Participation Analysis Project

Computer Science 683

Professor Sabine Graf

Athabasca University

Aidan Polese

Email: [apolese1@athabasca.edu](mailto:apolese1@athabasca.edu)

Student Number: 3594858

Abstract

During the first 6 weeks of students in the course COMP 683 were encouraged to participate in weekly discussions to demonstrate their understanding of the current weeks course content as well as discuss their opinions on course topics as well as talk to their peers about the content presented that week. For this project I took the information available from the course Moodle page over the 6 week study period. I analyzed each of the student’s participation statistics to understand the relationships between student participation and interactions with each other. I also aim to compare other students’ participation levels with my own to see how I have contributed to other people’s learning and course activity levels.

Introduction

Motivation

Student interaction in a university setting is integral to student understanding and perspective on course material. I personally have experienced this both in an in-person and online setting. As this is an online course, I will showcase how students interact in an online setting and which information is conceived to be the most important or most interesting to students. The goal is to find which information communication strategies are the most effective and how different students go about achieving this.

Goals

The research I am going to be conducting focuses on how students interact with each other and if any patterns can be extracted from this data. I am to display which topics students perceive to be the most important and interesting. I will also be comparing this data with my statistics to see how I contribute to student engagement and participation.

The data I will be examining is as follows:

**RQ1:** How much do students contribute to discussions?  
**RQ2:** What topics are students most interested in?  
**RQ3:** Which students are most interested in other students’ discussions?

Overview

I decided to focus on a macro perspective on the course information as I believe it to be most indicative of student interest and participation. I will be covering how my research is conducted, including how I manipulated and analyzed data for this research and the results of the analysis. I will finally summarize and make conclusions about my research and discuss the results.

Research Methodology, Design, Execution Summary

Given my research questions and goals, each question can be quantitively calculated, displayed and analyzed. Each question can also be quantitatively assessed by drawing conclusions from correlated data, as well as make extrapolations based on the results.

**RQ1:** How much do students contribute to discussions?

This **RQ** will show which students contribute the most to discussion forums on a post-by-post basis. This will include the metrics, information contributed, and average information contributed overall.

**RQ2:** What topics are students most interested in?

This **RQ** will be conducted by extracting keywords and topics within posts and replies to examine which topics come up the most. This could be indicative of student interest. I will then see who mentions the most topics to see which student is the most well rounded. I will also be categorizing students based on their interests to see if it is possible to form clusters of like-minded students.

**RQ3:** Which students are most interested in other students’ discussions?

This **RQ** will be conducted by examining how students reply to each other. This will include examining frequency of replies, as well as who responds to who. This will be done by creating a network graph of people post and reply connections.

Data Sources

I have opted to only use the data provided by the course Moodle page as I believe it to be the most suited to my study. The Moodle focuses on mainly course content and discussions related to course topics and interests. The landing information was not exclusively course material discussions, and because there is a lot of information present that needs to be examined, the landing was excluded from this study. I did a bit of data cleaning to remove words and sentences that were people signing their names or saying things like “thank you for your post” to keep the data more cohesive. I also removed any non-alphanumeric symbols.

Tools

For this study, I used quite a few tools. The main tool I used was a personally written Python script [1] that allowed me to both consolidate, organize, and analyze the data. The code has been included and is commented to explain how it works. I used quite a few Python packages as tools to visualize and analyze the data including, Matplotlib [2], Pandas [3], Scikit-learn [4], Numpy [5], AdjustText [6], and Wordcloud [7]. I also used Netlytic [8].

Data Organization and Tool Usage

Data Storage

Originally, data is presented on the course Moodle page in an html format. The information can also be extracted in an rss format. I personally am not a fan of either of these formats when doing data manipulation, so I converted the weekly discussion data into a json format as I believe it is the easiest to work with. An example of this data can be found on the GitHub where my code is stored [1]. There is also a sample in Fig 1.

Side Note:

Some figure text can be quite small since usernames can be so large, accommodating them through code results in some unorganized text locations, so I have left them as is. Files for the figures have been included on the GitHub as they are easier to read.

Fig 1: Sample Post-Reply Chain Text

Description automatically generated

From here, I used the script *Extract-Information.py* [1] to conduct my analysis for my research questions.

Tool Usage

My personal python tool is used as a data extraction tool from the discussion forum information json. It then organizes and analyzes the data. This analyzed data is then funneled into different Python libraries which serve as both visualization tools as well as analysis tools. Matplotlib [2] is a library which is used to create data visualizations. In my case I used in primarily to create graphs. I used Pandas [3] simply as an array storage tool to hold my data. I used Scikit-learn [4] to reduce the features of n-dimensional feature data down to a 2-dimensional data point to make visualizing plots and clusters easier. NumPy [5] is used to organize data formats. AdjustText [6] is used to space text out on Matplotlib properly. Wordcloud [7] is used to create word clouds. Netlytic [8] is used to create cluster visualizations. Exactly how I used these libraries within my code can be complicated at times, in each **RQ** I will summarize how it works for each section. I encourage readers to read through the documented code to fully understand how I conducted this study. I believe it to be the best way.

Research Methodology and Results

**RQ1:** How much do students contribute to discussions?

This **RQ** will show which students contribute the most to discussion forums in both posts and replies. This question shows who has contributed the most in terms of volume to discussion posts as well as who is answering and inquiring in the reply sections of people’s discussion posts.

To measure this, I went through each week’s discussion posts and replies and then counted how many words each person had contributed over the 6-week study period. The results of which are shown in Fig 2. I then also calculated the average words per post for a person each week, and the average words they contributed each week including replies. This is shown in Fig 3.

Fig 2: Total Word Contribution Per Person

Chart, bar chart

Description automatically generated

Fig 3: Average Word Contribution Per Person

Chart, bar chart

Description automatically generated

For each person, the total information contributed varies quite a bit. With the lowest value being 544, and the highest being 5891. We can see that Robert-Kemp has contributed the most by far in terms of raw information provided. However, it should be noted that by looking at Fig 3, we can see that most of this information comes from the amount of information he contributes in the form of replies. Summing the data, we can see that the amount of information contributed from Fig 1 gives 43,065 words, the total average being 2,692 words. We can see that 7 people contribute over the average, atulch, Robert-Kempt, kennethmc39, janineis1, Sheryl-Girffith, George-Adyrhaiev and myself. I stand above the average at 2970 words over the study period. There is quite a disparity between how many words per person is contributed from Fig 2. We can also see that in Fig 3, not may people have a very high average of words contributed including replies. Of course, replies are much smaller compared to posts, but this indicates that most people in the class might consider replying more to encourage more conversation to aid with learning.

We can also see that in Fig 3, that including my replies, my average information contributed is only slightly higher from 417 average words per post, to 495 average words overall each week. This would indicate that I personally should try and reply more often or be more detailed. I know that I try to reply a few times per week. However, to catch up with my peers such as Robert-Kemp, Kennethmc39 and George-Adyrhaiev, I should be replying more to be an exemplary student and encourage conversation.

**RQ2:** What topics are students most interested in?

This question aims to find which topics students are most interested in. This includes the most popular topics between all students in their discussion posts, as well as their replies to other students. This way, each student demonstrates what they know, what they are interested in, and what they are discussing in replies.

This question is where the process becomes a little more involved. If you were to see which words were most common, you would get a list of words like *a, of, the, which, there, I,* etc. To remedy this problem, I used a list from EspressoEnglish [9] of the 100 most common words to add to a restricted list of words not to be used as keywords for the word cloud. I also know that there are phrases that are important to this class like *educational data mining, machine learning, internet of things, big data,* etc. I then went through all posts and replies searching for the most common words and phrases being used and compiled them into an auto generated list of all significant words, 2-word and 3-word phrases. I then used Wordcloud [7] to generate a word cloud based on these phrases, see Fig 4. I also created a list of what I deemed ***Important words and phrases*** called ***IMPORTANT\_TOPICS***that have been hand qualitatively filtered from the most common words and phrases. These were excluded when generating the word cloud and for other parts of this study.

Fig 4: Discussion Forum Word Cloud

A picture containing text

Description automatically generated

From the word cloud wee can see that people are most interested in things like educational data mining, linked data, internet of things, student experience, tools that are being used to conduct research and much more. From Fig 5, we can see that I was most interested in e-learning and machine learning by far, and then a little bit into how they interact with big data.

Fig 5: Personal Keyword and Key-Phrase Usage Statistics

Text

Description automatically generated

Compared to other students, I seem to be much more interested in machine learning and e-learning as they do not appear in a large way in the word cloud. This makes my discussions a little more targeted and unique, which might explain why my posts and replies don’t gain a lot of traction.

Fig 6: Total Topics and Important Topics Covered

Chart, bar chart

Description automatically generated

Going by cleaned phrases and keywords alone, we can see that people have differing interests based on the blue bar heights. We can see that atulch covers the most information, followed by Rober-Kempt, janineis1 and then George-Adyrhaiev. This falls fairly in line with how much information they have contributed in their posts and replies. However, something interesting to note is that in number of topics covered, atulch is leading by a large margin. Going into the data, we can see that his posts from each week cover a lot of varied information. However, we can also see from Fig 3, that he has not replied to anybody. Perhaps this student might consider replying and discussing information with other students, like the second information leader, Robert-Kemp who has replied to many people and covers the most keyworded information. 10,801 is the total amount of keyword topics covered, with the average being 675. Only 5 students sit above this number, including myself. This might indicate that other students might consider delving a little deeper into other topics. For me personally, I believe this number should indicate that I should reply more to other students, based on my conclusion from **RQ1**, to encourage discussion with other people on more varied topics.

I do recognize that by going on raw phrases covered can be quite ambiguous, so referencing my curated ***IMPORTANT\_TOPICS*** list, I then compared to see which students touched upon these topics. The total number of phrases that I deemed important is 245, which can be seen in the code as a constant at the top. The total number of important phrases mentioned overall is 476, with the average per person being 30. I sit close to 30 at 29. We can see that most students are quite close to this number, so it can be deducted that students are all making the same important takeaways that I am also concluding from the course content. This also indicates that all students have similar ideas of what the key information to understanding this class is.

Fig 7: Total Topics Mentioned Scatter Plot

Chart

Description automatically generatedI have also clustered students based on their interests. Using each keyword and phrase as a feature I created points for each student. From there, I wanted to visualize the data but data points consisting of hundreds of phrases are not immediately easy to plot. To do this, I used Scikit-learn [4] to use PCA to transform the data down to 2 dimensions. From here I was able to plot people’s keyword and phrase usage to a plane that could be easily visualized, see Fig 7. We can see that most students examine similar topics when looking at all phrases except for kennethmc39, George-Adyrhaiev and atulch. This is likely because of the both the number of topics they bring up, as well as the uniqueness of them. In this total view, I too fall within the major cluster.

Fig 8: Total Important Topics Mentioned Scatter Plot

Chart

Description automatically generated

Moving on to measuring the important topics, we can see that the clusters are more varied with smaller sub clusters forming, as well as individuals who move far away from other people becoming outliers. For myself, I move far away as the topics I am interested in like machine learning and e-learning (see Fig 5) are not popular with my peers which moves me far away from the main cluster, but not so far away as atulch and George-Adyrhaiev who we have established brings up varied topics due to the volume of information they put on to Moodle.

**RQ3:** Which students are most interested in other students’ discussions, can they be categorized?

This question is designed to see which students participate in discussions with other students and if they share the same interests. The end goal is to cluster students in terms of what they are interested in based on replies. Previous **RQs** have explored this information throughout all posts and replies, regardless of if they were directed at a specific person or not. This **RQ** will analyze and cluster people who have made conscious decisions to discuss and interact.

Starting by counting total replies in Fig 9, we can see again that Robert-Kemp is leading the pack with 26 replies over the 6-week period followed by kennethmc39 and then Sheryl-Griffith. We can see that there is a total of 122 replies made total, with an average reply rate of 8 per person. 7 students sit at, or above that average. This indicates that perhaps students might consider replying more to encourage conversation and discussion on topics. I land right on the average, so I too need to begin to reply more to encourage discussion on topics that I am most interested in as I have more niche interests as shown in Fig 8.

Fig 9: Total Replies Per Person

Chart, bar chart

Description automatically generated

Fig 10: Netlytic Post-Reply Network Graph

Chart, radar chart

Description automatically generated

Using data based on who replies in each post, see file *PostReplyNetworks.csv*, I used Netlytic [8] to create Fig 10. This shows who has replied to each person’s posts within each week’s discussion forums. We can see that there are many clusters indicated by each colour. Weirdly enough, the largest cluster is the red cluster of which I am part of. This indicates that me and the people in my cluster with larger circles reply to the most varied amount of people. However, people with small dots have replied to a very small number of people. This would indicate that people in my cluster, including myself, have gone out of our way to respond to less active posts. This overall might indicate that students might consider replying to more people instead of people that they already communicate with a lot. I am doing well in this regard, so I am happy with my own performance.

Discussion

I believe my main goal of this study has been adequately answered as I have compared all students across the study period. When communicating, there seems to be a large disparity between students in the course. Some students have most of their information contained within replies, like Robert-Kemp. Some may not have replies at all but cover vast amounts of information like atulch.

I believe it is reasonable to say that largely, when posting and talking to each other, students tend to cover the same *important* material. However, when it comes to more niche material, students tend to vary quite a bit. This seems to be related to post length as large posts will naturally cover more content.

It does seem that replies don’t have a big impact on information variance as they appear to be talking about the same subjects. Mostly agreement and minor variations of the topic. From a pure quantitative standpoint, this makes sense. I believe it might be worth it for students to introduce ideas that are within the same topic but tackle it from a different angle which could increase variance and provide a better spread of information for discussions.

I also am not surprised by the fact that the most active students tend to mostly talk to each other and not to students who don’t interact fully with others like atulch. It can be hard to talk to people who don’t look like they have an interest in replying, so some might not see the point in commenting on their posts. I do understand that perspective, but I personally do see some merit in it as people like atulch cover many topics. So, even replying as a thought experiment can be constructive. However, a lot of people do like talking to George-Adyrhaiev because he also covers a lot of material and replies to people’s inquiries.

Similarly, Robert-Kemp contributes large amounts of information as well as replying to many people. It can be argued that he is a main driver in many discussions on the forums due to just due to his reply volume to many different people containing lots of information.

Conclusion

This report has covered the following research questions:

**RQ1:** How much do students contribute to discussions?  
**RQ2:** What topics are students most interested in?  
**RQ3:** Which students are most interested in other students’ discussions?

The main goal was to examine each question by looking at all students and then with myself. I do believe that they have all been adequately answered.

For **RQ1**, my contribution rate sits above most students however, my reply information volume is lacking, so I look to improve this in the future by replying to more people as well as making more detailed replies.

In **RQ2**, the topics which I am interested in don’t seem to be totally aligned with what most students are interested in, but I am still covering the more important topics which still leaves me having the same takeaways as others.

For **RQ3**, I am lacking in number of replies to other students, but I do have one of the larger reply spreads. Many students tend to talk to more active students but I do like replying to threads that don’t have any replies just so the poster might be able to see them, even if they don’t answer.

On almost all participation stats, I sit above average or at average. However, when it comes to replies, this seems to be my largest weakness. Even though my stats are still average or above, I do stand to make some improvement when compared to my peers who have very high stats.

I think overall I have contributed a lot to this course in the form of unique content that is related to the course but not always directly presented. I think this makes a good comparison point for others.

References

[1] Polese, A. (n.d.). *Aidanpolese/Participationanalysis*. GitHub. Retrieved from https://github.com/AidanPolese/ParticipationAnalysis/tree/master

[2] Matplotlib Development team. (n.d.). *Visualization with python*. Matplotlib. Retrieved from https://matplotlib.org/

[3] NumFOCUS. (n.d.). Pandas. Retrieved from https://pandas.pydata.org/

[4] Microsoft. (n.d.). scikit-learn. Retrieved from https://scikit-learn.org/stable/

[5] NumPy. (n.d.). NumPy. Retrieved from https://numpy.org/

[6] Flyamer , I. (n.d.). *Phlya/adjusttext: A small library for automatically adjustment of text position in matplotlib plots to minimize overlaps.* GitHub. Retrieved June 26, 2022, from https://github.com/Phlya/adjustText/

[7] Mueller, A. (n.d.). *Amueller/word\_cloud: A little word cloud generator in python*. GitHub. Retrieved from https://github.com/amueller/word\_cloud

[8] Netlytic. (n.d.). Netlytic. Retrieved from https://netlytic.org/home/

[9] Espresso English. (n.d.). *Learn English fast in a few minutes a day!* Espresso English. Retrieved from https://www.espressoenglish.net/