A Sneak Peek at a Collaborative CS Project: What part do female students play?

ANONYMOUS*

We present an exploratory analysis of how different genders collaborate while working on programming projects in a Java Programming Course of a south-eastern research-intensive US university. In the projects we studied, when students collaborated on Github, they declared what aspects of the project they worked on by including a tag in each commit message. Some aspects of the projects involved intensive coding whereas other did not. Our study analyzed these tags to determine what parts of the project each gender is prone to working on. We also performed predictive analysis to determine if elements of commit messages and tags are significant predictors of gender and vice versa. Our findings from the class we studied, show that gender is not a factor when it comes to what each student worked on. This could serve as a motivation for female high students who are considering choosing a career in Computer Science.

CCS Concepts: • **Social and Professional Topics** → **Professional Topics**; User Characteristics; User Characteristics ; • **Computing Methodology** → *Machine Learning* .

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Anonymous. 2018. A Sneak Peek at a Collaborative CS Project: What part do female students play?. In *CSCW 2021.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Computing is a male dominated field and many studies have reported on the gender gap that exists in Computing related courses. Beyer et al. 2008 in their book titled: "Women in Computer Science or Management Information Systems Courses: A Comparative Analysis" asked why so few women major in computer science?' They explained that gender imbalance is a complex issue and the reason for the gender gap is yet to be understood despite numerous publications on the subject [2, 4, 12, 13, 17]. Kamberi (2017) analyzed over 30 research studies investigating women in computing and produced the 4 Es model (expose, engage, encourage, empower) to entice women and spur their interest in computing [11]. One challenge is that computer programming is negatively viewed by most female students and they consequently avoid programming related fields[16]. Another challenge is the lack of confidence of women who are curious about going into computing related fields [11]. Kamberi recommended that programming should be introduced to girls at an earlier age. They also suggested that educators should encourage high-school female students to pursue computing majors in college [11]. Another solution involves introducing a gender-inclusive college curriculum which is a curriculum that acknowledges the different learning patterns and backgrounds of all genders [13]. Lessons learnt from Kamberi's survey suggested that early exposure, engagement, encouragement of female students by educators and empowerment

*

CSCW, Virtual, 2021

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

through mentoring would lead to an improved representation of women in computing related fields [11].

In this current study, we seek to understand what goes on during programming assignments in a second-year programming class in College. We are specifically interested in exploring what part female students play while working on a programming project in multi-gender teams in this class. A programming project consists of several parts such as the implementation of the project requirements, bug fixing, unit and system testing, code cleanup and documentation. In this class, students work in teams to achieve a programming task. Teams collaborate over Github, a decentralized version control platform for projects that allows different members of the project team to collaborate seamlessly by resolving changes to be merged and resolving resulting conflicts in an efficient manner [3]. GitHub has been embraced by the open source communities and is being employed as the major online learning tool for collaboration in programming related college courses [19]. The platform is increasingly vital for managing software projects. Several research studies have shown the effectiveness of Github in transforming the learning experience [19]. In addition, Github provides a system where every team member has a version of the project as well as a traceable project repository where each member can commit code changes to the project repository. Each commit is required to have a commit message which provides a basic description about the code being added to the repository. We aim to explore the contributions and portions of the code worked on by male and female students in mixed-gender teams based on their commit messages on Github.

Terrel et al. in their paper: "Investigating the Effects of Gender Bias on GitHub" discovered that female contributions on Github tend to be accepted more often than contributions by their male counterparts. However, when a female's gender can be identified, they are rejected more often [10]. Seeing that a bias exists in how gender efforts are perceived when gender is known, we want to know if this bias in turn affects females' choice of programming tasks while working in teams consisting of both genders in a blended learning environment. This paper reports tasks each gender chooses to work on while working in a multi-gender team. We investigate programming projects in one semester and seek to detect evidence for whether female students are prone to performing less programming intensive tasks, such as documentation, or styling, as opposed to the main implementation, debugging, resolving static analysis notifications or testing tasks. In this work, we investigate commit messages for team projects in the Fall 2020 semester of a CS 2-OOP class [14]. Prior research on previous offerings of the class, showed that the commit messages sometimes included random texts that do not really explain the code committed [9]. Therefore, in the beginning of the Fall 2020 semester, we asked the class instructor to hint students on some good examples of meaningful commit messages. Hence, in addition to the commit messages the instructor included a tag between brackets. For example: if a commit message says: "committing a fix for linked list error", the students were recommended to commit the following instead: "[Debugging] committing a fix for linked list error". Another example is the commit message "adding code for listview" becomes "[Implementation] adding code for listview". We want to see if a particular gender is prone to performing one type of task relative to the others. Prior research also discovered three teamwork habits among teams namely cooperative, collaborative and free-rider [5, 8, 9]. Coman et al. described two forms of habits Cooperative, where each team member supports each other while they work on different tasks in a project and Collaborative, where the team members share the same tasks while completing a project [5]. There is a third teamwork habit described by van der Duim called free-rider where one team member does the majority of the tasks while the other members do little or nothing [8]. We want to detect what habits are common among multi-gender teams. If Cooperation exists, we want to know what task each gender agreed to focus on. A positive result, which is collaboration on all tasks or cooperative habits where the

programming tasks are not divided along gender lines, would serve as a great encouragement to female students with a negative view of programming and programming related fields. A negative result which is either cooperative habits with females working on less programming intensive tasks or a free-rider situation where a female team member does nothing might signal the need for an intervention, perhaps males should learn to share work more or more preparedness of female students before entering programming intensive fields.

It is important to note that academic performance and level of programming experience are other factors that affect a student's choice of task to work on but we want to see if there is a relationship between gender alone and the choice of tasks in the team project of the class we are investigating. This study is broken into two parts:

- Exploratory Data Analysis of commit messages
- Predicting Gender based on commit messages and vice versa

Our exploratory data analysis (EDA) answers the following research questions:

- RQ 1: How did each gender respond to the recommendation to tag their commit messages?
  - (a) How many students actually tagged their commit messages?
  - (b) Were females prone to following that recommendation or not?
- RQ2: In the class we studied, were tasks assigned to team members biased along gender lines?
  - (a) Did female students do more documentation and less programming?
  - (b) Did one gender generally do more of one task than the other?

  For our prediction task, we ask:
- RQ3: Can we predict the gender (M/F) of the authors using details of the commit messages as predictor variables and can gender predict the commit tag?

2 RELATED WORKS

2.1 Gender differences in Choosing Computer Science

In a research study done by researchers from the University of Wisconsin, female students had less confidence in their ability to learn computer science (CS) concepts than male students in CS. In fact, they felt less confident than the male students in non-computer science majors. The study also found that female students who have supportive friends and family members are more likely to major in computer science. The authors then concluded that they speculate the reassurance of one's competence by others may increase the probability that a woman musters the courage to enter into a male dominated major such as CS [2]. A study by Du et al. showed that males were exposed to programming at an earlier age than females. They also found that after a one hour coding tutorial, males were more likely to take a programming course than females even though females performed better than or equaled males in the programming quizzes after the hour of code [7]. This current work studies what parts females play in a collaborative CS college project. An insight into how females work on programming projects might embolden other female students to take up CS in college.

2.2 Gender differences in Collaborative Projects

In their paper, Imtiaz et al. demonstrated the capability of women in the open source community. The findings showed that women tend to have their pull requests accepted at a higher rate than men [10]. In fact, women's pull requests are larger, indicating more lines edited. Vasilescu et al studied the effect of gender and tenure on productivity in Github Open source software (OSS) development communities. They hypothesized that gender diversity has a positive effect on productivity and a negative effect on turnover while tenure diversity has a positive effect on both productivity and turnover. To prove their thesis, they conducted surveys and measured productivity by the number

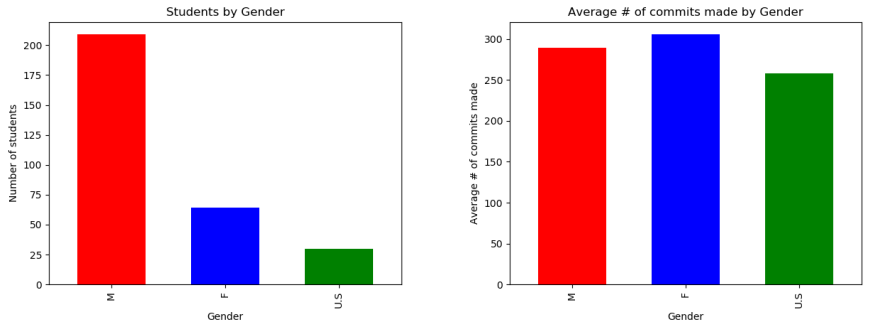
of commits by each team member. By using mixed effects analysis to measure significant predictive abilities of gender on productivity and turn over, they confirmed part of their thesis that gender diversity has a positive effect on productivity. However, they are unable to prove the negative effect of gender diversity on turnover [6, 18]. This current study applies an exploratory and predictive approach to exploring the productivity of female students while working in multi-gender teams.

3 METHOD

3.1 Pre-processing

This work explores the programming activities of students in the Fall 2020 semester of a CS 2-OOP class [14]. Our dataset consists of two files, Gender file and Commit Messages file. The Gender file provided by the course instructor contains mappings of author_IDs to students' gender. Missing values in the gender column were labeled "Unspecified" (U.S.). As shown in Figure 1(a), there were 303 students in the class. 209 males and 64 females, and 30 student's gender were U.S.. While this makeup seems highly unbalanced at first glance, these metrics are representative of the current departmental makeup, which thereby assure us that our analysis population/sample was representative of the department as a whole.

Fig. 1. (a) Statistics of the class by Gender (b) Average commit message for each Gender



(a) (b)

We collected the dataset containing commit messages directly from Github. This contains the author_ID, repo_name, commitID, commitTime, commitMessage, lines added, deleted, modified, copied, removed and typeChanged. Using the author_ID, we merged the two files into one dataframe. We dropped the columns typeChanged and commitID since they weren't relevant to our research question and sorted the dataframe by repo_name to make it easier to narrow our research on teams later on. There were a total of 1,942 unique repositories and 87,765 commits throughout the semester. 27 duplicates observations were removed resulting in 87,738 observations. Figure 1(b) shows the average number of commit made by gender. Female made more commits on the average than males in the class we studied.

3.2 RQ1: How did each gender respond to the recommendation to tag their commit messages?

3.2.1 Commit messages, cleaning & Tag Categorization. Of the 303 students in the class only 99% of both male and female students made commits to the course projects on Github and 90% of U.S. gender made commits on the platform. The minimum commits within the class for each gender were roughly similar, 2 for U.S., 10 for females and 9 for males. Males had the maximum number

Table 1. Statistics of Commit by Each Gender

Gender	Min	Max	Mean Lines Added	Mean Lines Modified	Mean Lines Deleted
Male	9	785	17,465	11,471	1,540
Female	10	693	18,741	12,479	1,647
U.S.	2	645	18,799	11,963	1,107

Table 2. Commit Categories, Keyword and Definition

New Tags	Original Tag Keywords	Definition
untagged		The commit message did not contain a tag
versioning	*git, version, merge*	The commit message is about versioning
documentation	*doc, deploy, comments, readme*	The commit message is about how the new update describes the code
style	*style, design, gui, uml*	The commit message is about front end tasks
implementation	*impl, impel, impe, optimization, integra, fsm, constructor, studentrecordio, tried, refactor*	The commit message is about main implementation tasks
debugging	*fix, bugs, revert, rollback, debug, exception, throw, error, deb, debug, degug, pmd, pls*	The commit fixes a bug, a.k.a bugfixes
testing	*test, random, hope, jenk, junit, unit, bb, box, black, coverage*	The commit tests the code update
skeleton	*skeleton, skel, libraries, initial, collection, setup, start*	This is usually the first commit or starter project
misc		The tag given by the student was not meaningful and did not fall into any category

of commits by a single student (785), maximum number of commits by a female student was 693 and U.S. gender, 645. On average, within the overall class female students made 5% more commits compared to males. As shown in Table 1, average male contributor added 17,465 lines of code, deleted 1,540 lines of code, modified 11,471 lines of code per male contributor, and removed 366 files.Females made more edits than males in every category. The average female contributor added 18,741 lines of code, deleted 1,647 lines of code, modified 12,479 lines of code and removed 772 files. The average U.S. gender contributor added 18,799 lines of code, deleted 1,107 lines of code, modified 11,963 lines of code and removed 348 files. The first gender to commit, commits the starter project provided by the instructor which has about 600 lines of code. The statistics above might be influenced by these, therefore we separated the commit messages into different tags/categories as described in Table 2. The category with of the starter project is called 'skeleton'.

In the Fall 2020 semester, students were expected to include 5 specific tags namely Implementation, Testing, Debugging, GUI, and Documentation, in their commit messages whenever they pushed to Github. We iterated over each commit and used a regex search to see if the student had tagged their commit. Only 51.6% of all commit messages were tagged. To answer RQ1a, 51.2% of all commit

messages made by females were tagged and 50.6% of all commit messages made by males were tagged. 24 students did not tag any commit message at all and all students forgot to tag at some point. Of the 24 students, 17 males (8.2% of males), 4 females (6.3% of females) and 3 U.S. genders (11% of U.S.) did not tag at all. We transformed the commit message column to lowercase and found that students used 604 unique tags instead of the 5 tags we suggested. Some of the tags were as a result of typographical and spelling errors while others used synonyms of the 5 tags. We attempted to clean and categorize the commit tags into the 5 tags but ended up with 9 categories as described in Table 2. Using Table 2, we created a new column that identified whether or not a student had tagged their commit and what category (out of the 9) their original tag fell under. If the student hadn't tagged their commit, we tag the commit 'untagged', if their tag didn't fall into the categories in Table 2, we tagged the commit message as 'misc' which means miscellaneous.

Table 3. Project Description

Project Name	**Topic**	**Number of Repos**
Guided Project 1 (gp1)	Composition and SE processes and practices	296
Guided Project 2 (gp2)	Inheritance	286
Guided Project 3 (gp3)	Testing	281
Project 1 (p1)	Finite State Machines (Implementation and Testing)	280
Project 2 (p2)	Linear Data Structures (Implementation and Test)	134 Individual repos, 71 Team repos, 13 multi-gender repos
Labs	Labs were on the following topics: InstallFest and Pair Programming, Project Creation, V&V Best Practices, Collections, Design, Debugging, FSMs, ArrayLists, Linked Lists, Stacks and Queues, Iterators, and Recursion and Lists	501 Individual repos and 93 multi-gender repos

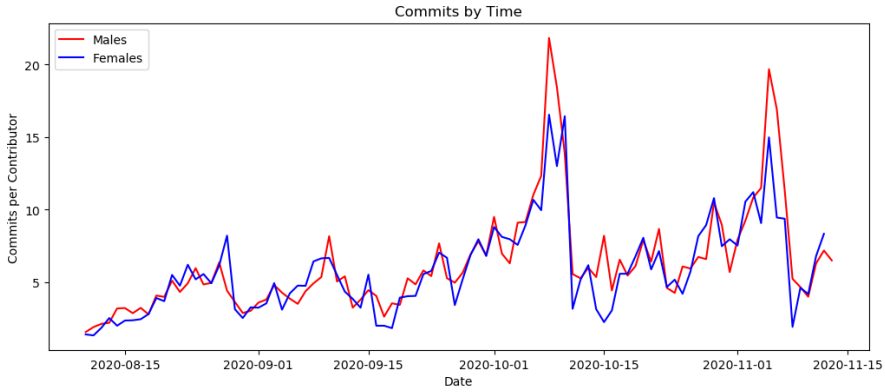
3.3 Repo Processing

Students worked on 5 unique projects and several lab sessions during the semester. We created a new column to indicate the project being worked on in each commit. From the column "Repo_name', we identified whether the commit was for Guided Projects 1 through 3, Project1, Project2, or Lab. These projects are described in Table 3.

3.3.1 Commit Timeline. 2 shows the average commits made per day by each gender. Interestingly, both genders had similar average commits per day. The earliest commit within the class was on August 10th. The last commit was on November 13th. The spike in Figure 2 shows the maximum number of commits. These were made on 10/08 (which was around the deadline for project 1). The second highest spike was on 11/05 (which was the deadline for project 2).The minimum number of commits was done on 9/17, 11/3 (around the start and end of the semester respectively).

3.3.2 Untagged Commit Messages. While preprocessing the students' commit messages, we noted that there were 43,557 untagged commits, which accounted for about half of all commits. The

Fig. 2. Mean commit messages per day for each gender



majority of these came from lab groups (18,553 commits) which also had the largest percentage of untagged commits (54%). The rest are as follows, GP1: 2,795 commits (50%), GP2: 1,123 commits (35%), GP3: 1,654 commits (39%), P1: 10,087 commits (50%), Team P2: 4,841 commits (45%), P2: 4,504 commits (47%), Labs: 18,553 (54%).

Males and females within the class tended to leave messages untagged at a roughly similar rate, which was about 50% of all commits made by each gender or an average of about 150 untagged commits. In Project P2 and some labs, students were allowed to work in teams. Within the Lab team setting, the proportion of untagged messages by male and female was 57% and 56% respectively. However, within the P2 team setting, males had a significantly larger proportion of untagged commits (73%) as compared to females (56%). To answer RQ1b, both female and male students responded in about the same way (51.2% of all commits by females and 50.6% of all commits by males were tagged) to the recommendation to tag their commit messages. Except within the P2 team setting where males responded to the recommendation at a much lower rate than females in the team.

3.4 RQ 2: In the class we studied, were tasks assigned to team members biased along gender lines?

3.4.1 Mixed Teams. In order to answer the second research question on whether the tasks assignments were biased along gender lines, we need to analyze multi-gender teams. In Project 2 students had the choice of working in teams and in a few Labs instructors required students to work in teams. We created a subset of the data which included commit messages for project 2 and labs by teams with both male and female students. Within the mixed teams, for project 2, there were 71 teams, 9 teams having students with U.S. gender, 40 all male teams, 9 all female teams and 13 2-member teams, each consisting of 13 males and 13 females. Labs had 93 mixed 2-4 member teams consisting of 111 males and 52 female students in different ratios. The average number of commits made by gender shows an interesting difference between mixed teams and the overall class. On average, male students made 18% more commits than females in Project 2 with mixed teams having equal numbers of male and female team members. In Labs with mixed teams having an unequal number of male and female members, females made 80% more commits on the average than males whereas there were fewer females (52) within Lab teams than males (111). Within the whole class, female students made 5% more commits than males on the average.

Table 4. Statistics of Total Commit Messages

	# of Students			# of Commit Messages		
Gender	Mixed Team Project 2	Mixed Team Labs	Whole class (including mixed teams)	Mixed Team Project 2	Mixed Team Labs	Whole class
Male	13	111	208	1,050	5,360	60,466
Female	13	52	64	807	452	19,551

Team P2 with an equal number of male and female members and the Labs with multi-gender are our *focus teams*. Males and females had a roughly similar involvement rate proportional to their makeup within all team commits. Multi-gender team P2 made 1586 commits while working on P2. Males made 55% of all commits, females committed 45% of all commits. Male and female added and modified lines of code in the ratio 55:45 and 54:46 respectively. In summary, males in team P2 made slightly more commits, additions and modifications than females. Interestingly, males and females within mixed groups had varying commit patterns as related to the day of the week. Both made the majority of their commits on Thursday (Males (38.9%) | Females (44.6%)) which is usually the day of the deadline for project 2 and Wednesday (Males (18.7%) | Females (14%)) which is a day before the deadline for Project 2. Females tended to prefer working on the weekend a bit more than males.

The Lab teams were of different sizes, and there were 14 teams consisting of 1 male and 1 female student, who made 1,179 commits. Males in these teams made 49% of all commits, females committed 51% of all commits. Male and female added and modified lines of code in the ratio 44:56 and 54:46 respectively. This shows that although females made more commits, they added lesser lines of code but modified more lines of code. It also shows that both genders in the 14 teams contributed almost equally to the project. 55 teams had 2 males and 1 female with 5703 commits. 1 teams had 3 males and 1 female with 123 commits and 23 teams had 1 male and 2 females with 2261 commits.

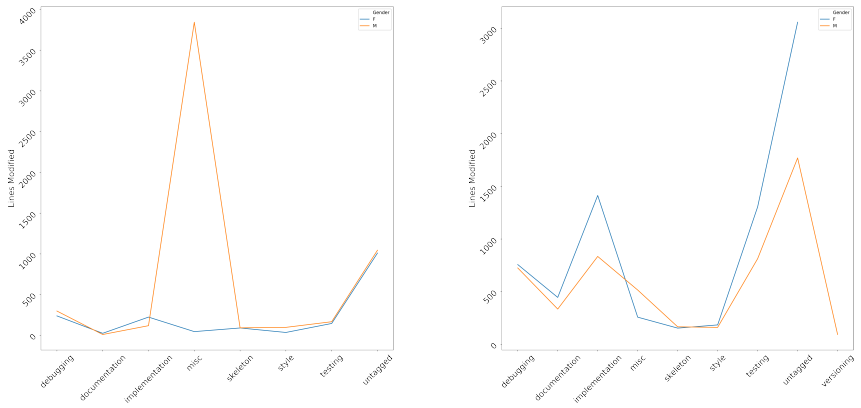
3.5 Visualizing Mean Number of Modified Files Committed in each category in Multi-Gender Teams

Each line chart in Figure 3 represents our focus teams and shows the mean of modified lines of code in each category as tagged originally by students. This contains untagged commit messages as well as misc tags. The red lines represent the male students and the blue represents the female students. The major inference from this figure is that due to the high number of untagged and misc messages, it is difficult to correctly conclude that one gender worked more on a certain category. In the first line chart, Figure 3(a) the female students in Team P2 performed slightly more implementation and documentation related tasks than the male student whereas the male students did more debugging, styling and testing. However, the majority of the commit messages by the male student had the miscellaneous tag. Similarly in Figure 3(b) the mean number of modified files by a female in the implementation category was higher than that of the male and since there were also a high number of untagged commits by both male and female, we cannot conclude on this result.

3.6 Tag Prediction using Natural Language Processing (NLP) techniques

By observing Figure 3, we cannot conclude what gender does more of one task simply because so many commit messages were not tagged. We used tf-idf NLP feature extraction techniques, three different classification techniques and an ensemble of the three techniques to predict the

Fig. 3. Each bin shows the mean number of modified files within each category as originally tagged by each gender. Focus Teams: (a) Team P2 and (b) Lab Teams with the original tags



(a) (b)

tags. The results are shown in Table 5. We applied One-vs-the-rest (OvR) multiclass strategy which divides a multi-class classification into one binary classification problem per class [1, 15]. We used logistic regression (LR) coupled with the One-vs-the-rest (OvR) multiclass strategy to classify the tags which resulted in an f-score of 0.631. We also used support vector classifiers (SVC) and naive bayes (NB) with OvR which resulted in f-score of 0.682 and 0.611 respectively. We tried to see if an ensemble of the three predictions would produce a better model but it resulted in an f-score 0.679 which was not better than the best model, SVM. Finally, we worked with the predicted tags from the model highest f-score which is OvR using SVC.

Table 5. Results of NLP Analysis

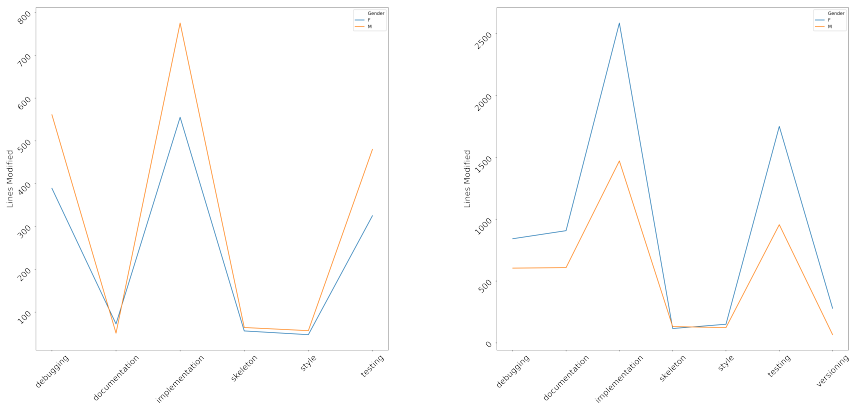
NLP Technique	**F-Score**	**Precision**	**Recall**
OvR using LR	0.631	0.631	0.631
OvR using SVM	**0.682**	**0.683**	**0.683**
OvR using NB	0.611	0.615	0.629
Ensemble of (LR, SVM, NB)	0.679	0.679	0.679

3.7 Visualizing Mean Number of Modified Files Committed in each NLP category in Multi-Gender Teams

Going by the NLP predicted tags, we can conclude based on Figure 4 that there was ample collaboration among genders. This means that each gender worked on a bit of each task and tasks assignment was not along gender lines in the projects we studied. Each line chart in Figure 4(a) represents the mean number of modified lines of code by each gender for the 13 P2 teams. We see that both genders worked the same tasks. On average, males modified slightly more lines of code than females in all tasks except documentation. Figure 4(b) represents the mean number of

modified lines of code by each gender for the 93 Lab teams and for each category as predicted by our NLP SVC model. Table 6 shows the percentage of commit message in each category by gender. Interestingly, although female students made slightly lower commits in most categories as shown in Table 6, on the average females modified more lines of code than male team members in all but the skeleton category while working on in Lab Teams. The lower average of modified lines under the skeleton category might be because more male teammates committed the starter project. Note that versioning tag was removed because some of the commit messages which include the keyword 'merge' were auto-generated by the platform. From the exploratory analysis, we can see that each gender worked on programming related tasks. To answer RQ2, by observing the line charts in Figure 4, female students collaborated with male students on each task.

Fig. 4. Each bin shows the mean number of modified files by each gender within each NLP classified tag Focus Teams: (a) Team P2 and (b) Lab Teams with the NLP-SVC tags



(a) (b)

Table 6. Percentage of commits messages in each category by gender O* - represents data from original tags. SVC - represents data from SVC - NLP tags.

	Multi-Gender P2 Teams				**Multi-Gender Lab Teams**			
Category	Male % O* tags	Female % O* tags	Male % SVC tags	Female % SVC tag	Male % O* tags	Female % O* tags	Male % SVC tags	Female % SVC tags
debugging	40	60	59	41	62	38	**56**	44
documentation	36	**64**	43	**57**	**53**	47	46	**54**
implementation	25	**75**	**54**	46	48	**52**	**54**	46
misc	**89**	11	-	-	**78**	22	-	-
skeleton	45	**55**	**52**	48	51	**49**	**58**	42
style	**56**	44	**52**	48	**53**	47	**54**	46
testing	48	**52**	**56**	44	50	50	**54**	46
untagged	**63**	37	-	-	**55**	45	-	-

3.7.1 Percentage of commits messages in each Category by gender. Furthermore, we analyzed the percentage of commit messages by each gender. Table 6, shows that each gender in Team P2 contributed almost equally to the same types of tasks. [9] defined collaboration as a working scenario where both members contributed between 30%-70% to two or more common parts of the project [9]. Going by the **NLP-SVC provided tags**, the male members did slightly more of debugging (59%), implementation(54%), skeleton(52%), styling(52%) and testing (52%) and females did slightly more of documentation (57%) in mixed P2 Teams with equal number of male and female members. Females performed above 40% of each task, hence there was a healthy collaboration between multi-gender teams with equal numbers of male and female students in the class we studied. Similar results were observed in Multi-Gender Lab Teams columns shown on Table 6. To answer RQ2(a) and (b), based on the definition of collaboration in [9], females did not significantly do less programming tasks than males. One gender did not significantly do more task than the other. Both members contributed over 40% to all tasks.

3.8 RQ3: Can we predict the gender (M/F) of the authors using details of the commit messages as predictor variables and can gender predict the commit tag?

3.8.1 Predicting Gender. To answer RQ 3, we classify the gender (M/F) of a contributor on a project with mixed gender teams by analyzing the data of multi-gender Teams in P2 and Labs. We wanted to see what features are important in predicting gender contribution within a mixed team. Specifically, we want to see if commit tag (e.g. testing, implementation etc.) is an important predictor of gender. For example, to what extent can the commit tag tell the commit author's gender? Before we began training our models, we performed data pre-processing to transform some of the data to a format that the algorithms could use. To this end, we used LabelEncoder() in Python to binarize both the Original and SVC commit tag values and gender values. We then split our data into 80% training and 20% testing for each of the models we used. We further split our training set into a training and validation set in order to optimize our hyperparameters. After achieving a satisfactory set of hyperparameters, we performed a final test on the testing set. When we created our models, we also included a 'class_weight = balanced' parameter to ensure that there were an equal number of male/female commits that would be used to create each classifier. We applied Random Forest (RF), Decision tree (DT), Logistic Regression (LR), Naive Bayes (NB), KNN, Radius, and SVM and compared their performance in predicting gender.

3.8.2 Predicting Gender using only Commit Tags. We trained 7 models to predict gender using only the original commit tags. We included only commit messages tagged by students in this process. The results of this analysis is shown in column A of Table 7 where we show the balanced accuracy (BA) and f-score (F-S) of each of the 7 models. We trained 7 more model to predict gender using only the predicted tags from the NLP - SVC process and reported our results in column B of Table 7.

3.8.3 Predicting Gender using more features. We trained additional 7 models using 6 features namely 'added', 'deleted', 'modified', 'removed', 'original commit tag' and 'dayOfWeek' to predict the gender of the contributor behind a given commit. Lastly, we trained an additional 7 models using the 6 features but used the NLP-SVC classified commit tag instead of 'original commit tag'. We trained our model on Team P2 and Labs which has an equal number of male and female students in the team. We also removed the commit messages with the tag 'skeleton' because it contains the initial starter code provided by the instructor and not the code produced by the students.

We explored random forests and decision tree classifiers due their simple yet powerful techniques. For RF, we used 100 trees which produce the optimal validation accuracy. For DT, we used the criteria of entropy so that branches would be built only if they reduced the entropy of a classification. We then used Logistic Regression with min-max normalization and a C-value of .3 as well as Gaussian

Naive Bayes. For the KNN and Radius classifiers, we assigned weights to the distance, used a power parameter of 3 to signify minkowski distance and tested k-values from 1 - 30. K=1 in columns C and D of Figure 6 provided the best results. We also tried the Radius neighbors classifier with R = 4 in column C and R =1 in D. Finally, we used SVM (LinearSVC) with a C-value of .005. The performances of these models are shown in Table 6. We determined the accuracy of a model by it's balanced accuracy (BA), which is the average recall obtained on both males and females. RF and KNN were our best classifiers for predicting gender using original tags with RF's accuracy of 83% and KNN's F-Score of 77%. Whereas RF and radius were best for predicting gender using NLP-SVC tags. Table 9 shows the feature importance of each feature in predicting gender. Each model gives the feature importance of each independent variable. We present feature importance for two top performing models RF and KNN. The lines of code added is the most important feature in predicting gender. The next most important is the lines of code modified. Neither the original commit tags nor the NLP classified tags were highly important and as we observed in Table 6 columns A and B, these features alone are not good predictors of gender.

3.8.4 Predicting Commit Tags using Gender as a feature. Finally, we trained models using gender to predict the commit tag. We trained RF and KNN models which were high performing models from Table 7 but we got a very low performance as shown in Table 8, signifying that gender is not an important feature in predicting commit type.

Table 7. Performance of Models in Predicting Gender O* - represents prediction using original tags as a feature. SVC - represents prediction using SVC - NLP tags as a feature. Number in bold signify highest performance in each column.

	A (1 Feature - O* Tags)		B (1 Feature - SVC Tags)		C (6 Features - O* Tags)		D (6 Features - SVC Tags)	
Models	BA(%)	F-S(%)	BA (%)	F-S(%)	BA(%)	F-S(%)	BA (%)	F-S(%)
Majority Class	50	50	50	**71**	50	50	50	71
Random Forest	50	3	50	68	**83**	76	**76**	**77**
Decision Tree	56	42	52	54	82	75	73	75
Logistic Regression	50	0	50	0	55	56	50	7.6
Naive Bayes	50	0	50	70	60	46	51	42
KNN	50	30(K=1)	**53**	62	82	**77(K=1)**	72	73(K=1)
Radius	50	0(R=1)	51	68	78	73(R=4)	74	**79(R=1)**
SVM	**60**	**50**	52	62	63	58	55	47

4 DISCUSSION

Table 7 reports the performances of our 28 models in predicting gender. The column A in Table 7, is the result of using one feature which is Original 9 Tags to predict gender. We excluded commit

Table 8. Accuracy of Models in predicting Commit Tags using Gender as a feature Highest values are in bold fonts

Models	**Original Commit Tags**	**NLP-SVC Tags**
Majority	**34**	**41**
Random Forest (Classify Commit)	23	20
KNN	24 (K=1)	20(K=4)

Table 9. Feature Importance of 6 features across top two performing models from column C and D - using permutation importance for each predictive analysis O* - represents prediction using original tags as a feature. SVC - represents prediction using SVC - NLP tags as a feature. Highest values are in bold fonts, commit tag features are in italics.

Model /Features	**Added (O*/SVC)**	**Deleted (O*/SVC)**	**Modified (O*/SVC)**	**Removed (O*/SVC)**	**Original Tags**	**Encoded NLP-SVC Tags**	**Day (O*/SVC)**
Random Forest	**0.33/0.32**	0.06/0.07	**0.30/0.31**	0.01/0.02	*0.12*	*0.13*	0.17/0.15
KNN	**0.38/0.32**	0.07/0.07	**0.38/0.36**	0/0.01	*0.15*	*0.11*	0.11/0.13

messages with the misc tags and untagged messages. In column B, we applied NLP classified tags to predict gender. We observe that these results were poor and barely performed better than the majority classifier. In column C, 6 features including original clean tags were used. Whereas in column D we used the 6 features but replaced original tags with NLP classified tags. Table 8 shows the result of predicting tags using only the gender as a variable. The low performance in columns A, B, and Table 8 show that the commit tag doesn't accurately predict gender and vice versa. Columns C and D show that the performance of gender prediction improved when other 5 features were included. To answer RQ3, we cannot predict gender based on commit tag alone and vice versa which means in the class we studied the type of task worked on was not along gender lines. In the second two columns we used the 6 features to predict gender. Table 9 shows the feature importance of the 6 features in predicting gender. Elements of a commit such as the lines added and modified had higher feature importance than the commit tags. This shows that each gender worked on similar tasks therefore the tags alone could not be used to differentiate the genders.

5 LIMITATIONS

Other factors such as competence of each student programmer irrespective of gender might affect the choice of tasks undertaken by each gender. If teams were requested by students instead of assigned by instructor, this might affect the distribution of tasks as team members in requested teams might already have a plan on how to work together.

6 CONCLUSION AND FUTURE WORK

We explored the team work habits by each gender in a Java programming class. In the class we studied, students were advised to tag their commit messages to indicate what type of task they worked on. Each gender responded about the same way to the recommendation to tag commit messages. In the multi-gender teams we studied, we did not find any team with negative teamwork habits such as cooperation where females solely perform less programming intensive tasks or female free-rider habit where the male team member does the majority of the tasks while the female does little or nothing. We found that both genders collaborate to complete the project. Future work will replicate this work for multiple semesters and years to confirm the validity of our results across a larger population. This study is beneficial for identifying harmful teamwork habits among multi-gender teams. An adaptive support platform that provides instructors with similar insights from our analysis in real-time could signal the need for interventions to promote better teamwork. The knowledge that both genders have important roles to play in the success of a programming project is encouraging to females who aspire to take up a career in computing.

REFERENCES

[1] 2021. https://en.wikipedia.org/wiki/Multiclass_classification

[2] Sylvia Beyer, Kristina Rynes, Julie Perrault, Kelly Hay, and Susan Haller. 2003. Gender Differences in Computer Science Students. 35, 1 (2003). https://doi.org/10.1145/792548.611930

[3] Kevin Buffardi. 2020. Assessing Individual Contributions to Software Engineering Projects with Git Logs and User Stories. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Portland, OR, USA) *(SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 650–656. https://doi.org/10.1145/3328778.3366948

[4] J. Cohoon and W. Aspray. 2008. *Women in Computer Science or Management Information Systems Courses: A Comparative Analysis*. 323–349.

[5] Irina D. Coman, Pierre N. Robillard, Alberto Sillitti, and Giancarlo Succi. 2014. Cooperation, collaboration and pair-programming: Field studies on backup behavior. *Journal of Systems and Software* 91 (2014), 124–134. https://doi.org/10.1016/j.jss.2013.12.037

[6] Sherae Daniel, Ritu Agarwal, and Katherine J. Stewart. 2013. The Effects of Diversity in Global, Distributed Collectives: A Study of Open Source Project Success. *Information Systems Research* 24, 2 (2013), 312–333. http://www.jstor.org/stable/42004307

[7] Jie Du and H. Wimmer. 2019. Hour of Code: A Study of Gender Differences in Computing. *Information Systems Education Journal* 17 (2019), 91–100.

[8] L. V. D. Duim, J. Andersson, and M. Sinnema. 2007. Good Practices for Educational Software Engineering Projects. *29th International Conference on Software Engineering (ICSE'07)* (2007), 698–707.

[9] Deleted for Blind review. [n.d.].

[10] N. Imtiaz, J. Middleton, J. Chakraborty, N. Robson, G. Bai, and E. Murphy-Hill. 2019. Investigating the Effects of Gender Bias on GitHub. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. 700–711. https://doi.org/10.1109/ICSE.2019.00079

[11] S. Kamberi. 2017. Enticing women to computer science with es (Expose, engage, encourage, empower). In *2017 IEEE Women in Engineering (WIE) Forum USA East*. 1–5. https://doi.org/10.1109/WIE.2017.8285609

[12] A. Kermarrec. 2014. Computer Science: Too Young to Fall into the Gender Gap. *IEEE Internet Computing* 18, 3 (2014), 4–6. https://doi.org/10.1109/MIC.2014.48

[13] J. E. Mills, M. E. Ayre, and J. Gill. 2008. Perceptions and understanding of gender inclusive curriculum in engineering education. In *SEFI Annual Conference*.

[14] Leo Porter, Daniel Zingaro, Cynthia Lee, Cynthia Taylor, Kevin C. Webb, and Michael Clancy. 2018. Developing Course-Level Learning Goals for Basic Data Structures in CS2. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 858–863. https://doi.org/10.1145/3159450.3159457

[15] scikit. [n.d.]. sklearn.multiclass.OneVsRestClassifier¶. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

[16] Bernadette Spieler, Libora Oates-Indruchovà, and Wolfgang Slany. 2020. FEMALE STUDENTS IN COMPUTER SCIENCE EDUCATION: UNDERSTANDING STEREOTYPES, NEGATIVE IMPACTS, AND POSITIVE MOTIVATION. *Journal of Women and Minorities in Science and Engineering* 26, 5 (2020), 473–510. https://doi.org/10.1615/jwomenminorscieneng.2020028567

[17] Beyer Sylvia and Michelle DeKeuster. 2008. Women in Computer Science or Management Information Systems Courses: A Comparative Analysis. *Women and Information Technology* (2008). https://doi.org/10.7551/mitpress/7272.003.0013

[18] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and Tenure Diversity in GitHub Teams. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2702123.2702549

[19] Alexey Zagalsky, Joseph Feliciano, Margaret-Anne Storey, Yiyun Zhao, and Weiliang Wang. 2015. The Emergence of GitHub as a Collaborative Platform for Education. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1906–1917. https://doi.org/10.1145/2675133.2675284