1. Analysis of activity and language on the forum

1.1. Activity of the participants from over time

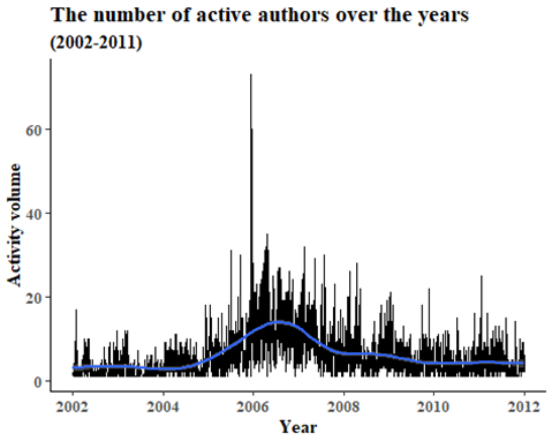


Figure 1.1. The number of active authors over years

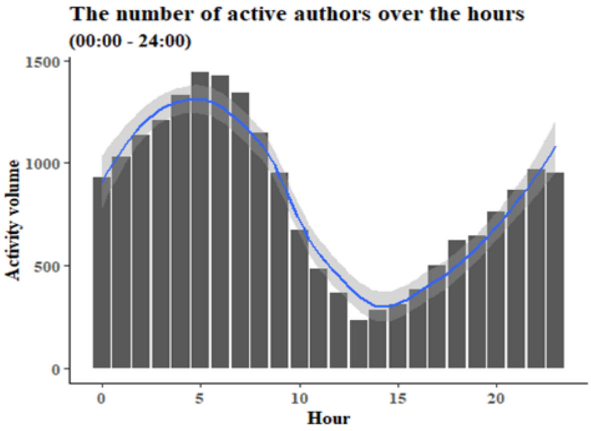


Figure 1.2. The number of active authors over hours

The trend is irregular for the activity over both the years and hours.

The fluctuation can be seen in Figure 1.1. There is an increasing trend from 2002 to 2007, with the utmost increase in 2006. After 2007, the trend decreased (Figure 1.1). That means the forum had been circulated for 4 years. Particularly, the peak in 2006 means that there were significant events being discussed around 2006 and then its popularity has dwindled.

During the day, the morning is usually busier than the afternoon as the trend between 00:00 and 10:00 a.m. has most of the traffic volume. The busiest traffic hour is at 5 a.m. (Figure 1.2). There might be the reasons that people stay up late to use the forums or check their messages once they wake up in the morning.

1.2.1 The trend of linguistic variables

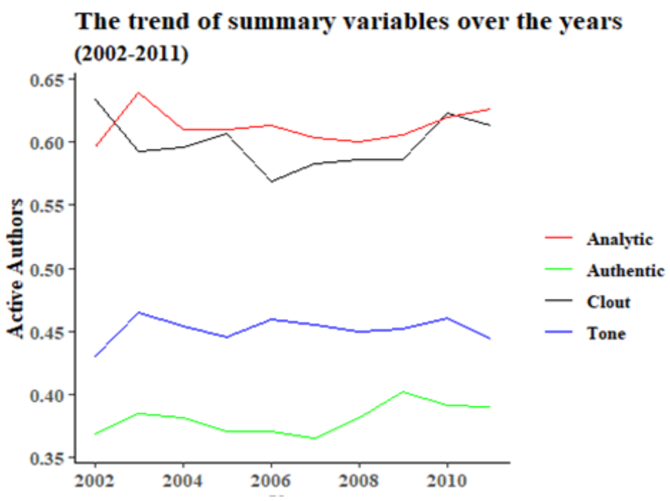


Figure 1.3. The trend of summary variables over the years

Based on Figure 1.3. we can observe:

1. The trends of summary variables are mostly steady, though the Clout's trend is the most turbulent.

2. The trending directions of Analytic and Tone are similar. It shows that they increased from 2002 to 2003, then decreased from 2003 to 2011, with similar shape.

3. The activities of Analytic and Clout were higher than Tone and Authentic over years. When authors make analysis, they tend to think from their expertise and post it with higher confidence.

1.2.2. The relationship of different linguistic variables

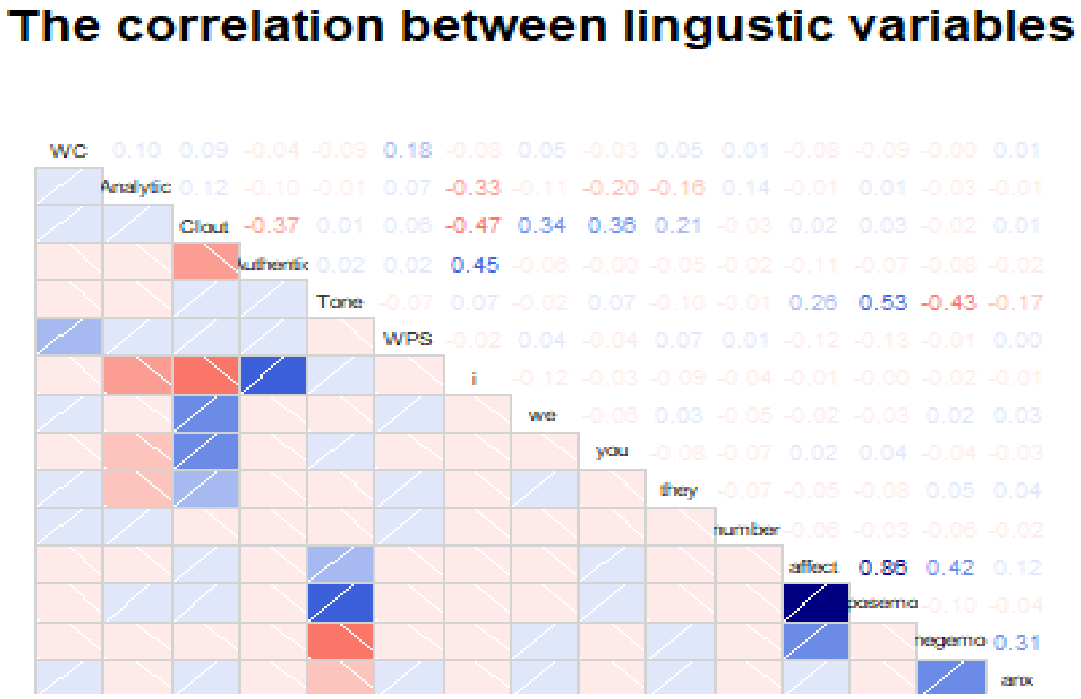


Figure 1.4. The correlation between linguistic variables

Although most of the variables have no correlation (Figure 1.4), there are 3 interesting insights found which will be listed below:

Insight 1: The relationship between Authentic and i is moderately positive (0.45).

In the forum, authors might want to contribute their genuine opinions which are shaped by their original experience. For example, Author 3 might have the opinion posted that "I believed A is killed by B as I saw the event happen". This authentic expression needs words like 'I'; therefore, it resonates the positive relationship Authentic and i.

Insight 2: The relationship among Affects, variables of emotions can be modelled using Multiple Regression.

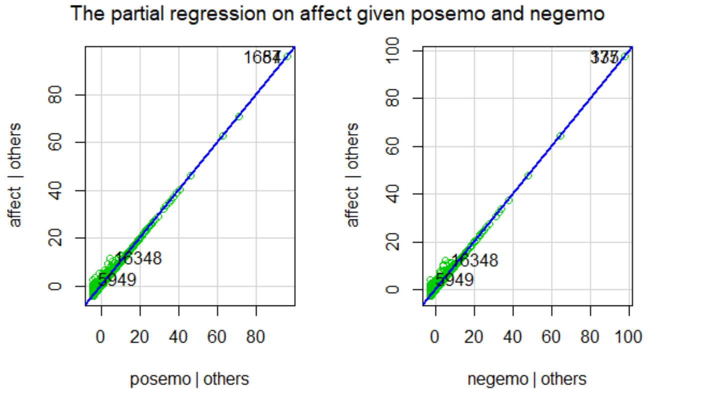


Figure 1.5. The partial regression of posemo and negemo on affect

In Figure 1.5, "posemo" (0.86), and "negemo" (0.42), have the greatest effect on "Affects", illustrating that the sentiment of expression, (Adjusted R-squared: 0.9984, p< 2e-16), given by:

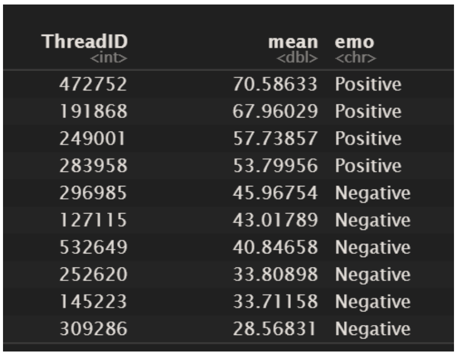
$$\widehat{Affect} = 0.0435734 \quad + 0.999\,(\,posemo) \; + (\text{negemo}) + \hat{\epsilon}$$

Insight 3: The relationships between Clout and the pronoun constitute the leadership style nowadays.

Higher Clout means more influential in which the Authors are more likely leaders. -> According to Harvard Business Review on leadership (Goleman, 2017) (!! HBR), effective leaders have higher emotional intelligence. They tend to make connections to others by showing their empathy. This could be justified by the correlation of Clout and the pronoun. It shows that the correlation between Clout and "I" is negative whereas Clout among "they", "you", and "we", are positive. Which means using more second-person and third-person pronouns have better care for others, thus having higher influence.

2. Analysis of language used by threads

2.1. The description of all the threads



ThreadID	mean	emo
<int>	<dbl>	<chr>
472752	70.58633	Positive
191868	67.96029	Positive
249001	57.73857	Positive
283958	53.79956	Positive
296985	45.96754	Negative
127115	43.01789	Negative
532649	40.84658	Negative
252620	33.80898	Negative
145223	33.71158	Negative
309286	28.56831	Negative

Figure 2.1. The 10 most active threads

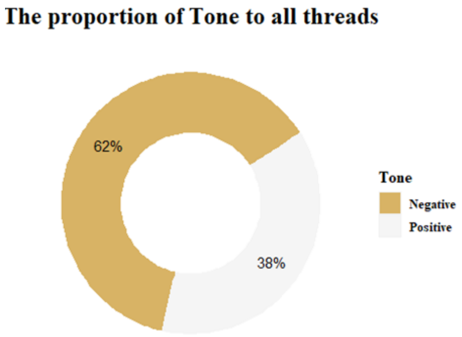


Figure 2.2 Statistically significant that there are more negative threads on average (p-value < 2.2e-16)

Figure 2.2. illustrated that the 62 % of the threads are negative. The median of Tone in overall threads 34.93, with 50 being the threshold value. The most positive and negative threads within the 10 most active threads are 472752 and 309286, respectively (Figure 2.1).

2.2. The consistency and changes of language used within threads over time

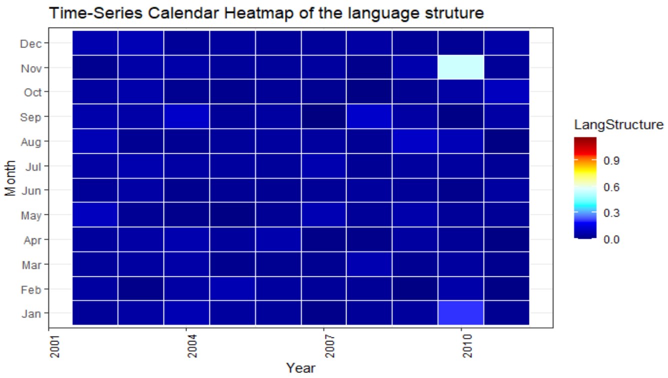


Figure 2.3. Time-Series Calendar Heatmap of the language structure

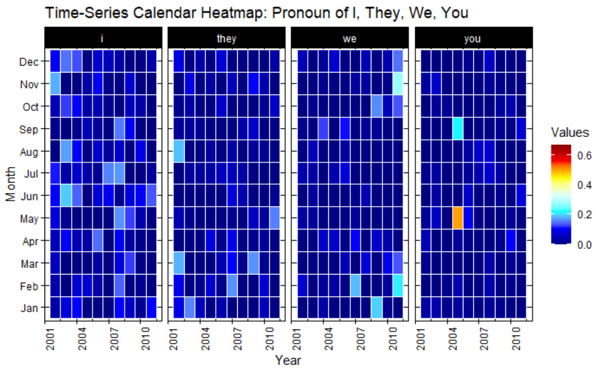


Figure 2.4. Time-Series Calendar Heatmap: Pronoun of I, They, We, You

The complexity is measured by language structures - LangStructure and different pronouns in the dataset. LangStructure consists of the values of WC and WPS. Generally, it is assumed that the higher the LangStructure, the longer the words per sentence and post, and hence the more on complexity.

The change of language structure was trivial. Almost all are of the same types of blue, with its values of around 0.25 (Figure 2.2). However, there is one white square in Nov 2010; the reason for this change might be due to the randomisation of the dataset. Moreover, the structure was rather consistent based on the similar colour types.

In terms of the pronouns used, even though the types of Blue in the squares are not as similar as the LangStructure one (Figure 2.3), generally the changes are less than 0.1. That means, "I", "we", "you", "they" only have little changes and been consistent. The only significant change of "you" was increased from 0.1 to 0.4 in May 2005. Based on the study (University of Michigan, 2017) !!, people use the word "you" to cope with negative experiences. In 2005, the Tone was decreasing (Figure 1.3) on which many authors had more negative experiences, perhaps.

2.3. The difference in the language between the most positive and negative threads

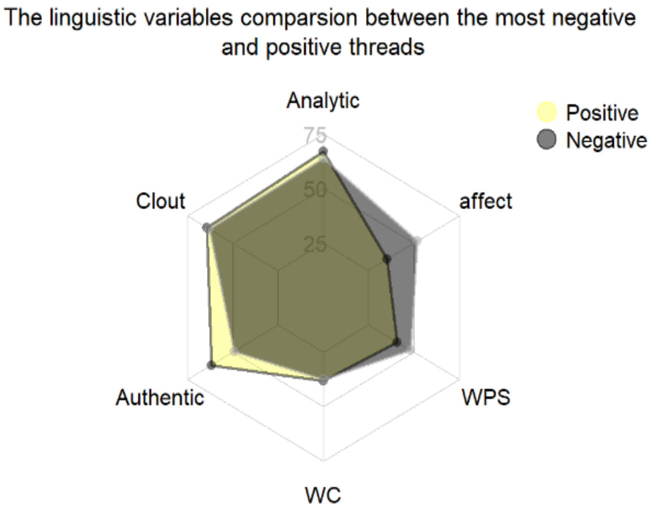


Figure 2.5. The linguistic variables comparison between the most negative and positive threads

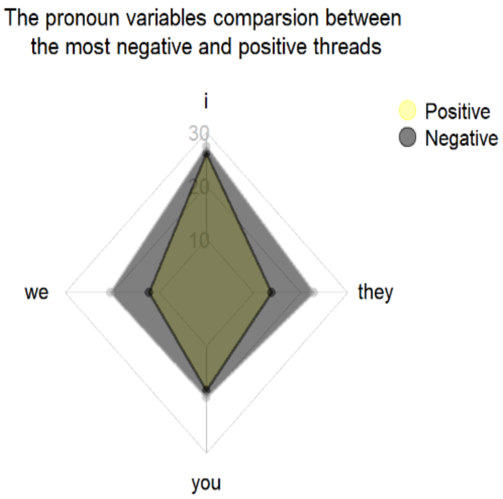


Figure 2.6. The pronoun variables comparison between

the most negative and positive threads

Since the most positive and negative threads are known, which are threads 472752 and 309286 respectively, it is worthy to have a comparison in terms of the language used.

Overall, both threads have similar Clout, Analytic, WC, WPS. But the positive one tends to have a higher Authentic as they perhaps shared more about their original experience or how life was leading genuinely. Whereas the negative one tends to have a higher Affect on which the reason might be that authors were used to vent their anger via negative experiences. In terms of the pronouns, the negative one used more "we" and "they"; perhaps they try to make a comparison or argue between themselves and others.

3. Social network comparison from February 2002 to March 2002

3.1. Definition of the networks

The network of authors is based on the interaction of different threads.

In terms of network structures, both networks are undirected. The network size is similar with 17 nodes and in February and 24 nodes. However, March had twice more interactions as it had 106 edges whereas February only has 48.

3.2 insight based on the network graphs (Figure 3.1 and Figure 3.2):

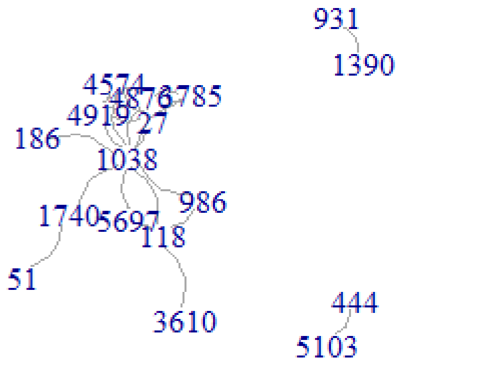


Figure 3.1. The network of authors in February 2002

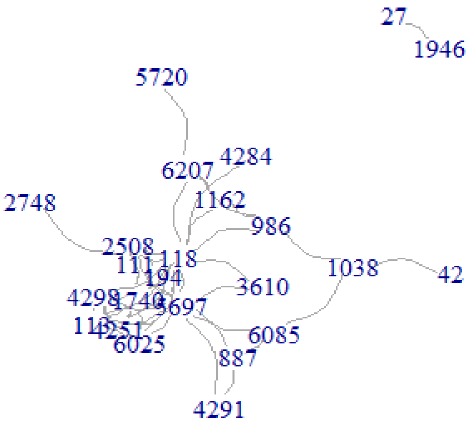


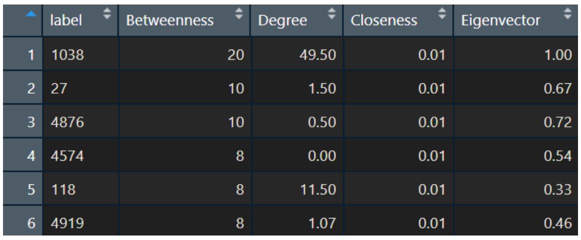
Figure 3.2. The network of authors in March 2002

3.2.1. Connection in the network graphs:

Although March seems more isolated (since it has 2 isolation), the connection is similar based on topology and interaction.

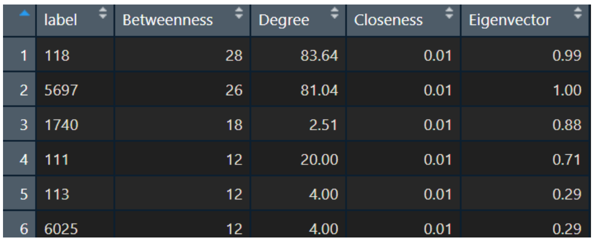
Both are star shaped. There are similar amounts of overlap in the centre of the star. The centralisation means the authors had similar topics to discuss. But the shape in March is bigger, with the corners of the star being more scattered. That shows some authors like 2748l or 5720 tended to have different interests, some of which were quite irrelevant to the main topics in the centre. Which is why being the corners of the star.

3.3. Centrality of node within the network



	label	Betweenness	Degree	Closeness	Eigenvector
1	1038	20	49.50	0.01	1.00
2	27	10	1.50	0.01	0.67
3	4876	10	0.50	0.01	0.72
4	4574	8	0.00	0.01	0.54
5	118	8	11.50	0.01	0.33
6	4919	8	1.07	0.01	0.46

Figure 3.3. The summary of centrality in February 2002



	label	Betweenness	Degree	Closeness	Eigenvector
1	118	28	83.64	0.01	0.99
2	5697	26	81.04	0.01	1.00
3	1740	18	2.51	0.01	0.88
4	111	12	20.00	0.01	0.71
5	113	12	4.00	0.01	0.29
6	6025	12	4.00	0.01	0.29

Figure 2.4. The summary of centrality in March 2002

Overall, the influence of Author 118, named as “A“, has been extended over the period. A ranked top 6th in February (Figure 3.3) and climbed to the top in March (Figure 3.4). The interpretation can be seen through the rank of the influence as below:

Note: It is assumed that the rank of the influence from February 2002 to March 2002 is based on the measure of centrality.

3.1. The degree of A ranked from 2nd to 1st. Having the greatest number of links to other nodes can be told that A is popular in the networks.

3.2. In terms of Betweenness, A climbed up from 4th to 1st. That means A influences the flow of the networks; it acts like a bridge of the communication dynamics. Perhaps A held authority in the network or talked more about popular topics.

3.3. Being ranked from 5th to 1st for eigenvector, A’s influence is spread over the whole network, not just those directly connected to it.

3.4. The score of closeness remained unchanged; A is in a highly connected network who can easily reach other authors.

4. Reflection on your investigation

This analysis is done using the methodology — Cross Industry Standard Process for Data Mining (CRISP-DM), which consists of 6 phases. There are 3 lessons learnt during my research process.

4.1. Lesson 1: The importance of having a clear project objective

4.1.1. Easier to investigate the problems

Having a clear project objective enhanced my understanding on what types of problems to investigate. For example, it is known that this is a linguistic analysis to assess the prevalence of certain thoughts, feelings and motivations used in communication. Knowing this objective helped the problem definition from the data mining perspective, which was to analyse authors' sentiments, language structures, social network and so on.

4.1.2. Easier to establish the narration of the story

It also helped with the narration of the story. By knowing what problems to investigate, it was easier to decompose the tasks. The application here is to analyse the dataset authors and threads to answer the problem mentioned.

4.2. Lesson 2: Data understanding and data preparation are the most frequent iteration

4.2.1. Understand the quality of the dataset:

It is clean with no NA values. Since the dataset was prepared by the FIT3152 teaching team who perhaps cleaned it before we received it. However,while understanding the data, for the variable of AuthorID, there was a peculiar value i.e., -1. Therefore, instead of ignoring the values of Author -1, it is assumed that Author -1 remained valid. I found that the number of posts from this author weighs more than other authors. I presume that those authors either posted anonymously or deactivated their accounts. This assumption is made based on the ongoing exploration of the values of Author -1, most of which are accurate to be analysed.

4.2.2. Prepare a dataset using data visualisation to answer questions and feed the regression model:

Since there were many ways to answer the questions, data visualisation helped find the right way. There are types of charts to represent the same data, but they could represent different meanings. For example, when answering question 2.3, I was intending to plot line charts for time series of linguistic variables. But considering that I needed to review the changes of linguistic variables over time and its consistency in which heatmaps could help.

To prepare the final datasets for feeding the regression model in question 2, initially, the boxplots by variables showed that there were many outliers, all of which were then removed. Moreover, the y axis of the boxplots shows that many dots were on different scales, particularly the four summary variables like Clout. Therefore, before making the model, the data needs to be transformed using min-max normalization.

All these conclusions are done by data visualisation as well as trial and error. The steps of data understanding, and data preparation are the second and third important phase of the analysis.

4.3. Lesson 3: Evaluation of the model

In terms of validity, the adjusted-r square of 0.98 and p-value <0 tells that the regression model is valid. However, when I fit the model, adjusted-r square score kept increasing. I realised that overfitting the model as it might lead to bias and invalidity.

Moreover, this model is further proving the answer to the question "Is there a relationship between variables?" After knowing the relationship between Affect and other emotion variables like Posemo, the regression model further justifies that Affect was the output from Posemo and Negemo.

Therefore, the model is valid, assuming that it is a linear relationship.