I.Introduction

People adopt similar patterns of language when they interact. This is proven in the theory of social science. In this report, we will be discussing if the language pattern used by members in an online forum changes over time. This online forum consists of over 20,000 posts from years 2002 to 2011. For this analysis, we are ignoring posts by unknown authors and posts with zero word count.

Since we want to investigate the change in language, we have decided to choose the 4 factors from the multivariate dataset. The 4 factors are Analytic, Clout, Authentic and Tone.

Next, we decided to filter the data by splitting it into two groups which are active and inactive threads. Active threads are determined by the maximum number of posts in each year from 2002 to 2011. And the remaining threads are grouped as inactive threads.

From the 10 active threads, we further investigated those which are interactive by looking at the frequency of posts by each author, which gave us 5 threads, as seen in figure 2.
In order to find out how people in a single thread communicate, we have chosen 5 interactive thread IDs and investigated how the language change over time for each of them.

Finally, we investigate if the language has changed for active and non-active threads over the years by using Hypothesis testing. From the hypothesis test, if it is not certain whether the language used in active threads is similar to the language used in non-active threads, we plot boxplots and observe the median to see if the language used in active threads is similar to the language used in non-active threads.

II. Language change across multiple active threads

As stated in the Introduction, active threads are determined by the maximum number of posts in each year from 2002 to 2011. We are aware that some threads lasts more than a year but in our analysis we have made an assumption for active threads that the number of posts for the year in a thread will be the total number of posts for that thread which lasts for that year. For example, ThreadID 127115 has the most number of posts in the year 2010 with 79 posts but ThreadID 127115 lasts for 7 years (2004 to 2011) and has a total of 311 posts. We are going to assume that ThreadID 127115 has a total of 79 posts in the year 2010 and ignore the original total number of posts (i.e. 311 posts). We are making this assumption because we are already picking threads with maximum number of posts which means that it contains higher proportion of posts for that thread in a particular year as compared to other posts from that thread in the other years. Therefore, analysing this subset is going to be very indicative of the overall performance of the thread. Non active threads are basically the remaining threads that are not considered active for each year.

With our 10 active threads, ranging from 2002 to 2011, we decided to analyse how each factor of language changes over the years, using a line graph as shown in figure 1. And we have chosen median to measure the proportion of language because median is the better central tendency and it is susceptible to extreme values such as outliers.

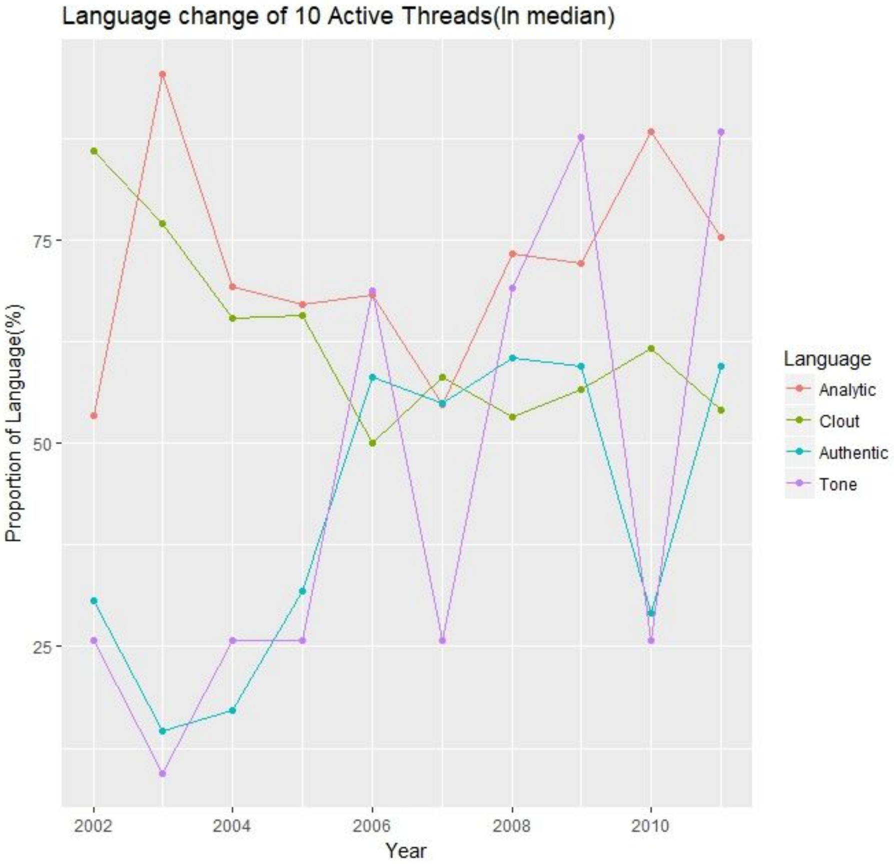

Figure 1: Language change of 10 active threads

As seen from figure 1, analytic levels have an overall increasing trend, it has increased from 53% to 75% approximately from 2002 to 2010. The largest increase takes place from 2002 to 2003, while the biggest decline is from 2003 to 2004. From 2004 to 2007, there is a gradual decrease in analytic level. It increases again from 2007 onwards till 2010, before it decreases in 2011.

Authenticity also has an overall increasing trend as the proportion rises from 30% to 60% approximately from 2002 to 2011. The level of authenticity decreases from 2002 to 2003, before it increases gradually from 2003 to 2005. From 2005 to 2006, there is a large increase in authenticity and it gradually increase from 2006 to 2009. The biggest decrease took place from 2009 to 2010 before another large increase took place from 2010 to 2011.

Tone also has an overall increasing trend as it increases from 25% in 2002 to 88% 2011. It originally decreases from 2002 to 2003, before it increases from 2003 to 2004. Second largest increase of approximately 40% took place from 2005 to 2006, and from 2007 to 2008, and the largest increase of approximately 60% took place from 2010 to 2011. There are also sudden declines taking place from 2006 to 2007 and from 2009 to 2010(approximately 60%).

However, clout has an overall decreasing trend. It decreases from 2002 to 2004, slightly increases from 2004 to 2005 before decreasing from 2005 to 2006. From 2006 to 2010, there is a gradual increase of approximately 12%. And from 2010 to 2011, the median percentage of clout decreases again.

Based on the results of our analysis, we can conclude that the language used across the 10 active threads have changed over time. We also found out that authenticity and use of tone has a similar trend from 2002 to 2011.

Now, we are going to analyse threads which are interactive by looking at the frequency of posts by each author for each Active thread as shown below in figure 2. We will be choosing threads with higher frequency of posts by a few authors.

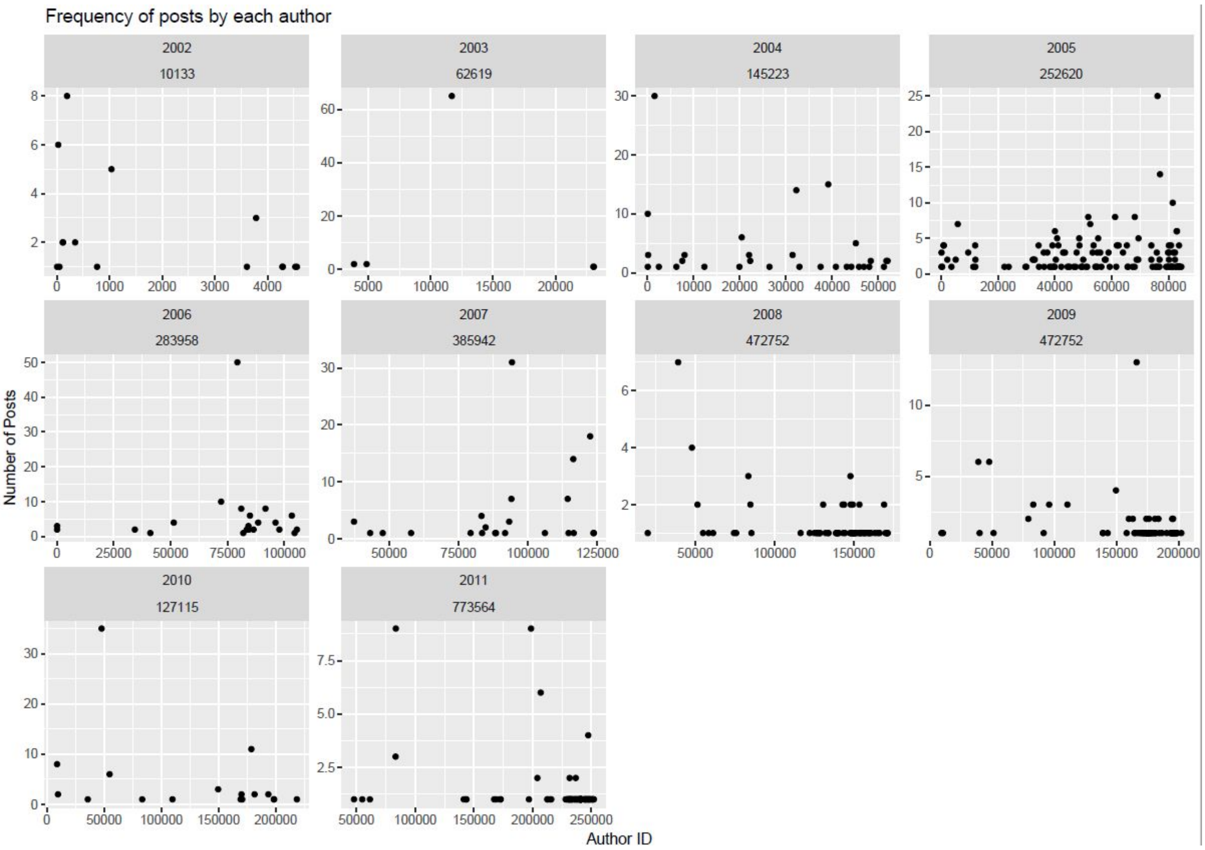

Figure 2: Frequency of posts by each author

From figure 2, we can see that for 2002, 2008 and 2011, the frequency of posts by each author are less than 10 posts. For 2003, there is only an author who posts more than 60 times but other authors posts less than 5 times. For 2009, the frequency of posts by each author is less than approximately 15 posts. So, we can conclude that ThreadID 10133 in year 2002, ThreadID 62619 in year 2003, ThreadID 472752 for year 2008 and year 2009 and ThreadID 773564 for year 2011 are not interactive because either the number of posts by each author are too little or only one author in the thread has higher number of posts while the other authors in that thread have little posts. Therefore, ThreadID 145223 in year 2004, ThreadID 252620 in year 2005, ThreadID 283958 in year 2006, ThreadID 385942 in year 2007 and ThreadID 127115 in year 2010 are interactive threads.

Using these 5 interactive threads, we want to find out how people in a single thread communicate and investigate how the language change over time for each of them. For each interactive threads, we will be taking into consideration the time span of each thread instead of restricting it to one particular year. This analysis will be done in section III below.

III. Language change in a single thread (5 Interactive threads)

With the 5 interactive thread IDs, we have plotted the proportion of language of each post in a thread ID against datetime to analyse how language changed in each thread. However, as we plotted them, we found out how data points are closely packed since we are using date and time slice as our x-axis. This could be due to some “noise” or weak signals in the dataset. It is hard to see how the levels of analytic, clout, authenticity and tone are changing and thus, we decided to perform linear regression on each of the factors. The linear equation is plotted in blue line which is shown in fig 3, fig 4, fig 5, fig 6, fig 7.

Here, we have taken into account that the number of words in each post may be too few for analytic and clout. However, we have assumed that despite the number of words in each post, it still contributes towards factors like authenticity and tone. Hence, we decided not to filter data based on word count.

Our first interactive thread ID,145223, ranges over 2002 to 2007, being most active in the year 2004. Its analytic level, clout and authenticity have decreased over the 6 years, while the level of tone increased as seen in figure 3. Therefore, we conclude that the language used in threadID,145223, has changed over time.

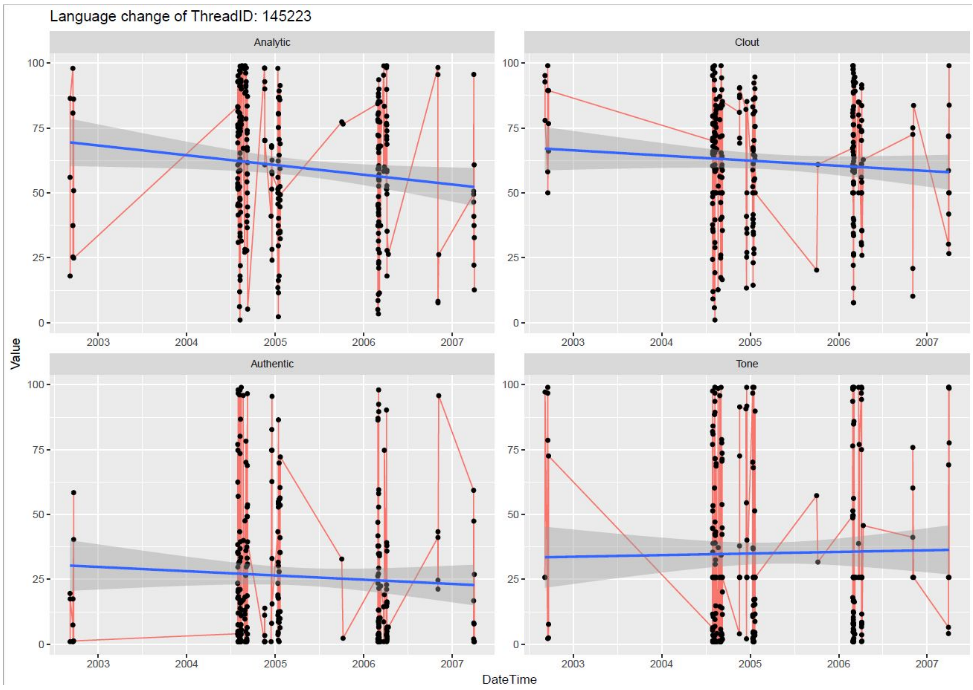

Figure 3: Interactive Thread ID - 145223

As for the threadID, 252620, we found out from the linear regression line that, analytic level, clout and tone has decreased over the time period, while its authenticity has increased as seen in figure 4.
Thus, we conclude that language used in this thread has changed over time.

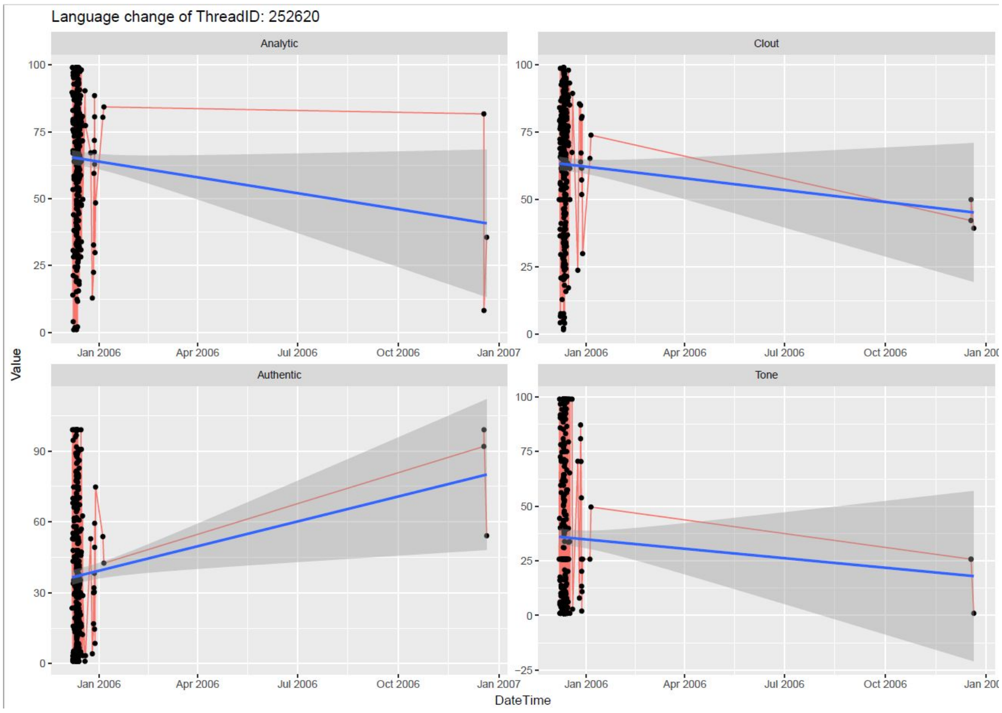

Figure 4: Interactive Thread ID - 252620

For the thread, 283958, we found out that the level of clout and tone have remained the same, while there is a decline in authentic. Only analytic level has an increasing trend over the time period as seen from figure 5. Hence, we conclude that the language used in this thread has a small change over the time period.

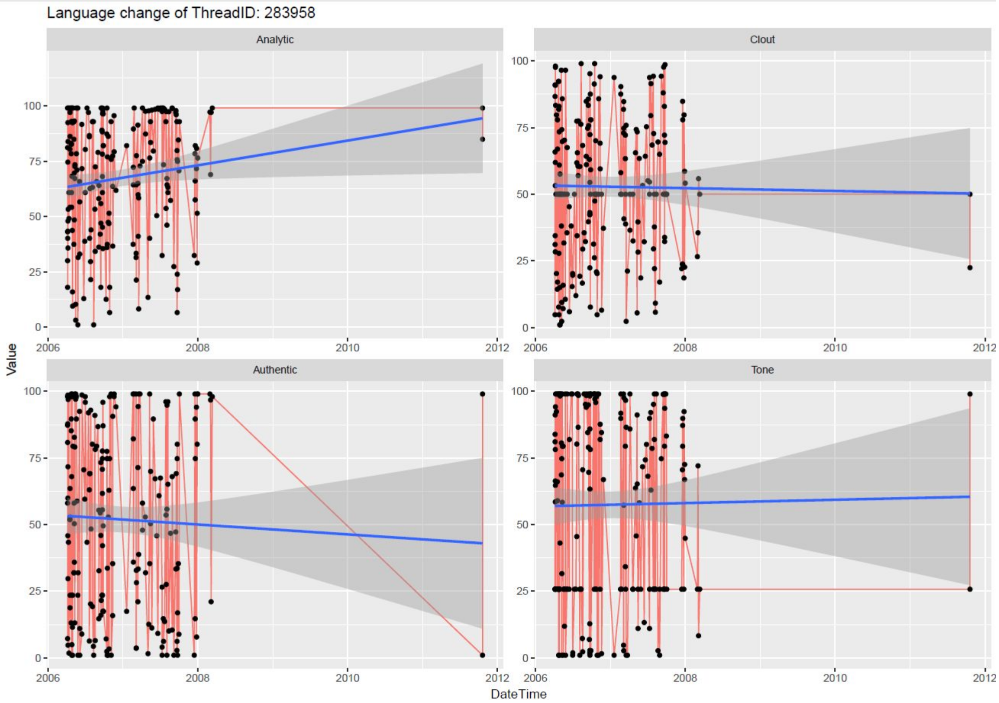

Figure 5: Interactive Thread ID- 283958

As for thread ID, 385942, there is a decreasing trend for analytic and authenticity, while the level of clout and tone increased over the time period as seen in figure 6. Hence, the language used in this thread has changed over the time period.

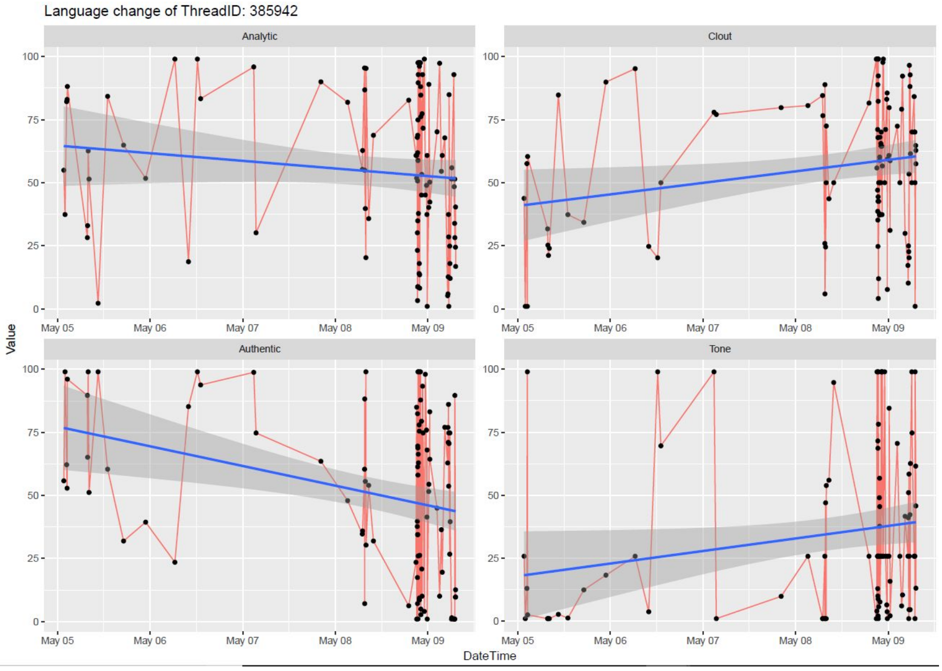

Figure 6: Interactive Thread ID- 385942

For the thread ID, 127115, we found out that the level of analytic, clout, authenticity and tone have seemed to remain constant over the time period as seen in figure 7. Hence, language used in this thread has not changed.

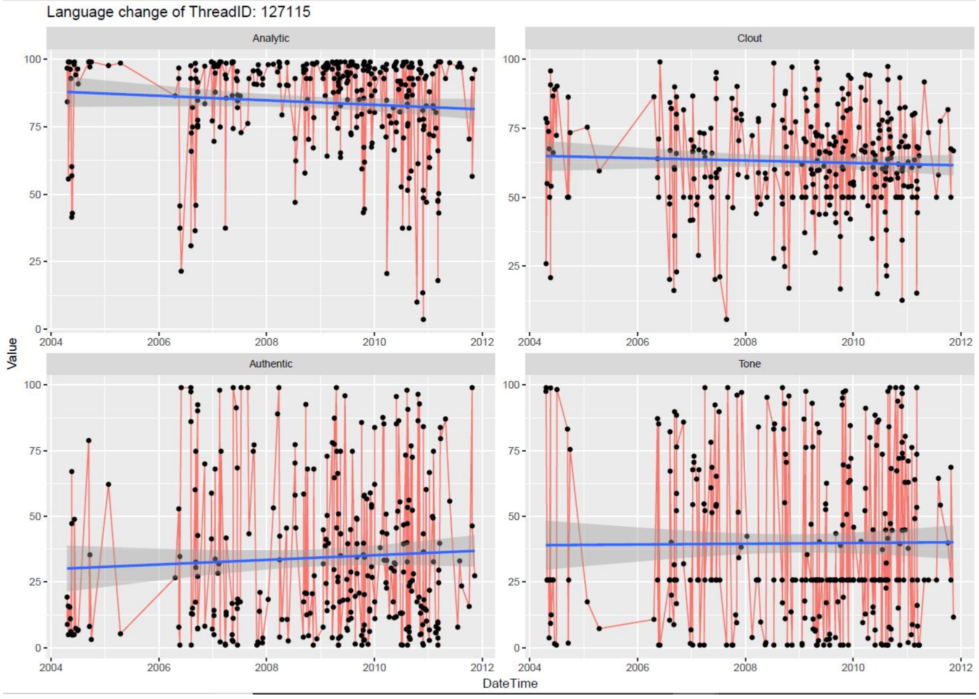

Figure 7: Interactive Thread ID - 127115

IV.Comparing Active with Non-Active threads

In this section, we will be comparing if the language used in active threads is similar to the language used in non-active threads over the years. The aim of this analysis is to observe if the language used in the most popular(i.e active) thread in that year is similar to other ordinary(i.e non active) threads in that year. We are performing hypothesis testing to aid us in this analysis which will be explained further in subsection A below.

(A) Hypothesis Testings

Null Hypothesis, H_0 : Average % of language used in active thread in that year is equal to the average % of language used in non-active threads in the same year.

Alternative Hypothesis, H_1 : Average % of language used in active thread in that year is not equal to the average % of language used in non-active threads in the same year.

Rejection criteria: p-value <= α (level of significance, 0.05)

We are assuming a 5% level of significance. We are measuring language in terms of Analytic, Clout, Authentic and Tone. Therefore, we will be doing hypothesis testing for Analytic, Clout, Authentic and Tone for active and inactive threads each year. Based on the results we can conclude if the average percentage of language on the whole used in active threads is equal to non-active threads in that year.

The table below shows the p-value computed for each year 2002 to 2011 for Analytic, Clout, Authentic and Tone.

	2002 **ThreadID 10133**	**2003** **ThreadID 62619**	**2004** **ThreadID 145223**	**2005** **ThreadID 252620**	**2006** **ThreadID 283958**
Analytic	p-value = 0.03266 <= 0.05, we reject H_0	p-value = 2.2 e-16 <= 0.05, we reject H_0	p-value = 0.03991 <= 0.05, we reject H_0	p-value= 1.313e-05 <= 0.05, we reject H_0	p-value=0.1416 > 0.05, we don't reject H_0
Clout	p-value = 0.0532 > 0.05, we	p-value= 2.384e-12 <= 0.05, we	p-value= 0.1265 > 0.05, we	p-value= 6.886 e-04 <= 0.05, we	p-value= 0.06002 > 0.05, we don't reject

	don’t reject H_0	reject H_0	don’t reject H_0	reject H_0	H_0
Authentic	p-value = 0.59 > 0.05, we don’t reject H_0	p-value = 2.833 e-07 <= 0.05, we reject H_0	p-value= 1.678e -04 <= 0.05, we reject H_0	p-value= 0.7229 > 0.05, we don’t reject H_0	p-value= 1.192e-06 <= 0.05, we reject H_0
Tone	p-value = 0.04578 <= 0.05, we reject H_0	p-value = 3.455e-07 <= 0.05, we reject H_0	p-value= 0.01011 <= 0.05, we reject H_0	p-value= 0.0005101 <= 0.05, we reject H_0	p-value= 8.745e-05 <= 0.05, we reject H_0

Table 1: p-values for each language factor from 2002 to 2006

	2007 ThreadID 385942	**2008 ThreadID 472752**	**2009 ThreadID 472752**	**2010 ThreadID 127115**	**2011 ThreadID 773564**
Analytic	p-value= 0.09837 > 0.05, we don’t reject H_0	p-value= 0.002495 <= 0.05, we reject H_0	p-value= 0.02292 <= 0.05, we reject H_0	p-value= 1.731e-11 <= 0.05, we reject H_0	p-value= 0.001587 <= 0.05, we reject H_0
Clout	p-value= 0.8086 > 0.05, we don’t reject H_0	p-value= 0.4623 > 0.05, we don’t reject H_0	p-value= 0.1912 > 0.05, we don’t reject H_0	p-value= 0.397 > 0.05, we don’t reject H_0	p-value= 0.3773 > 0.05, we don’t reject H_0
Authentic	p-value= 0.0001594 <= 0.05, we reject H_0	p-value= 1.023e-05 <= 0.05, we reject H_0	p-value= 5.686e-05 <= 0.05, we reject H_0	p-value= 0.1295 > 0.05, we don’t reject H_0	p-value= 2.449e-06 <= 0.05, we reject H_0
Tone	p-value= 0.001024 <= 0.05, we reject H_0	p-value= 5.983e-08 <= 0.05, we reject H_0	p-value= 1.837e-12 <= 0.05, we reject H_0	p-value=0.1825 > 0.05, we don’t reject H_0	p-value= 6.017e-15 <= 0.05, we reject H_0

Table 2: p-values for each language factor from 2007 to 2011

After performing hypothesis testing, for 2002, average percentage of Analytic and Tone used in ThreadID 10133 is not equal to the average percentage of Analytic and Tone used in non-active threads. But the average percentage of Clout and Authentic used in ThreadID 10133 is equal to the average percentage of Clout and Authentic used in non-active threads. This can be seen in Table 1. Since Analytic and Tone of ThreadID 10133 and non-active threads are equal but not the case for Clout and Authentic, we can’t say with certainty if the average percentage of language used in ThreadID 10133 is equal to the average percentage of language used in non-active threads in 2002.

In 2003, average percentage of Analytic, Clout, Authentic and Tone used in ThreadID 62619 is not equal to the average percentage of Analytic, Clout, Authentic and Tone used in non-active threads. This can be seen in table 1. Therefore, we can conclude that the average percentage of language used in ThreadID 62619 is not equal to the average percentage of language used in non-active threads in 2003 because Analytic,Clout, Authentic and Tone are as such.

In 2004, average percentage of Analytic, Authentic and Tone used in ThreadID 145223 is not equal to the average percentage of Analytic, Authentic and Tone used in non-active threads. Average percentage of Clout used in ThreadID 145223 is equal to average percentage of Clout used in non-active threads. This can be seen in table 1. Therefore, we can conclude that the average percentage of language used in ThreadID 145223 is not equal to the average percentage of language used in non-active threads in 2004 because Analytic, Authentic and Tone are as such and only Clout is the exception.

In 2005, average percentage of Analytic, Clout and Tone used in ThreadID 252620 is not equal to the average percentage of Analytic, Clout and Tone used in non-active threads. Average percentage of Authentic used in ThreadID 252620 is equal to average percentage of Authentic used in non-active threads. This can be seen in table 1. Therefore, we can conclude that the average percentage of language used in ThreadID 252620 is not equal to the average percentage of language used in non-active threads in 2005 because Analytic, Clout and Tone are as such and only Authentic is the exception.

For 2006 and 2007 average percentage of Authentic and Tone used in ThreadID 283958 and ThreadID 385942 is not equal to the average percentage of Authentic and Tone used in their respective non-active threads. But the average percentage of Analytic and Clout used in ThreadID 283958 and ThreadID 385942 is equal to the average percentage of Analytic and Clout used in non-active threads for 2006 and 2007 respectively. This is evident in table 1 and 2. Since Authentic and Tone of ThreadID 283958 and ThreadID 385942 and non-active threads are equal but not the case for Analytic and Clout, we can’t say with certainty if the average percentage of language used in ThreadID 283958 and ThreadID 385942 is equal to the average percentage of language used in non-active threads in both 2006 and 2007.

For 2008, 2009 and 2011, average percentage of Analytic, Authentic and Tone used in ThreadID 472752 and ThreadID 773564 is not equal to the average percentage of Analytic, Authentic and Tone used in non-active threads. Average percentage of Clout used in ThreadID 472752 and ThreadID 773564 is equal to average percentage of Clout used in non-active threads for 2008, 2009 and 2011 respectively. This can be seen in table 2.

Therefore, we can conclude that the average percentage of language used in ThreadID 472752 and ThreadID 773564 is not equal to the average percentage of language used in non-active threads in 2008, 2009 and 2011 because Analytic, Authentic and Tone are as such and only Clout is the exception.

In 2010, the average percentage of Clout, Authentic and Tone of ThreadID 127115 is equal to the average percentage of Clout, Authentic and Tone used in non-active threads. Average percentage of Analytic used in ThreadID 127115 is not equal to the average percentage of Analytic used in non-active threads. This can be seen in table 2. Therefore, we can conclude that the average percentage of language used in ThreadID 127115 is equal to the average percentage of language used in non-active threads in 2010 because Clout, Authentic and Tone are as such and only Analytic is the exception.

Based on the results, we can say that for years 2003, 2004, 2005, 2008, 2009 and 2011, the language used in active threads is not similar to the language used in non-active threads in their respective years. For 2010, however, the language used in active threads is similar to the language used in non-active threads.

For years 2002, 2006 and 2007, we are not certain if the language used in active threads is similar to the language used in non-active threads based on looking at its mean. In this case, median happens to be the next best central tendency after mean. So, we are going to use boxplots to compare if the median proportion of language used in active thread is similar to the median proportion of language used in non active threads. By using median, we hope to get a more definite answer on whether the language used is similar for 2002, 2006 and 2007. This will be explained more in depth in subsection B below.

(B) Boxplots

In this section, we are using Median to compare if the language used in active threads is similar to the language used in non-active threads.We are going to do boxplots for year 2002, 2006 and 2007.

As shown in Figure 8, median percentage of Analytic level of other non active threads (~60%) is higher than the median percentage of Analytic level for Thread 10133 (~53%). Median percentage of Authentic level of Thread 10133 (~30%) is higher than the median percentage of Authentic level of other non active threads (~27%). Median percentage of Clout level of Thread 10133(~87.5%) is higher than the median percentage of Clout level of other non active threads (~70%). Median percentage of Tone level of Thread 10133 and other non active threads (~27%) are approximately equal. Therefore we can conclude that the median percentage of language used in Thread 10133 is not similar to the the median percentage of language used in other non active threads.

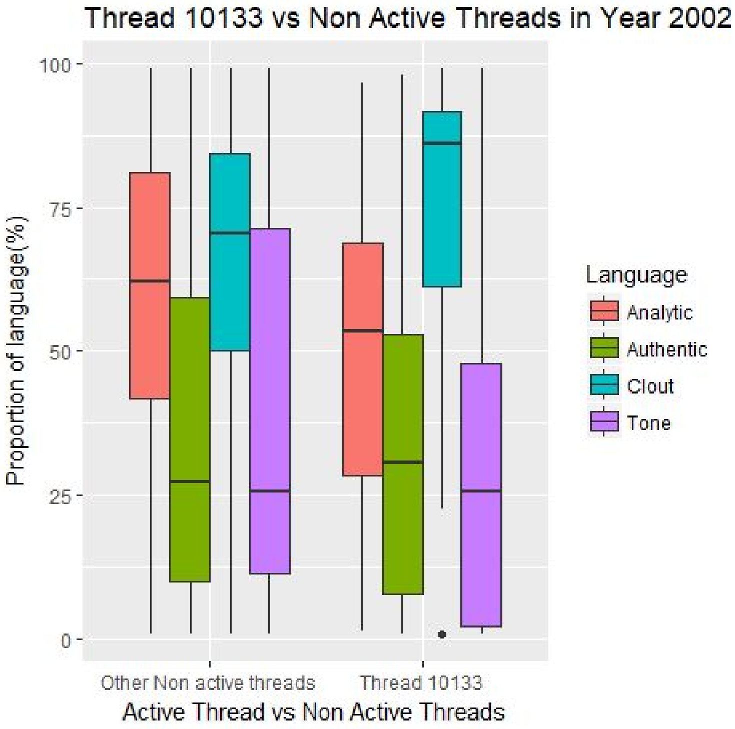

Figure 8: Thread 10133 vs Non-Active Threads in 2002

As shown in Figure 9, median percentage of Analytic level for Thread 283958 (~72%) is higher than the other non active threads (~65%). Median percentage of Authentic level for Thread 283958 (~ 58%) is higher than other non active threads (~28%). Median percentage of Clout level for Thread 283958 (=50%) is lower than the median percentage of Clout level for other non active threads (~58%). Median percentage of Tone level for Thread 283959 (~70%) is higher than the median percentage of Tone level for other non active threads (~26%). Therefore we can conclude that the median percentage of language used in Thread 283958 is not similar to the the median percentage of language used in other non active threads.

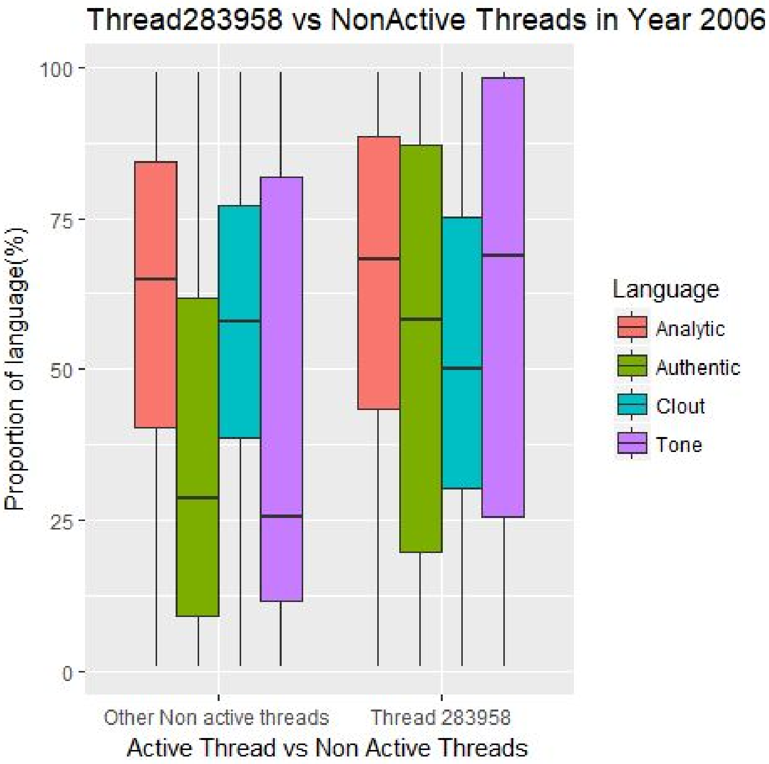

Figure 9: Thread 283958 vs Non-

As shown in Figure 10, median percentage of Analytic levels for Thread 385942 (~55%) is lower than the median percentage of Analytic levels for other non active threads (~ 62.5%). Median percentage of Authentic levels for Thread 385942 (~55%) is higher than the median percentage of Authentic levels for other non active threads (~30%). Median percentage of Clout level (~60%) for Thread 385942 is higher than the median percentage of Clout level for other non active threads (~58%). Median percentage of Tone level for Thread 385942 (~26%) is the same as median percentage of Tone level for non active thread(~26%). Therefore we can conclude that the median percentage of language used in Thread 385942 is not similar to the the median percentage of language used in other non active threads.

Active Threads in 2006

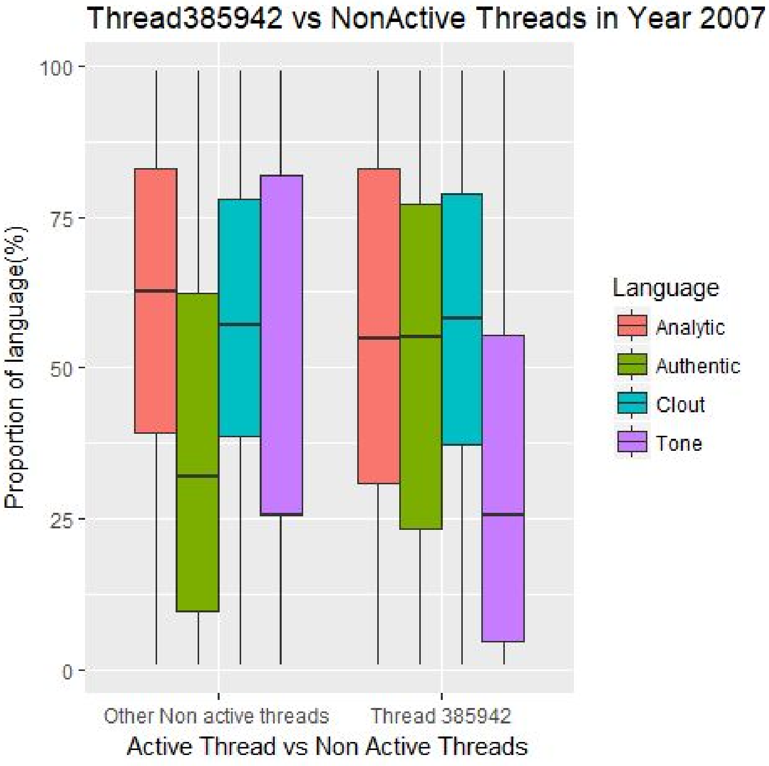

Figure 10: Thread 385942 vs Non-Active Threads in 2007

Based on the results, we can say that for years 2002,2006 and 2007 the language used in active threads is not similar to the language used in non-active threads in their respective years.

V. Conclusion

In general, when looking at this online forum, we can conclude that the language pattern used by members who are communicating directly with each other in a thread over time is not similar. We proved this by investigating 5 interactive threads individually, whereby the language used in the thread has changed over time for almost all threads except ThreadID 127115. Also, when we analysed if the language used in active threads is similar to the language used in non-active threads for each year, we learnt that for most of the years, the language used in active threads is not similar to the language used in non-active threads except for year 2010. What makes these results interesting is that the active thread in 2010 happens to be ThreadID 127115. This actually goes to show that in the year 2010, the language patterns used across different threads over time are similar and also the language patterns used in ThreadID 127115 is similar over time. Despite the exceptions, when we look at it on the whole, we can deduce that the language pattern used by members in a thread over time is not similar.