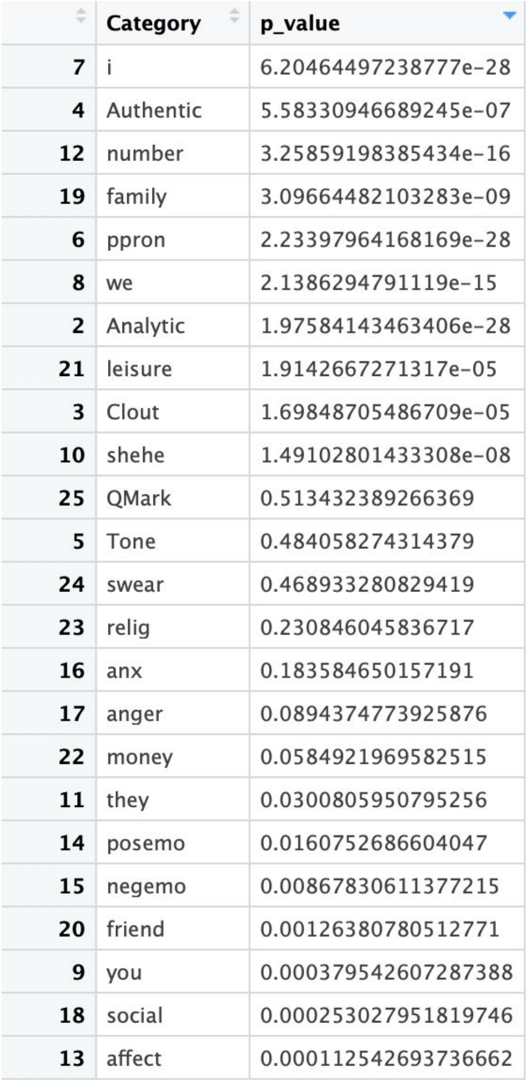


Boxplot of anger in year 2007 group by threadID

Appendix 16

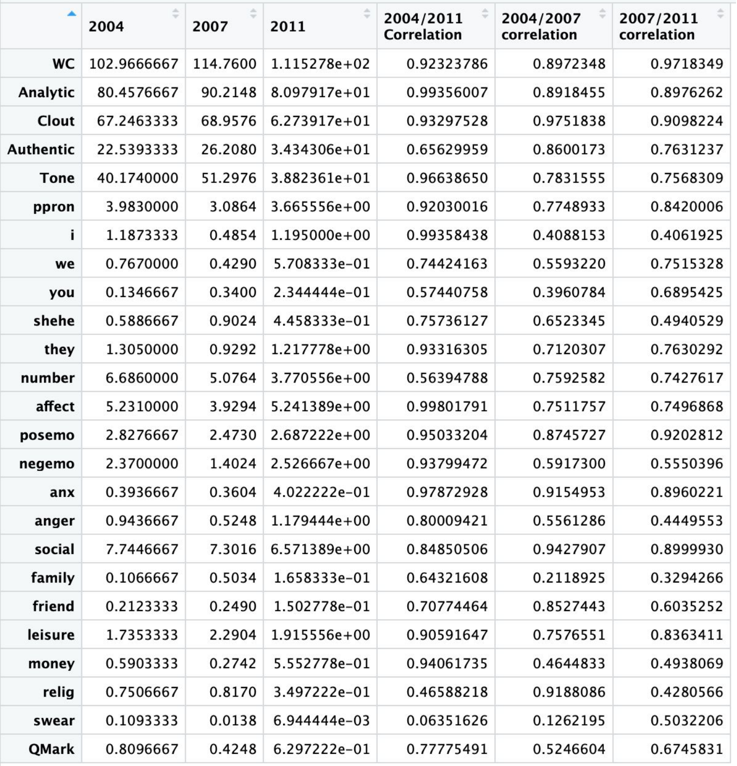

Correlation Matrix between variables

Appendix 17



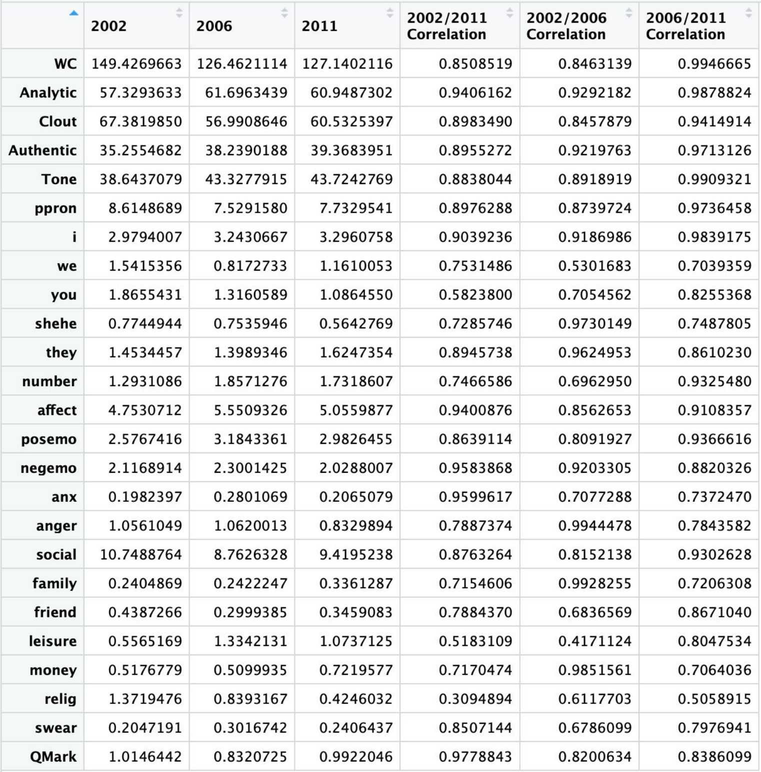
	Category	p_value
7	i	6.20464497238777e-28
4	Authentic	5.58330946689245e-07
12	number	3.25859198385434e-16
19	family	3.09664482103283e-09
6	ppron	2.23397964168169e-28
8	we	2.1386294791119e-15
2	Analytic	1.97584143463406e-28
21	leisure	1.9142667271317e-05
3	Clout	1.69848705486709e-05
10	shehe	1.49102801433308e-08
25	QMark	0.513432389266369
5	Tone	0.484058274314379
24	swear	0.468933280829419
23	relig	0.230846045836717
16	anx	0.183584650157191
17	anger	0.0894374773925876
22	money	0.0584921969582515
11	they	0.0300805950795256
14	posemo	0.0160752686604047
15	negemo	0.00867830611377215
20	friend	0.00126380780512771
9	you	0.000379542607287388
18	social	0.000253027951819746
13	affect	0.000112542693736662

Appendix 18



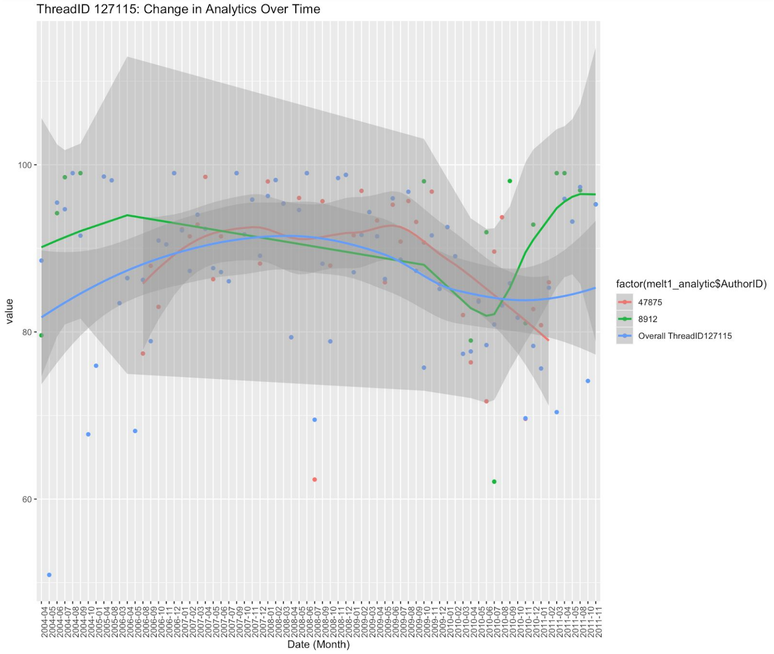
	2004	2007	2011	2004/2011 Correlation	2004/2007 correlation	2007/2011 correlation
WC	102.9666667	114.7600	1.115278e+02	0.92323786	0.8972348	0.9718349
Analytic	80.4576667	90.2148	8.097917e+01	0.99356007	0.8918455	0.8976262
Clout	67.2463333	68.9576	6.273917e+01	0.93297528	0.9751838	0.9098224
Authentic	22.5393333	26.2080	3.434306e+01	0.65629959	0.8600173	0.7631237
Tone	40.1740000	51.2976	3.882361e+01	0.96638650	0.7831555	0.7568309
ppron	3.9830000	3.0864	3.665556e+00	0.92030016	0.7748933	0.8420006
i	1.1873333	0.4854	1.195000e+00	0.99358438	0.4088153	0.4061925
we	0.7670000	0.4290	5.708333e-01	0.74424163	0.5593220	0.7515328
you	0.1346667	0.3400	2.344444e-01	0.57440758	0.3960784	0.6895425
shehe	0.5886667	0.9024	4.458333e-01	0.75736127	0.6523345	0.4940529
they	1.3050000	0.9292	1.217778e+00	0.93316305	0.7120307	0.7630292
number	6.6860000	5.0764	3.770556e+00	0.56394788	0.7592582	0.7427617
affect	5.2310000	3.9294	5.241389e+00	0.99801791	0.7511757	0.7496868
posemo	2.8276667	2.4730	2.687222e+00	0.95033204	0.8745727	0.9202812
negemo	2.3700000	1.4024	2.526667e+00	0.93799472	0.5917300	0.5550396
anx	0.3936667	0.3604	4.022222e-01	0.97872928	0.9154953	0.8960221
anger	0.9436667	0.5248	1.179444e+00	0.80009421	0.5561286	0.4449553
social	7.7446667	7.3016	6.571389e+00	0.84850506	0.9427907	0.8999930
family	0.1066667	0.5034	1.658333e-01	0.64321608	0.2118925	0.3294266
friend	0.2123333	0.2490	1.502778e-01	0.70774464	0.8527443	0.6035252
leisure	1.7353333	2.2904	1.915556e+00	0.90591647	0.7576551	0.8363411
money	0.5903333	0.2742	5.552778e-01	0.94061735	0.4644833	0.4938069
relig	0.7506667	0.8170	3.497222e-01	0.46588218	0.9188086	0.4280566
swear	0.1093333	0.0138	6.944444e-03	0.06351626	0.1262195	0.5032206
QMark	0.8096667	0.4248	6.297222e-01	0.77775491	0.5246604	0.6745831

Appendix 19

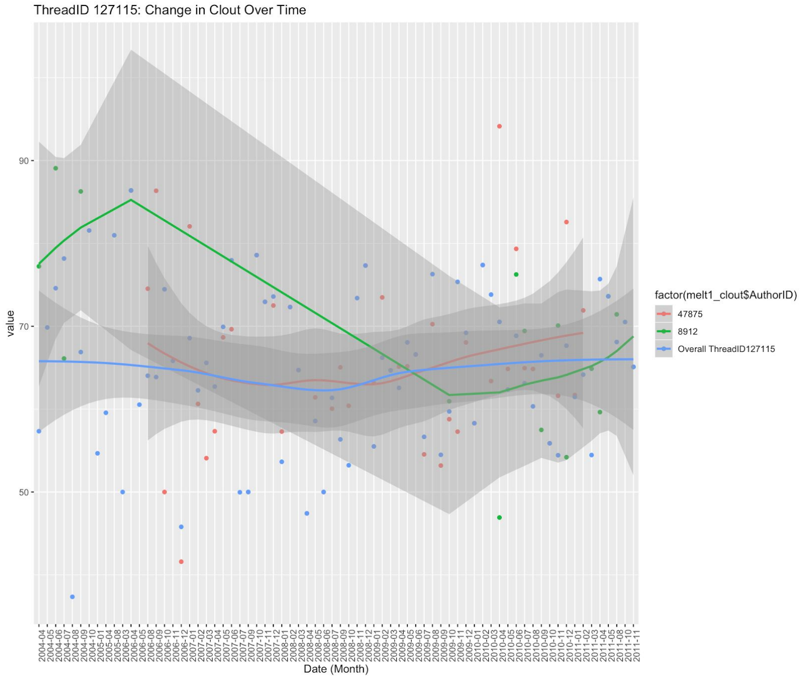


	2002	2006	2011	2002/2011 Correlation	2002/2006 Correlation	2006/2011 Correlation
WC	149.4269663	126.4621114	127.1402116	0.8508519	0.8463139	0.9946665
Analytic	57.3293633	61.6963439	60.9487302	0.9406162	0.9292182	0.9878824
Clout	67.3819850	56.9908646	60.5325397	0.8983490	0.8457879	0.9414914
Authentic	35.2554682	38.2390188	39.3683951	0.8955272	0.9219763	0.9713126
Tone	38.6437079	43.3277915	43.7242769	0.8838044	0.8918919	0.9909321
ppron	8.6148689	7.5291580	7.7329541	0.8976288	0.8739724	0.9736458
i	2.9794007	3.2430667	3.2960758	0.9039236	0.9186986	0.9839175
we	1.5415356	0.8172733	1.1610053	0.7531486	0.5301683	0.7039359
you	1.8655431	1.3160589	1.0864550	0.5823800	0.7054562	0.8255368
shehe	0.7744944	0.7535946	0.5642769	0.7285746	0.9730149	0.7487805
they	1.4534457	1.3989346	1.6247354	0.8945738	0.9624953	0.8610230
number	1.2931086	1.8571276	1.7318607	0.7466586	0.6962950	0.9325480
affect	4.7530712	5.5509326	5.0559877	0.9400876	0.8562653	0.9108357
posemo	2.5767416	3.1843361	2.9826455	0.8639114	0.8091927	0.9366616
negemo	2.1168914	2.3001425	2.0288007	0.9583868	0.9203305	0.8820326
anx	0.1982397	0.2801069	0.2065079	0.9599617	0.7077288	0.7372470
anger	1.0561049	1.0620013	0.8329894	0.7887374	0.9944478	0.7843582
social	10.7488764	8.7626328	9.4195238	0.8763264	0.8152138	0.9302628
family	0.2404869	0.2422247	0.3361287	0.7154606	0.9928255	0.7206308
friend	0.4387266	0.2999385	0.3459083	0.7884370	0.6836569	0.8671040
leisure	0.5565169	1.3342131	1.0737125	0.5183109	0.4171124	0.8047534
money	0.5176779	0.5099935	0.7219577	0.7170474	0.9851561	0.7064036
relig	1.3719476	0.8393167	0.4246032	0.3094894	0.6117703	0.5058915
swear	0.2047191	0.3016742	0.2406437	0.8507144	0.6786099	0.7976941
QMark	1.0146442	0.8320725	0.9922046	0.9778843	0.8200634	0.8386099

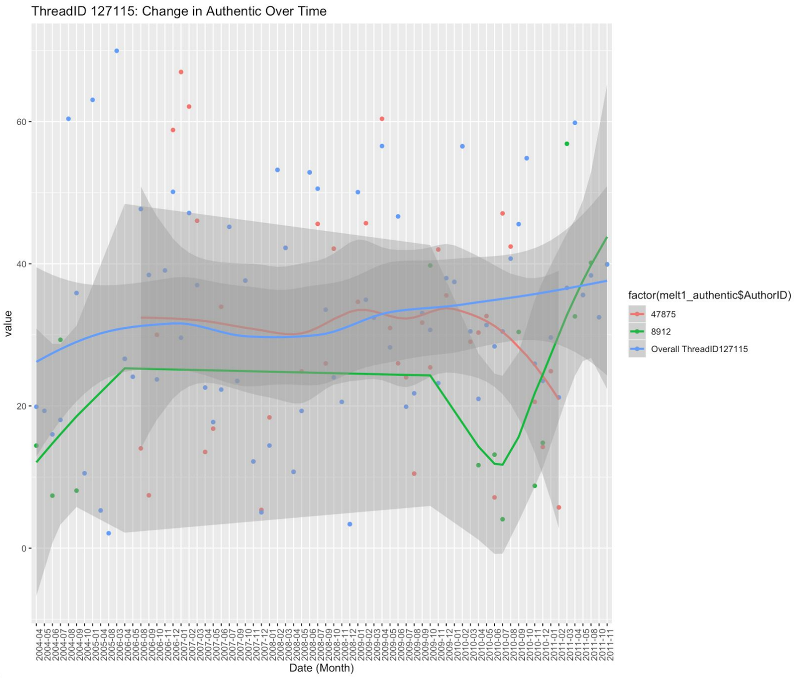
Appendix 20



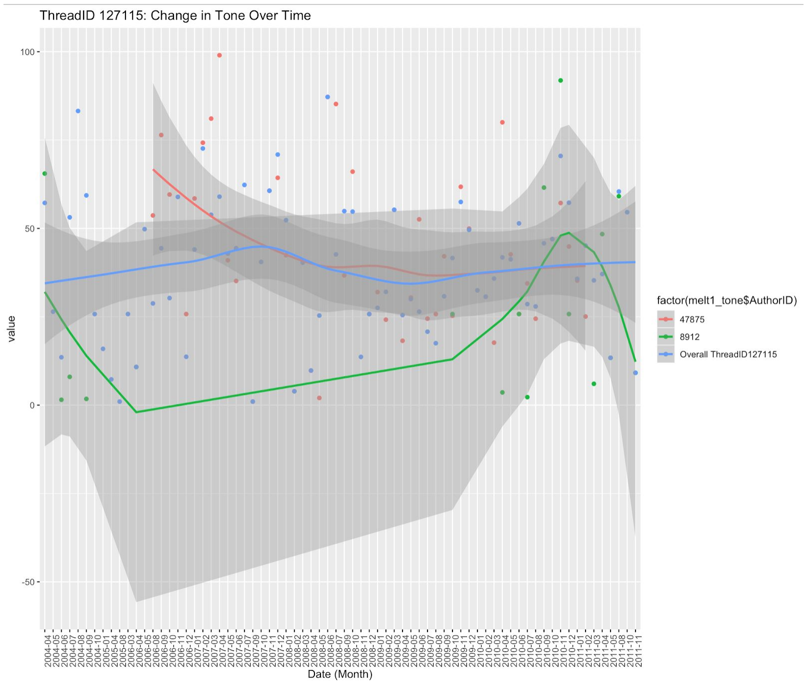
Appendix 21



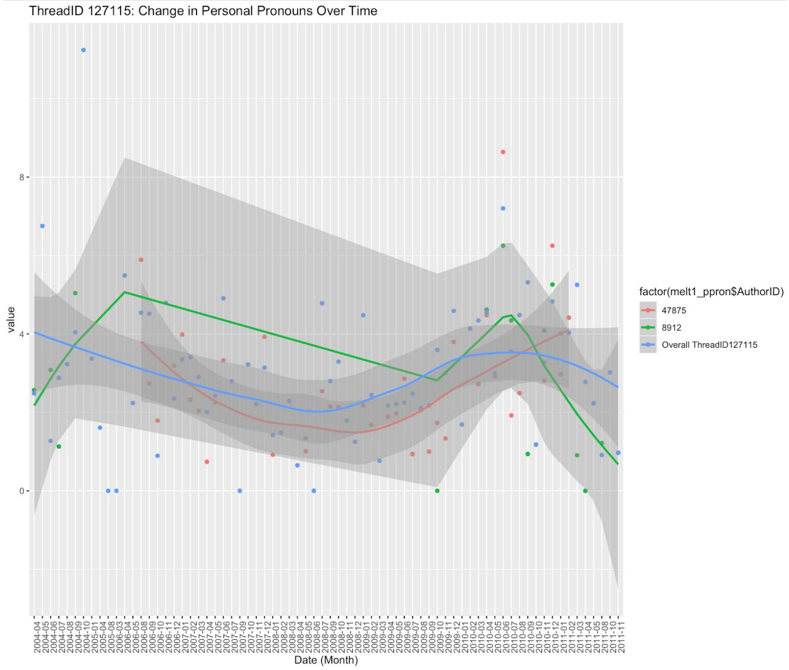
Appendix 22



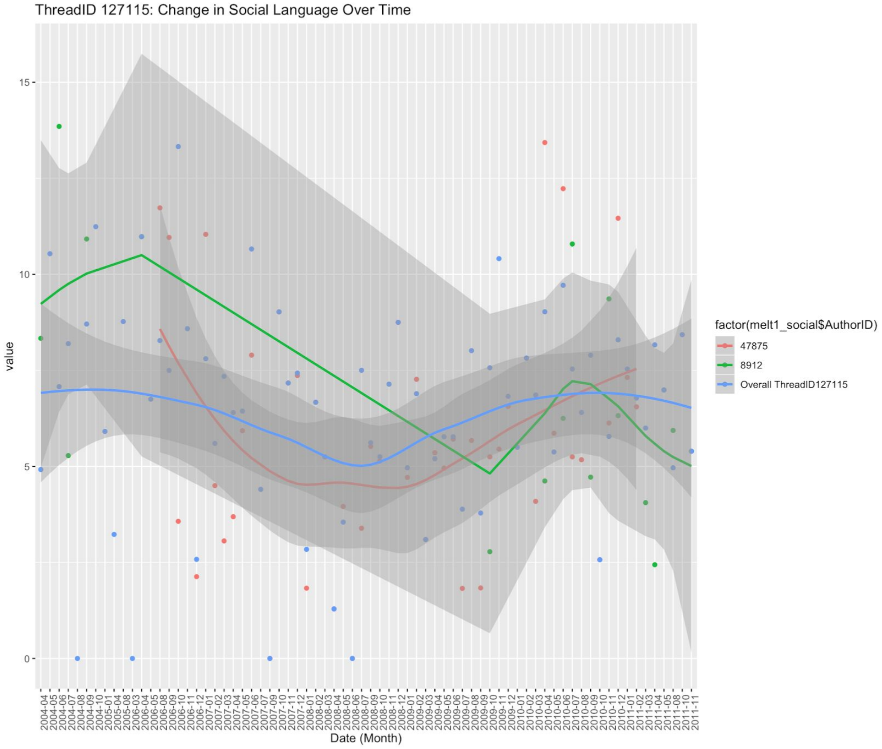
Appendix 23



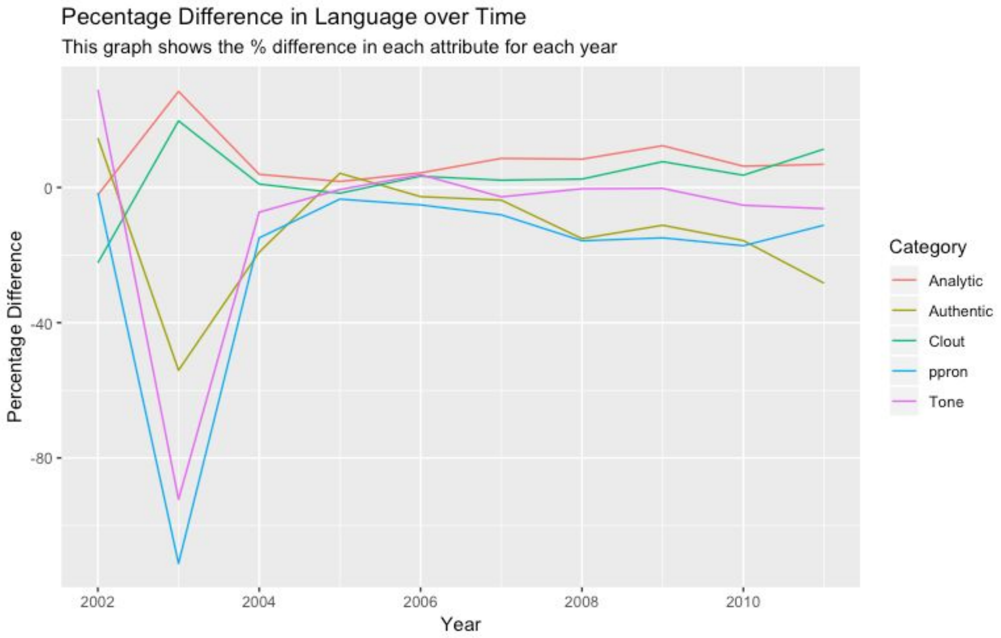
Appendix 24


ThreadID 127115: Change in Personal Pronouns Over Time

Appendix 25


ThreadID 127115: Change in Social Language Over Time

Appendix 26



Appendix 27

Link to Google Drive folder containing the R code used in analysis:
https://drive.google.com/drive/folders/1pbdC8ucoWwwCUHw94esRJdKQU3Bsta5X?usp=sharing