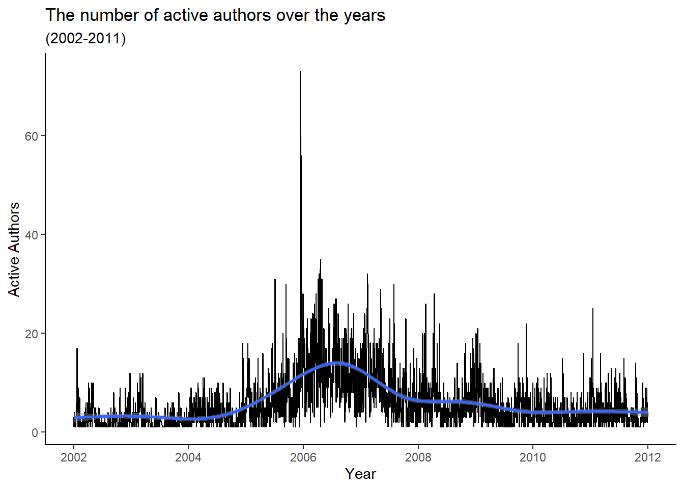
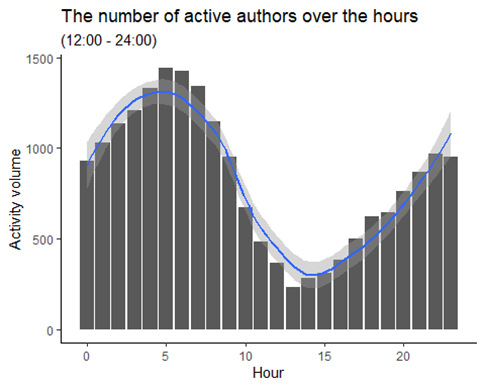
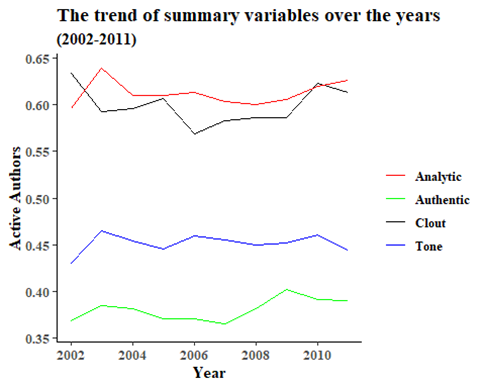
Introduction:

**1. Analysis of activity and language on the forum**





* 1. The activity of the participants from over time

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

According to !!, there is an increasing trend from 2002 and 2007, with the utmost increase in 2006. After 2007, the trend is decreased.

Insight: That means the forum has been circulated for 4 years and its popularity had been dwindled. There is a significant event that being discussed and boosted around in 2006.

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.

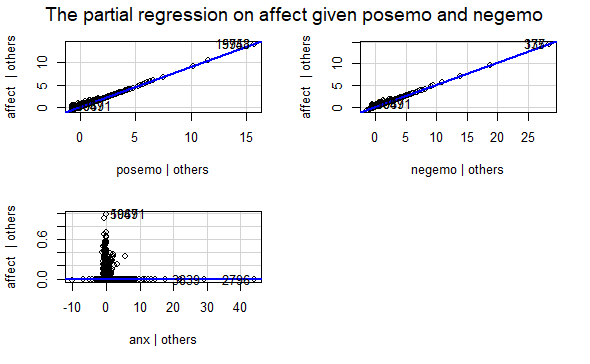
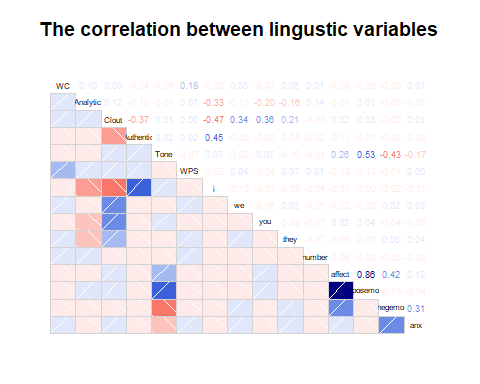
Insight : There might be the reasons that people stay up late to use the forum or check their messages once they wake up in the morning

* 1. The linguistic variables' trend and its relationship

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

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

The activity of Analytic and Clout were higher than Tone and Authentic.

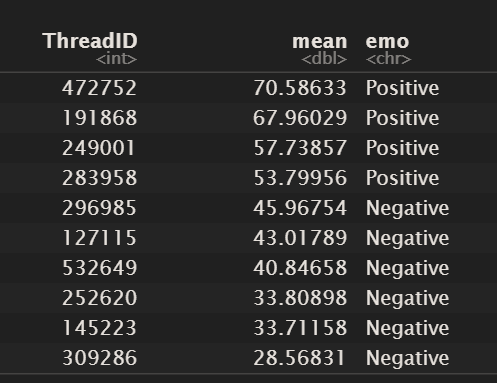


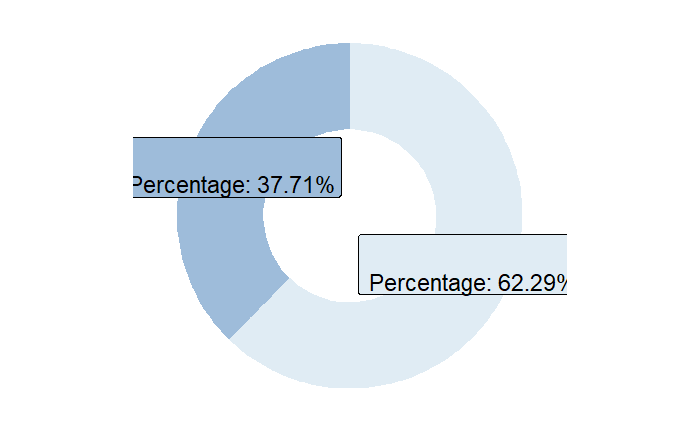
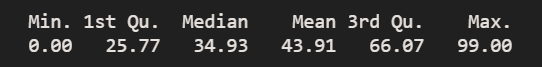
Although most of the variables have no correlation, there are 3 interesting insights found which will be listed below:

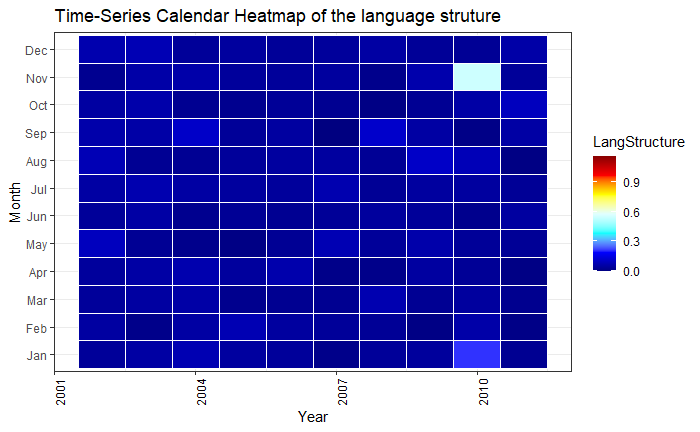
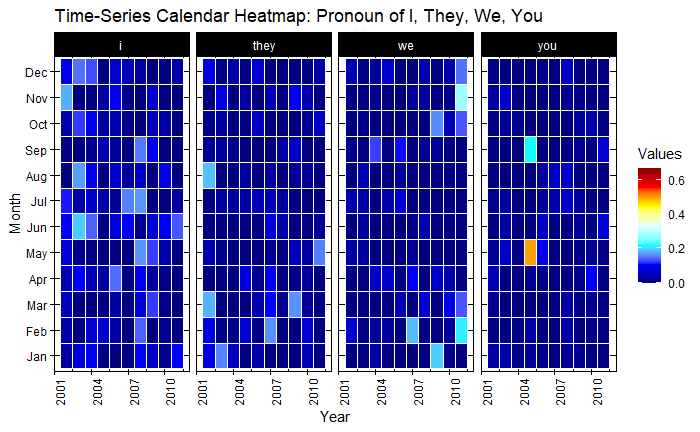
Correlation 1: The relationship between Authentic and i is moderately positive (.45). In the forum, Authors might want to contribute their genuine opinions which are shaped by their original experience. For example, Author 1 might have the opinion posted that “I believed A is killed by B as I actually saw the event happened”. Therefore, this authentic belief with the word of ‘I’ would resonate the positive relationship Authentic and i.

Correlation 2: The relationships between Clout and the pronoun constitute the leadership style nowadays. Higher Clout means more powerful in which the Authors are more likely leaders. -> According to Daniel Goleman’s review on leadership (!! [HBR](https://hbsp.harvard.edu/product/10101-PDF-ENG?Ntt=&itemFindingMethod=Search)), effective leaders have higher emotional intelligence. They tend to make connection to others by showing their empathy. This could be justified by the correlation of Clout and the pronoun. It is shown that the correlation between Clout and I is negative and Clout among they, you, and we are positive. -> Which means using more second-person and third-person pronouns have better care on others, thus having higher influence.

Correlation 3: The relationship among Affects, variables of emotions can be modelled using Multiple Regression. In graph !!,the correlations between Affects and posemo (.86), Affects and negemo (.42) , as well as Affects and anx (.12) illustrate that the sentiment of expression has a relationship, ( Adjusted R-squared: 0.9981 ; p<0), given by :

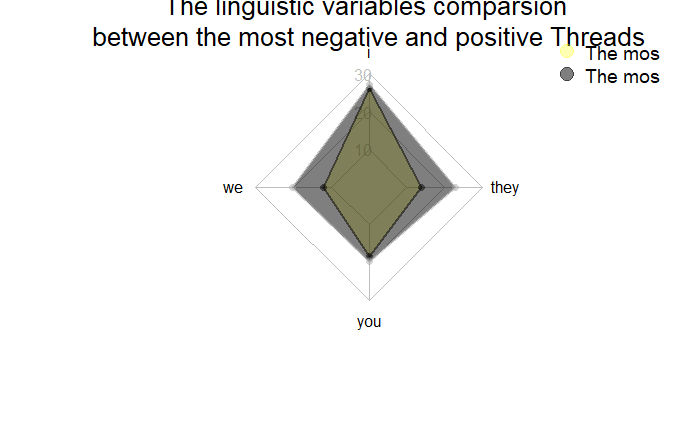
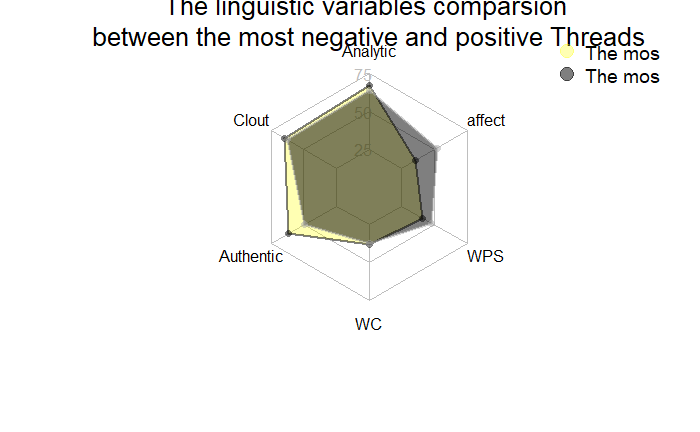
Affect = -0.00008346 + 0.9085 posemo + 0.5112 negemo + 0.0001879 anx.

******2. Analysis of language used by threads**

* 2.1. Description of all the threads
  + Based on the !!, it is illustrated that the 62 % of the threads are negative, with the median of overall threads 34.93. The most positive and negative threads are 472752 and 309286 respectively.
    - 2.1.2. The consistency and change of the language used within threads over time
    - The complexity is measured by LangStructure — its language structure and different pronouns in the dataset. LangStructure consists the values of WC and WPS. Generally, it is assumed that the higher the values, the longer the words per sentence and post, and hence more on complexity.
    - In terms of language structure, almost all are of the same types of blue, with its values of around 0.5. However, there is one white square in Oct 2010. That means the value increased to 0.6 during that point of time then decreased afterwards. The reason of this change might be due to the randomisation of the dataset. Also, the structure had been 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, generally the changes are less than 0.1. That means , I, we, you, they only have little changes. The significant change of “you” was increased from 0.1 to 0.4 in May 2005. Based on study [!!,](https://www.sciencedaily.com/releases/2017/03/170323141411.htm) people use the word 'you' to cope with negative experiences. Perhaps that period many authors had more negative experiences.

2.2. The difference in the language used between the most positive and negative threads? Since the positive and negative threads are known, 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 going on 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 and negative experiences.



1. **Social network comparison from February 2002 to March 2002** 
   1. Definition of the networks
      1. The network of the authors is based on the interaction of the threads. (!! If hav space, give an example).
      2. Network structures:
         1. they are undirected.
         2. The network size is similar with 17 nodes and in February and 24 nodes. However, March has more interaction as it has 106 edges whereas February only has 48.
   2. insight based the network graphs (figure!! and figure!!):
      1. Connection in the network
         1. Although March seems more isolated (since it has 2 isolation), the connection is similar based on the topology and interaction.
         2. Both are star shaped. There are similar amounts of overlap in the centre of the star. That means this is the authors have similar topics discussed. Then the corner of the stars being scattered means that some authors like 3610 tend to have different interests.
         3. >> (why? The meaning of star?)
   3. Centrality of node within the network
      1. Overall, the influence of Author 118, named as “A“, has been extended over the period. The interpretation can be seen through rank of the influence (!figure table!) as below:
         1. Assumption: The rank of the influence from February 2002 to March 2002 is based on the measure of the centrality in which A ranked top 6th in February and climbed to the top in March.

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

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

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

The score of closeness remained unchanged is; A is in the highly connected network who can easily reached to other authors.

1. 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.

Lesson 1: The importance of having a clear project objective

1.problems to investigate and story to tell

This enhanced my understanding on what types of problems to investigate. For example, it is known that this is a linguistic analysis to assesses 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.

* 1. narration of the story

It also helped 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.

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

Understand the quality of the dataset :

It is cleaned with no NA values. Since the dataset is prepared by FIT3152 teaching teams who perhaps clean it before we received it. However, in the variable of AuthorID, there was a peculiar value i.e. -1. Therefore, instead of ignoring values of Author -1, it is assumed that Author -1 remained valid. Probably, the intention of the Author -1 might signify Author 1. This assumption is made based on the ongoing exploration of the values of Author -1, most of which are accurate to be analyse.

Prepare the dataset by using data visualisation for answering questions and feeding the regression model :

Since there are many ways to answer the problems, data visualisation helped find the right way. There are types of charts to represent the same data but they can represent different meaning. For example, when answering question 3, I was intending to plot line charts for time series of linguistic variables. But considering that I need to review the changes of linguistic variables over time and its consistency, on 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 then are removed. The y axis of a scatter chart show that many dots were on different scales, particularly the summary variables like Clout. Therefore, before making the model, the data needs to be transformed using min-max normalization.

All of these conclusion are done by trial and error. The steps of data understanding and data preparation are the second and third important phase of the analysis.

* 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. The caution I realised was not to overfit the model as it might lead to the bias and invalidity.

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

Therefore, the model is valid, assumed that is a linear relationship.