2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
Paper ID #39986

# An analysis of relationships between course descriptions and student enrollment patterns

**Dr. Agoritsa Polyzou, Florida International University**

Agoritsa Polyzou is an Assistant Professor at the Knight Foundation School of Computing and Information Sciences in Florida International University (FIU), Miami. Agoritsa received the bachelor's degree in computer engineering and informatics from the University of Patras, Greece, and her Ph.D. degree in computer science and engineering from the University of Minnesota. Next, she was a Postdoctoral Fritz Family Fellow with the Massive Data Institute of McCourt School of Public Policy at Georgetown University, Washington, DC. She is involved in projects in the intersection of education, data mining, machine learning, ethics, and fairness. Her research interests include data mining, recommender systems, predictive models within educational contexts, and the fairness concerns that arise from their use. Her goal is to help students succeed using data and machine learning models.

**Joaquin Molto, Florida International University**

Joaquin Molto is a Florida International University student who has earned his B.S. in Computer Science with a Minor in Mathematical Sciences. He is currently pursuing his M.S. in Computer Science and is passionate about Software Engineering, AI, and Machine Learning. Throughout his academic career, Joaquin has demonstrated a keen aptitude for programming, developing his skills in numerous programming languages, including Python, Java, C++, and C. He has also gained practical experience working on various software engineering projects, including designing and implementing efficient algorithms, creating user-friendly interfaces, and optimizing application performance. Joaquin is particularly interested in the applications of AI and machine learning to solve complex problems, and he has already started exploring these areas through his coursework, personal projects, and research.

**Nicholas Sean Gonzalez, Florida International University**

Graduate AI/ML Researcher at Florida International University

**Dr. Trina L. Fletcher, Florida International University**

Dr. Fletcher is currently an Assistant Professor at Florida International University. Her research focus equity and inclusion within STEM education, STEM at HBCUs and K-12 STEM education. Prior to FIU, Dr. Fletcher served as the Director of Pre-college Pr

**Sophia Tavio Perez**

# An analysis of relationships between course descriptions and student enrollment patterns in engineering and computing undergraduate education


## Abstract

Over the last years, there have been great efforts to improve the diversity of engineering and computing education. Our work contributes to these efforts by following a data-driven approach to analyze easily accessible data and provide insights that could potentially lead to meaningful and impactful interventions. In particular, we focus on course selection and how we can balance the percentage of women in courses across engineering and computing courses. Deciding which courses to take every semester can be challenging for students. One of the factors influencing their decisions is the descriptions of the available courses. By reading these, the students get a first impression of the type and content of a course. This study reviews course descriptions offered by the college of engineering and computing. We employ natural language processing (NLP) approaches to identify patterns in the language used in course descriptions and how this relates to the student enrollment and descriptive characteristics of the different departments and courses. Our ultimate goal is to identify and quantify how different course descriptions are from different majors as these relate to the student gender distribution. Our language analysis indicates that verbs, adjectives, and adverbs have the most significant impact on differentiating course descriptions and highlighting differences across the different programs and across the different variables of focus. Implications of this work and the impactful dissemination include sharing results with faculty and staff within the college during departmental and college-wide meetings to encourage meaningful course description changes for their courses. This research adds significantly to the literature as there is very little research on the impact of course descriptions on students' course selection process.


## 1. Introduction

Efforts to broaden the participation of women and people of color within engineering and computing education have made incremental improvements in the U.S. but, overall, continue to have challenges. The under-representation of women and ethno-racial minorities in STEM fields is nothing new [1]. Based on a 2021 report, the share of women in computer occupations declined from 30% in 2000 to 25% in 2016 and has remained stable until 2019 [2]. Women continue to be vastly underrepresented in the ranks of engineers and architects (15% in 2019), but their share has increased since 2000 (12%). Research has shown that diversifying faculty, increasing financial support for these populations, and improving culture and climate within departments all can help contribute towards improvements in the efforts [3]-[5]. However, each of these recommendations can take years to be implemented and oftentimes require adequate resources and individuals in positions of power (i.e., department, school, or college-level leadership) to advocate, push for, and prioritize those recommendations.

Based on this, our team sought to find ways to use available data and machine learning (ML) methodologies as a pathway for positive classroom-level changes can contribute towards broadening participation efforts within engineering and computing undergraduate programs. The first student contact with a course is through its course description which is publicly available in the department's undergraduate course catalog. While there are other factors influencing student decision making that are out of our control, there is very little research on the impact of course descriptions on students' course selection process. That first impression may have a considerable effect on student enrollment. We believe that the language used in the course descriptions may have an impact on how female students perceive computing and engineering courses.

We review and analyze course descriptions offered by the college of engineering and computing at one of the nation's largest universities. By employing NLP approaches, we identify patterns in the language used in course descriptions across programs with varying female student enrollment. We examine the programs of biomedical engineering, electrical engineering, computer science, and civil and environmental engineering. Our goal is to identify how different course descriptions are from these different majors in relation to the student gender distribution. We identify that the use of verbs, adjectives, and adverbs varies a lot depending on the department. Departments with higher percentages of male student enrollment in courses had significantly less, as they were using more nouns.

In course descriptions, we see an opportunity to improve attracting, bridging, and retaining women in engineering and computing. The outcomes of this work highlight the importance of using language that does not create additional barriers for women to consider courses that are traditionally considered to be male dominated. Faculty and staff can consider the insights from our work to adjust existing course descriptions or to create descriptions for new offerings.

## 2. Background and motivation

*Course selection*

Course selection in higher education has been studied in a fragmented way. Different works examine specific criteria. DellaGioia investigated how other students' opinions could impact the likelihood of enrolling in or recommending a course after students had read the courses description [6]. The lecturer's style, the learning value and difficulty of the courses, as well as other course section attributes, such as meeting time have also been considered [7], [8]. Babad assumes that course selection is a sequential process, as students first select one course, usually because of its intellectual level, expected quality of teaching, and students' potential learning and occupational gains, while the last course is selected because it is convenient and easy [9]. On a slightly different setting, Gaskell looked into course selection and gender in high school [10].

Overall, there has been little research on the impact of course descriptions on students' decision making. The exception is a work by Mourey et al., where they examine the conceived difficulty on the presentation of course information [11]. The idea is that the same course can be described in different ways that might sound easy or complex to the students. Mourey et al. study how the course description difficulty affects student enrollment. That also motivates our work. We believe that we can re-write course description in a way that is welcoming to all students, without propagating existing biases. This paper is the first step towards identifying differences in course descriptions with respect to student enrollment that can drive our future interventions.

Course descriptions have certain limitations (limited number of words, only plain text, often unclear expected entry knowledge). Alone, they might not be adequate to describe a course or inform student selection [12]. For that reason, the problem of information asymmetry exists, meaning that the available information is either insufficient or not fully used when students select courses [13]. However, course catalogs are published to disseminate relevant information about the courses to all students. It is a resource that is readily available and as such, any minor intervention may have a significant effect, especially to students from underrepresented or minority groups, who usually struggle more to get useful information and advice to support their decision making.

*Course descriptions usage*

Course descriptions offered by a university course catalogue have been used to answer different research questions. Course descriptions have been used to depict the current state of data science education in the U.S. [14], [15]. In the educational data mining area, course recommendation systems have been built that use course descriptions to extract information about concepts or topics covered by each course [16] - [20]. Another example is the problem of next-term grade prediction [21].

Course syllabi (a more detailed text-based description of a course) have been investigated to understand how they can form the students' perceptions of the instructor and the class [22]. Other researchers work towards understanding and improving the communication value of course syllabi [23]. Course syllabi include more information about the instructors and the expectations, requirements, learning goals, assignments, and policies of courses. As our focus is on better informing course selection, course syllabi are not ideal sources of data as usually students have access to them after they register for courses.

In this space, we identify a gap in the prior work, as course descriptions have not been studied yet on how they influence student decision making (apart from the recent paper from Mourey et al., 2022). Even further, there is no existing work connecting course descriptions with student enrollment with respect to students' gender or other protected and sensitive characteristics.

*Prior text-based work in the educational domain*

The educational domain has rich textual information. Recently, researchers have been using text processing techniques to solve research questions in an educational setting. The system SMART was proposed which can identify latent skills from existing instructional text on existing online courseware [24]. Text mining techniques have been used to connect job descriptions and course descriptions and curricula descriptions [25]. In another application, course descriptions are compared against the resume of a faculty in order to assign adjuncts to courses [26]. Gomez et al.

analyzed the text of learners' reviews of courses in Massive Open Online Course (MOOC) platforms to better support course selection in MOOCs [27].

*Gender Differences in writing (text data)*

Numerous empirical studies have been conducted examining text-level linguistic features related specifically to men and women and the effects of gender on linguistic behavior [28] - [32]. These works examine various characteristics, both at the word-level as well as on the style of the writing. However, they are not directly applicable in our setting, as the nature of the text of course descriptions is quite unique and different from essays, responses to assignments, or spoken communications. Additionally, we do not have information about who exactly wrote the course descriptions and their gender. The gender of the instructor could be an indicator, but course descriptions do not change based on who teaches the course.

## 3. Methodology/Methods

The research question that we will investigate is: are there linguistic differences in the vocabulary of course descriptions between departments with different gender distributions?

*Data*

We are using data from two different sources for our analysis. First, we collected all course descriptions from the undergraduate course catalogs of four different departments during the 2020-2021 academic year in the College of Engineering and Computing at Florida International University, a big, public, minority-serving institution. More specifically, we used the Department of Biomedical Engineering (BME), Civil and Environmental Engineering (CEE), Computing and Information Sciences (CIS), and Electrical and Computer Engineering (ECE). Second, we also used course enrollment statistics by gender for the same academic year. BME is known to have more women enrolled and graduating from those programs. On the contrary, CEE, CIS and ECE are heavily male dominated, particularly the ECE department. We will collectively signify the heavier presence of females as most females (MF), and the thinner presence of females in the last three departments as less females (LF). You can find more statistical information for each department in Table 1.

**Table 1. Department statistics**

| Department | Male students | Female students | Total Students |
|---|---|---|---|
| BME | 49% | 51% | 350 |
| CEE | 75% | 25% | 1K |
| CIS | 81% | 19% | 2.6K |
| ECE | 88% | 12% | 1.2K |

From all the collected course descriptions, we removed those that refer to special courses, e.g., “Research Experience for Undergrads”, “Graduate Research”, “Project Research”, “Capstone”, “Cooperative Education in Computing”, “Special Topics”, “Independent Study”, “Vertically Integrated Projects”. If a course has multiple sections, we aggregate them into one, and consider them as a single offering, since the course description will be the same for all sections. We also removed courses that had less than five students enrolled, as in that case, the percentage of male versus female would be less meaningful and could have an unintended effect when aggregating them with other courses with higher enrollment numbers. In the departments of BME, CEE, CIS, ECE, we have 31, 62, 91, and 102 course remaining, respectively. In our text analysis, we did not consider any additional information related to prerequisites, co-requisites, or number of credits. We also removed special characters and acronyms.

*Text Analysis*

We used Python to open and read the course description files, and we filter out the necessary courses and special characters. We transform our text so that all of it is lowercase. The next step is to use the python package spaCy and its trained NLP pipelines to perform our analysis [33]. More specifically, we identify and remove stop words, which are words like “and”, “or”, etc. that do not hold any special meaning. Then we perform part-of-speech (POS) tagging, which will return the type of word. While spaCy returns many different tags, we will focus our analysis on the most important ones, i.e., nouns, verbs, adjectives, adverbs. These are the most commonly used types of words used in our dataset.

## 4. Results

First, we did not identify any variation on the number of words per sentence or the number of words per course description. On average, all departments had around 7 words per sentence and 34 words per course description. This was expected, as there are certain standards within the universities that the course descriptions need to follow.

*POS tags*

For each department, Table 2 shows the average percentages over all the courses remaining in our dataset. We grouped together adverbs and adjectives, as both describe or modify other parts of speech. We see a clear distinction between the use of nouns, adverbs and adjectives across departments. More specifically, the BME department uses less nouns and more verbs or adverbs and adjectives than the LF departments. The high number of nouns might point to a “dry” description, including a lot of words covering the different topics. This is more prominent of departments with higher number of male students. So, interestingly, our insights regarding the use of nouns follow the same trend as the male-enrollment in the corresponding departments. ECE that has the highest percentage of male students has the highest number of nouns, adverbs and adjectives.

**Table 2. The average percentage (± standard deviation) of nouns, adverbs and adjectives, verbs, and omitted special characters and acronyms for each department in our dataset.**

| Department | Nouns | Adv and Adj | Verbs | Omitted Characters & Acronyms |
|---|---|---|---|---|
| BME (MF) | 60.1 (±6.0) | 16.8 (±4.4) | 10.9 (±0.7) | 13.1 (±3.4) |
| CEE (LF) | 60.6 (±4.7) | 13.7 (±2.4) | 19.1 (±3.2) | 6.4 (±2.7) |
| CIS (LF) | 67.0 (±3.3) | 13.7 (±2.5) | 15.5 (±3.8) | 4.0 (±1.6) |
| ECE (LF) | 66.5 (±2.6) | 13.1 (±2.0) | 16.2 (±2.4) | 4.1 (±1.9) |

While the CEE department seems to have a low number of nouns, that is not true. After careful inspection of the POS tagging results, we noticed that many words in that particular department were misclassified as verbs, while they were abbreviations or nouns, for example, “filter”, “cod”, “toc”, “env”, “chm”. As a result, the number of verbs is not a very trustworthy quantity to consider. On the other hand, the adverbs and adjectives seem to have way less false positives (tokens that were misclassified as adverbs and adjectives while they were actually something else), so, these numbers are more trustworthy.

*Statistical significance*

In order to verify that our findings are meaningful, we used a t-Test. We want to determine whether the differences between two samples are significant, or if they are likely to have come from the same two underlying populations that have the same average. In our case, the samples are the percentages of each POS tag in courses descriptions of different departments (BME, CEE, CIS, ECE). We want to verify that the differences we observe in the average percentage of nouns, and adverbs and adjectives are significant. We form a null hypothesis; the average percentage of a POS tag is similar across a pair of departments. We conduct a t-test to test if the hypothesis is plausible. If we can reject that hypothesis, that means that the mean of the two population sets have intrinsic differences which are not by chance. In particular, we use the unequal variance t-Test since the number of courses across departments is different, and the variances we observe are also different.

We performed the t-Test for the percentage of nouns and adverbs and adjectives for each pair of departments. We set the significance level to 5% (i.e., confidence level to 95%). For the average number of nouns, we were able to reject the null hypothesis when we compared the BME department with CIS or ECE. The differences were less significant compared to the CEE department. For the adverbs and adjectives, we were able to able to reject the null hypothesis when comparing BME with any other department, indicating that our findings are significant with 95% confidence. The CIS and ECE departments were the pair of departments that were the most similar, as we could not reject the null hypothesis for any POS tags.

*Most common words*

Table 3 presents in more detail the most common words that are unique in each department. We can see that these words are specific to the corresponding area of the department. Even in that list, we see that BME has 6 adjectives, respectively, while CEE, CIS and ECE have 4, 3, and 4 adjectives, respectively.

We also examined the most common words per department that were also shared by other departments. These were very general words, like “course”, “system”, “design”, “analysis”, and “application”. BME and CEE each have one adjective in that list, which CIS and ECE have only nouns. This outcome (that nouns were the words most commonly shared across departments) was expected, as those do not capture domain-specific concepts but are relevant for all departments.

**Table 3. Top 10 most common unique words per department, and their type.**

| BME | CEE | CIS | ECE |
|---|---|---|---|
| Biomedical - Adj | Civil - Adj | Additional - Adj | Embedded - Adj |
| Tissue - Noun | Transportation - Noun | Fee - Noun | Electronic - Adj |
| Cardiovascular - Adj | Structural - Adj | Database - Noun | Electronics - Noun |
| Clinical - Adj | Highway - Noun | Acceptable - Adj | Forensics - Noun |
| Physiology - Noun | Air - Noun | Credit- Noun | Radar - Noun |
| Organ - Noun | Wastewater - Noun | Unix - Noun | Typical - Adj |
| Physiological - Adj | Waste - Noun | Parallel - Adj | Transmission - Noun |
| Optic - Adj | Hazardous - Adj | Server - Noun | Autonomous - Adj |
| Optical - Adj | Concrete - Noun | Intelligence - Noun | Analog - Noun |
| Biomechanics - Noun | Pollution - Noun | Window - Noun | Order - Noun |

## 5. Limitations

While in this work we only used four different departments, we plan to expand our analysis to other departments from other colleges, where we might find even more prominent differences in student enrollment based on gender. The outcomes of this study will inform and guide future research that is needed within departments.

Our filtering might have disproportionally affected the courses remaining in the BME departments, which in general have fewer enrolled students. There are less courses in that department that survived the pruning.

Another important consideration that we do not cover in our work is the instruction mode. We used data from the 2021-2022 academic year, where the students and the whole educational system was still affected by the COVID-19 pandemic. Many courses were offered online, some were hybrid, while others were taught in-person. We believe that this is an important factor that might have influenced course selection, and consecutively, our findings.

## 6. Conclusion & Future Work

Course selection is a difficult process for undergraduate students. One of the most easily accessible resources that can guide them is the course catalog, where students can find the course descriptions and have a first look at the topics covered. At the same time, historical and preexisting biases shape course selection, which leads to unbalanced student enrollment in terms of gender across departments. Our work is the first that explores the characteristics of course descriptions as those relate to male/female student enrollment. After data collection and careful data analysis, we identify some characteristics that are similar across departments (e.g., number of words per sentence or course description), while we find that the use of nouns and adjectives/adverbs is closely related to the percentage of male versus female enrolled students. This is our first step in that direction. In our future work, we want to expand our analysis to other departments, as well as to other linguistic characteristics. We also plan on examining differences within each department, to identify differences in course characteristics or topics that have unbalanced student enrollment.

## References

[1] T. Ross, G. Kena, A. Rathbun, A. KewalRamani, J. Zhang, P. Kristapovich, and E. Manning. “Higher Education: Gaps in Access and Persistence Study (NCES 2012-046)”. U.S. Department of Education, National Center for Education Statistics. Washington, DC: Government Printing Office, 2012.

[2] R. Fry, Kennedy, B. and C. Funk, “STEM jobs see uneven progress in increasing gender, racial and ethnic diversity”. Pew Research Center, 2021, pp.1-28.

[3] S. James, S. Singer. “From the NSF: The National Science Foundation's Investments in Broadening Participation in Science, Technology, Engineering, and Mathematics Education through Research and Capacity Building”. CBE Life Sci Educ. 2016 Fall; 15(3):fe7, doi: 10.1187/cbe.16-01-0059.

[4] J. Peckham, L.L. Harlow, D.A. Stuart, B. Silver, H. Mederer, and P.D. Stephenson, “Broadening participation in computing: issues and challenges”. *ACM SIGCSE Bulletin*, 39(3), , 2007, pp. 9-13.

[5] T.K. Holloman, W.C. Lee, J.S. London, A. Hawkins, and B.A. Watford, “The assessment cycle: Insights from a systematic literature review on broadening participation in engineering and computer science”. *Journal of Engineering Education*, 110(4), pp.1027-1048, 2021.

[6] M. DellaGioia. “Student Opinion and Student Course Selection”. *Journal of Undergraduate Psychological Research*, 3, 2008.

[7] E. Babad, and A. Tayeb, “Experimental analysis of students' course selection”. *British Journal of Educational Psychology*, 73(3), pp.373-393, 2003.

[8] C.L. Brown, and S.M. Kosovich, “The impact of professor reputation and section attributes on student course selection”. Research in Higher Education, 56, pp.496-509, 2015.

[9] E. Babad. “Students' course selection: Differential considerations for first and last course,” *Research in Higher Education*, pp. 469-492, 2001.

[10] J. Gaskell, “Gender and course choice: The orientation of male and female students,” *Journal of Education*, 166(1), pp. 89-102, 1984.

[11] J.A. Mourey, M.M. Markley, and S.K. Koernig, “Dazzling descriptions and tantalizing titles: How simple versus complex course information influences course selection,” *Journal of Marketing Education*, 44(1), pp. 100-112, 2022.

[12] O. Simpson, “Access, retention and course choice in online, open and distance learning”. *European Journal of Open, Distance and E-learning*, 7(1), 2004.

[13] M. Scott, and D.A. Savage, “Lemons in the university: asymmetric information, academic shopping and subject selection”. *Higher Education Research & Development*, 41(4), pp. 1247-1261, 2022.

[14] D. Bukhari, “Data science curriculum: Current scenario”. *International Journal of Data Mining & Knowledge Management Process*, Vol. 10, 2020.

[15] D. Li, E. Milonas, and Q. Zhang, “Content Analysis of Data Science Graduate Programs in the US,” *2021 ASEE Virtual Annual Conference Content Access*, 2021.

[16] Z. Chen, X. Liu, and L. Shang, “Improved course recommendation algorithm based on collaborative filtering,” in *International Conference on Big Data and Informatization Education*, pp. 466-469. IEEE, 2020.

[17] H. Ma, X. Wang, J. Hou, and Y. Lu, “Course recommendation based on semantic similarity analysis,” in *3rd IEEE International Conference on Control Science and Systems Engineering*, pp. 638-641. IEEE, 2017.

[18] Y. Ng, and J. Linn, “CrsRecs: A personalized course recommendation system for college students,” in *8th International Conference on Information, Intelligence, Systems & Applications*, pp. 1-6. IEEE, 2017.

[19] Z. Pardos, and W. Jiang, “Designing for serendipity in a university course recommendation system,” in *Proceedings of the tenth international conference on learning analytics & knowledge,* pp. 350-359, 2020.

[20] R. Morsomme, abd S.V. Alferez, “Content-based Course Recommender System for Liberal Arts Education,” *Educational Data Mining*, 2019.

[21] S. Morsy, and G. Karypis, "Cumulative Knowledge-based Regression Models for Next-term Grade Prediction," In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 552-560. Society for Industrial and Applied Mathematics, 2017.

[22] R.J. Harnish, and K.R. Bridges, "Effect of syllabus tone: Students' perceptions of instructor and course", *Social Psychology of Education*, 14, pp. 319-330, 2011.

[23] M.F. Smith, and N.Y. Razzouk, "Improving classroom communication: The case of the course syllabus," *Journal of Education for Business*, 68(4), pp.215-221, 1993.

[24] N. Matsuda, J. Wood, R. Shrivastava, M. Shimmei, and N. Bier, "Latent Skill Mining and Labeling from Courseware Content," *Journal of Educational Data Mining*, 14(2), 2022.

[25] A. Fortino, Q. Zhong, W.C. Huang, and R. Lowrance, "Application of Text Data Mining To STEM Curriculum Selection and Development," In *2019 IEEE Integrated STEM Education Conference*, pp. 354-361, IEEE, 2019.

[26] A. Fortino, Q. Zhong, L. Yeh, and S. Fang, "Selection and Assignment of STEM Adjunct Faculty Using Text Data Mining". In *2020 IEEE Integrated STEM Education Conference*, pp. 1-7, IEEE, 2020.

[27] M.J. Gomez, M. Calderón, V. Sánchez, F.J.G. Clemente, and J.A. Ruipérez-Valiente, "Large scale analysis of open MOOC reviews to support learners' course selection". *Expert Systems With Applications*, 210, p.118400, 2022.

[28] Y. Ishikawa, "Gender differences in vocabulary use in essay writing by university students". *Procedia-Social and Behavioral Sciences*, 192, pp. 593-600, 2015.

[29] C.M. Bell, P.M. McCarthy, and D.S. McNamara, "Using LIWC and Coh-Metrix to investigate gender differences in linguistic styles". In *Applied natural language processing: Identification, investigation and resolution*, pp. 545-556, IGI Global, 2012.

[30] S. Jones, and D. Myhill, "Discourses of difference? Examining gender differences in linguistic characteristics of writing," *Canadian Journal of Education/Revue canadienne de l'éducation*, pp. 456-482, 2007.

[31] J.A. Boser, "Gender Differences: Let's See Them in Writing", 1991.

[32] M.L. Newman, C.J. Groom, L.D. Handelman, and J.W. Pennebaker, "Gender differences in language use: An analysis of 14,000 text samples," *Discourse processes*, 45(3), pp. 211-236, 2008.

[33] M. Honnibal, and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing", 2017.